Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 14
TPCEE 2026

Contactless Electric-Field Sensing for Privacy-Preserving ASL Finger-Spelling Recognition

Katelyn L. Pei [1, *]

[1] Walt Whitman High School, 7100 Whittier Boulevard, Bethesda, MD 20817, US

* Corresponding Author Email: katielqp@gmail.com

Abstract. Hearing loss affects hundreds of millions of people worldwide and is projected to increase substantially by 2050. Yet sign-language interpreters remain scarce, and few people can communicate using sign language, creating persistent barriers in healthcare, education, and everyday life. Existing recognition technologies also have significant limitations: vision-based systems require continuous image capture and perform poorly under unfavorable lighting, whereas instrumented gloves are intrusive and prone to mechanical fatigue. Consequently, no current approach simultaneously provides contactless operation, durability, privacy preservation, and offline capability. This work presents a camera-free, glove-free system that recognizes American Sign Language (ASL) finger spelling using contactless quasi-static electric-field sensing. Two orthogonal three-axis sensor boards measure hand-induced electric-field perturbations, producing a six-channel time series for each finger-spelled letter. A Transformer encoder classifies these signals into 27 classes: the 26 alphabet letters plus an end-of-word token, compared with a chance level of 3.7%. The system is evaluated on a single-subject dataset containing 2,756 recorded clips under two evaluation protocols. Using the conventional random train-test split, the model achieves an accuracy of 86.12% ± 2.06% across five random seeds. However, this protocol is shown to be inappropriate for the dataset because each class was recorded as a single uninterrupted sequence with a median inter-clip interval of 1.535 seconds. Temporally adjacent clips exhibit a median cosine similarity of 0.997, and an untrained one-nearest-neighbor classifier achieves 88.26% accuracy under the same random split, indicating substantial near-duplicate leakage. When evaluated using a time-blocked protocol that keeps temporally adjacent samples on the same side of the train-test split, the Transformer's accuracy decreases to 63.66% ± 1.56%, a reduction of 22.5 percentage points that directly quantifies the effect of temporal leakage. Under neither protocol does the trained encoder significantly outperform the nearest-neighbor baseline. Nevertheless, the central modality finding remains unchanged and is independent of the classifier: six-channel contactless electric-field sensing enables reliable 27-class letter-level discrimination well above chance for a single user under fixed sensor placement. Per-class performance metrics and complete data acquisition metadata are provided.

Keywords: contactless gesture sensing; electric-field sensing; privacy-preserving assistive technology; finger-spelling recognition; evaluation protocol.

1. Introduction

1.1. A Persistent Communication Gap

Hearing loss affects hundreds of millions of people worldwide, and the number of people affected is projected to increase substantially by 2050 [1]. Many deaf and hard-of-hearing individuals rely on sign language and fingerspelling for communication, yet relatively few hearing people are proficient in sign language, and qualified interpreters remain scarce. Consequently, communication barriers persist in healthcare, education, workplaces, and everyday interactions. An effective gesture-to-language system should therefore be affordable, responsive, privacy-preserving, and capable of operating entirely offline. (See Figure 1, Figure 2)

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.



Figure 1. Conceptual illustration of the Whisper-Hands system. A contactless, camera-free, and glove-free device recognizes an in-air ASL fingerspelling gesture and vocalizes the intended sentence. The illustration was generated using an AI image model for conceptual visualization only.



Figure 2. Real-world context of the communication challenge addressed by this work, illustrated using openly licensed photographs. (a) The American Sign Language (ASL) manual alphabet, which serves as the target for the fingerspelling recognition system (public domain; Cwterp, Wikimedia Commons). (b) A sign-language interpreter facilitating communication between deaf and hearing individuals (photo by Stephan Roehl / Heinrich-Böll-Stiftung; CC BY-SA 2.0). (c) A hearing aid, representative of conventional assistive hearing technology (public domain; Jonas Bergsten, Wikimedia Commons).

1.2. Sensing Paradigms Involve Inherent Trade-Offs

Current gesture-recognition systems rely primarily on vision-based sensing or wearable devices, each of which offers distinct advantages and limitations [2]. Vision-based approaches can achieve high recognition accuracy but continuously capture images, raising privacy concerns and making performance sensitive to lighting, occlusion, and complex backgrounds. Wearable systems, such as flex-sensor gloves, can provide robust measurements of hand configuration and motion but require users to wear specialized hardware, reducing comfort while increasing cost and maintenance requirements. Alternative sensing modalities—including radar, radio-frequency (RF) sensing, surface electromyography, and ultrasonic sensing—can improve robustness in certain settings but typically require more specialized hardware, greater computational resources, or higher system cost.

These trade-offs motivate the exploration of alternative sensing modalities that can combine contactless operation with low cost, real-time processing, privacy protection, and fully offline deployment. In particular, electric-field sensing offers the potential to detect hand motion without capturing images or requiring physical contact with the user. However, its ability to resolve the fine-grained spatial configurations required for fingerspelling recognition has not been established.

1.3. Contactless Electric-Field Sensing

Whisper-Hands addresses these challenges through a contactless gesture-to-sentence system designed to operate entirely offline on a Raspberry Pi 5. The system uses quasi-static electric-field sensing

based on Microchip's MGC3130 GestIC technology [3,4]. Two orthogonally mounted three-axis sensing boards are integrated into a 3D-printed L-shaped bracket. The horizontal board measures in-plane hand motion along the X and Y axes, while the vertical board captures depth motion along the Z axis. Together, the two boards generate a six-channel spatiotemporal signal, which is acquired by an Arduino Uno through I²C and streamed to the Raspberry Pi for real-time processing.

The central research question is whether a sensing technology originally developed for proximity detection and simple gesture interactions can resolve the fine-grained hand configurations required for 27-class ASL fingerspelling recognition. Prior electric-field sensing interfaces have primarily demonstrated relatively coarse gestures and interactions [3,4]. This work therefore investigates whether the same sensing principle can be extended to letter-level discrimination across the complete 26-letter ASL manual alphabet, together with a period/end-of-word token.

1.4. A Two-Stage Deployability Study

For gesture recognition, Whisper-Hands employs a Transformer encoder [5], referred to as the ProGestureTransformer, to classify the six-channel electric-field signal into one of 27 output classes: the 26 ASL fingerspelling letters (A–Z) and a period/end-of-word token (pd). Rather than introducing a new neural-network architecture, this stage evaluates whether low-resolution electric-field sensing contains sufficient information for fine-grained letter recognition.

The model was trained on 2,756 gesture clips and achieved an accuracy of 86.12% ± 2.06% under a conventional random train-test split. However, when temporally adjacent recordings were constrained to remain within the same partition using a time-blocked split, accuracy decreased to 63.66% ± 1.56%. This 22.5-percentage-point difference highlights the distinction between recognizing highly correlated samples and evaluating generalization to temporally separated data. The source of this evaluation bias and its quantitative effect are examined in the Results section.

The second stage demonstrates end-to-end deployment by converting recognized letter sequences into natural-language text entirely offline. Letter sequences terminated by the pd token are expanded into complete sentences; for example, a sequence corresponding to a phrase can be converted to I need water. To enable on-device execution, the language model was adapted using LoRA fine-tuning [6] and knowledge distillation [7], using a Llama-3.1-8B-Instruct teacher and a Llama-3.2-1B-Instruct student [8], followed by 4-bit GGUF quantization [10].

This stage is intended primarily as a demonstration of deployability rather than as a novel language-generation method. In the available in-sample evaluation, the distilled student achieved performance comparable to, but did not exceed, that of a supervised TinyLlama-1.1B-Chat baseline [9]. Because the comparison involved different base models and training data, the result should not be interpreted as a controlled benchmark of language-model quality.

1.5. Contributions

This work makes three principal contributions.

1. Contactless fingerspelling recognition. We demonstrate a contactless six-channel electric-field sensing system capable of recognizing all 27 target classes, comprising the 26 ASL fingerspelling letters and a period/end-of-word token. The results show that a sensing modality traditionally used for proximity detection and coarse gesture interaction can support fine-grained, letter-level discrimination substantially above chance.

2. Rigorous evaluation of continuously recorded gesture data. We identify a substantial evaluation bias associated with conventional random train-test splitting of continuously recorded gesture datasets. Because samples from each class were collected in uninterrupted recording sessions, random partitioning can place highly correlated neighboring samples in both the training and testing sets, inflating measured accuracy by approximately 22.5 percentage points. We therefore use a time-

blocked evaluation protocol, together with an untrained nearest-neighbor control, to provide a more stringent assessment of temporal generalization.

3. End-to-end offline deployment. We demonstrate a complete offline gesture-to-sentence pipeline by integrating electric-field gesture recognition with a compressed language model capable of running on a Raspberry Pi 5.

The primary scientific contributions of this work are the demonstration of fine-grained recognition using contactless electric-field sensing and the identification of an important evaluation bias in continuously recorded gesture datasets. The language-generation stage serves primarily as a proof of practical deployment rather than as an advance in language modeling. The current system was validated using a single participant with fixed sensor placement; therefore, cross-user generalization and robustness to changes in sensor placement remain important areas for future work.

2. Related Work

Whisper-Hands builds on four complementary research areas: sign-language recognition (SLR), gesture-sensing modalities, sequence modeling for short multichannel time series, and language-model compression for edge deployment.

2.1. Vision-Based Sign-Language Recognition

Most SLR systems are vision-based, evolving from hand-crafted features to convolutional, recurrent, and attention-based models applied to RGB or RGB-D video [2]. Because these systems use visual inputs and datasets with class sets that differ from those used in Whisper-Hands, they are not directly comparable in terms of sensing modality or recognition task. Two limitations are particularly relevant. First, camera-based systems can be sensitive to illumination, occlusion, skin tone, camera viewpoint, and background clutter. Second, continuous image capture can introduce privacy concerns, particularly for systems intended for everyday or sensitive environments.

2.2. Wearable and Flex-Sensor Gloves

A second line of research captures hand articulation using wearable gloves equipped with flex, inertial, capacitive, or related sensors. These approaches avoid many of the illumination and background limitations of vision-based systems and can provide direct measurements of finger and hand motion. Their principal drawbacks, however, are physical and operational: the user must wear and, in some cases, calibrate the device, while flexible sensing elements may experience mechanical fatigue after repeated bending. Reported fatigue lifetimes vary substantially with sensor construction and operating conditions; therefore, claims of a specific lifetime should be interpreted as device-dependent rather than universal.

2.3. Contactless Electric-Field Gesture Sensing

A less common alternative detects hand motion through perturbations of a quasi-static electric field, enabling contactless estimation of hand position without cameras or physical sensors attached to the user [3]. The Microchip MGC3130 GestIC controller is a representative commercial implementation of this approach [4]. Electric-field sensing is inherently contactless and solid-state, thereby addressing both camera-related privacy concerns and the mechanical fatigue associated with wearable flex sensors. Existing applications, however, have primarily focused on coarse directional, proximity, or a relatively small number of predefined gestures [3, 4]. This leaves open the more demanding problem of fine-grained, multiclass discrimination required for alphabet-level finger-spelling recognition.

2.4. Transformers for Short Multichannel Time Series

The Transformer architecture replaces recurrent state updates with self-attention, allowing information from different time steps to be modeled directly and processed efficiently in parallel [5].

Transformer encoders can also be applied to multichannel time-series data by representing each time step as a feature vector and incorporating temporal or positional information. WhisperHands adopts this established architecture in the form of the ProGestureTransformer, which is a standard Transformer encoder adapted to a new sensing modality rather than a novel neural-network architecture. The contribution therefore lies primarily in demonstrating that the proposed electric-field signal contains sufficient information for fine-grained gesture classification.

2.5. Adaptation and Distillation for Edge Language Models

Running a generative language model fully offline on a single-board computer requires reducing both adaptation cost and model size while preserving sufficient linguistic capability. Low-Rank Adaptation (LoRA) addresses adaptation cost by freezing the pretrained model weights and learning a small set of low-rank parameter updates [6]. Knowledge distillation further enables model compression by transferring information from a larger teacher model to a smaller student model, typically by combining hard-label supervision with soft output distributions [7]. These techniques can be combined with open-weight model families such as Llama 3 [8] and Tiny-Llama [9] and deployed through lightweight inference runtimes such as llama.cpp [10]. In WhisperHands, the on-device language-model stage is presented primarily as a deployability and compression study rather than as evidence of a new language-model architecture or a general improvement in language-generation quality.

2.6. Positioning of WhisperHands

WhisperHands makes two principal contributions. The first is empirical: it demonstrates that a relatively low-dimensional, six-channel quasi-static electric-field signal can contain sufficient information for 27-way, letter-level discrimination. This finding is the central technical contribution because electric-field sensing, Transformer-based classification, LoRA adaptation, knowledge distillation, and 4-bit quantization are each established techniques rather than individually novel methods.

The second contribution is methodological. When each gesture class is recorded as a continuous burst, randomly splitting individual clips can place highly similar or temporally adjacent samples in both the training and test sets. Such a split can therefore measure the retrieval of near-duplicate samples rather than generalization to independently collected gestures. WhisperHands explicitly examines this distinction and treats the temporally separated evaluation as the more meaningful estimate of deployment-oriented generalization.

Relative to vision-based systems, WhisperHands eliminates optical failure modes and avoids continuous image capture. Relative to wearable gloves, it requires no device to be worn on the hand and avoids mechanical components that undergo repeated bending. The resulting system therefore occupies a distinct point in the design space: contactless, camera-free, and glove-free sensing combined with lightweight sequence classification and offline language processing.

3. Methods

3.1. System Overview

WhisperHands converts contactless hand motion into spoken sentences through a five-stage edge pipeline—Sense, Synchronize, Classify, Expand, and Speak—without a camera or glove. Two orthogonal electric-field sensing boards capture perturbations in the electric field caused by hand motion. An Arduino Uno reads the sensor outputs over I²C and streams the data to a Raspberry Pi 5, which constructs a six-channel time series for each gesture. The ProGestureTransformer then classifies each clip into one of 27 classes: the 26 letters a–z plus a period/end-of-word token, pd. The recognized letter codes are subsequently passed to a compact on-device language model for sentence generation and then converted to speech using text-to-speech (TTS). The first three stages constitute

the recognition stage, which is the core focus of this work, while the final two stages constitute the semantic-expansion stage, included as a deployability proof of concept. (See Figure 3)

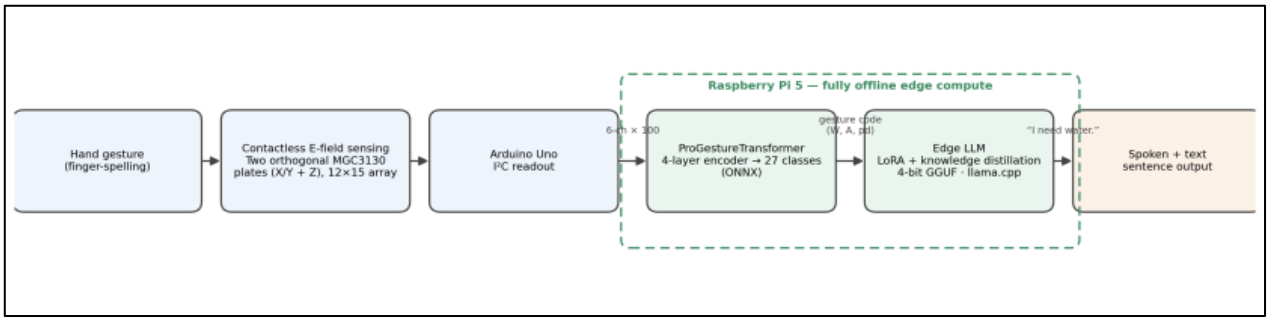

Figure 3. High-level architecture of the Whisper-Hands pipeline, from contactless electric-field sensing through the Pro-Gesture-Transformer classifier to the offline edge language model and spoken output.

At runtime, a per-letter loop appends each recognized letter to a buffer; upon detecting pd, the accumulated sequence is passed to the language model, and the buffer is then reset. (See Figure 4)

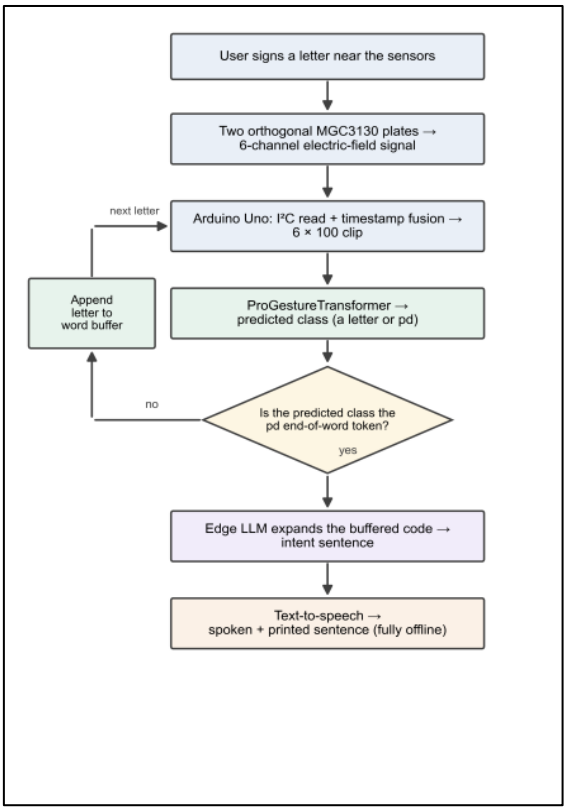

Figure 4. Runtime operation flowchart. Recognized letters accumulate in a word buffer until the pd end-of-word token triggers offline sentence generation and speech output.

3.2. Contactless Electric-Field Sensing

The Figure 5 sensing system uses the Microchip MGC3130 GestIC controller [4], which estimates three-dimensional hand position from perturbations in a quasi-static electric field. A transmit electrode establishes a low-frequency electric field, while a 12 × 15 receive-electrode array measures field distortions caused by an approaching hand [3]. The resulting six input channels comprise three-axis position estimates from each of the two orthogonal sensing boards, rather than raw electrode voltages. The key novelty is the use of these relatively coarse positional outputs for fine-grained gesture classification.

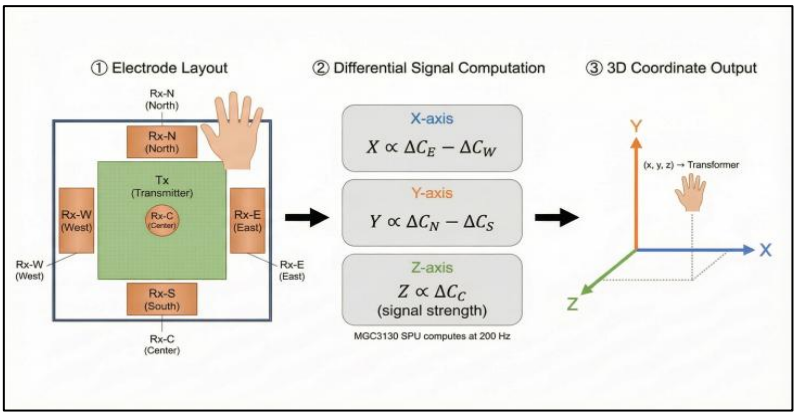

Figure 5. Contactless electric-field sensing principle of the MGC3130, illustrating the transmit-receive electrode layout, the differential signal computation for the X, Y and Z axes, and the resulting three-dimensional coordinate output.

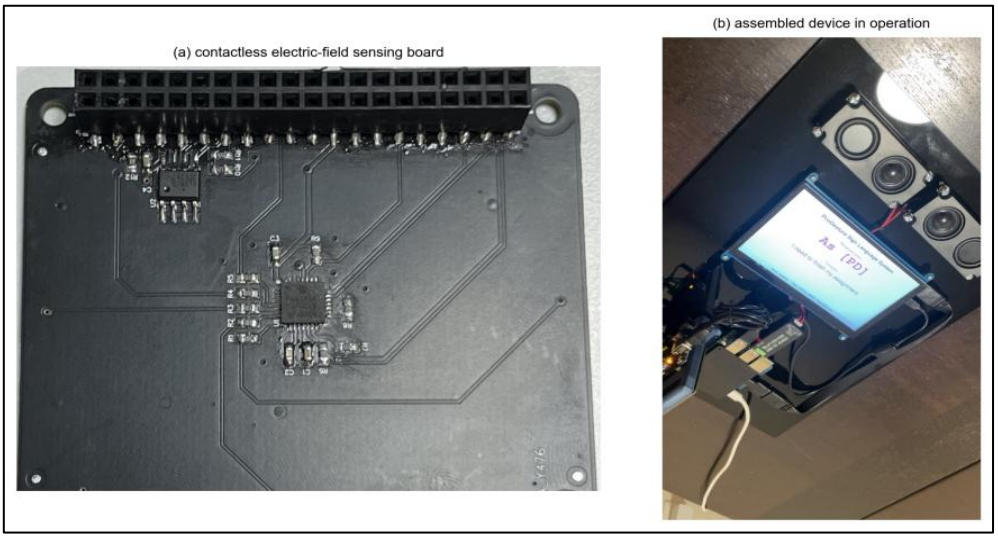

Figure 6. Assembled Whisper-Hands prototype (authors' own photographs): (a) the contactless electric-field sensing board, featuring the GestIC controller and fan-out receive-electrode traces; and (b) the integrated device, with the touchscreen, loudspeakers, Raspberry Pi 5, and Arduino Uno mounted on an acrylic base plate.

Two boards are mounted orthogonally in a 3D-printed L-shaped bracket to capture full three-dimensional hand motion: the horizontal board measures in-plane X and Y, while the vertical board captures depth Z, yielding the six channels M_X, M_Y, M_Z and A_X, A_Y, A_Z. (See Figure 7)

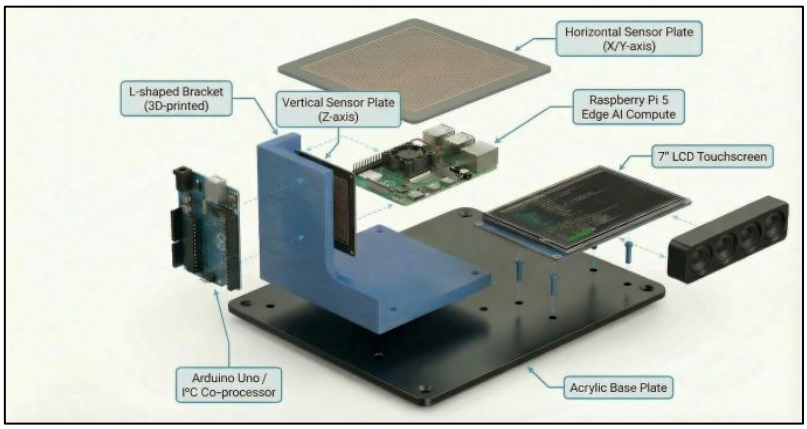

Figure 7. Exploded view of the orthogonal sensing hardware, showing the horizontal (X/Y) and vertical (Z) sensor boards, the 3D-printed L-shaped bracket, the Arduino Uno co-processor, the Raspberry Pi 5, and the acrylic base plate.

3.3. Data Acquisition and Preprocessing

The dataset comprises 2,756 gesture clips spanning 27 classes, with approximately 100–114 clips per class: 100 clips for six classes and 114 clips for class m. The pd token marks word completion and is treated as one of the 27 gesture classes. Each clip is resampled by linear interpolation onto a fixed 100-step temporal grid, producing a 6 x 100 input array. The six channel values are normalized by dividing by 65,535, corresponding to the full 16-bit measurement range. Representative traces for a, e, w, and pd are shown in Figure 8; the A_Y channel reaches the 16-bit measurement ceiling.

All clips were recorded by a single subject—the author—over two consecutive evenings, 26 and 27 December 2025, across four recording sessions. The session assignments were reconstructed from the millisecond-resolution acquisition timestamps embedded in the filenames. Session 1, recorded on 26 December from 20:12 to 20:52, contained 1,417 clips covering classes a–i and k–o. A subsequent 69-second supplementary recording at 21:27 contained 17 clips from classes b, m, and n. Session 2, recorded on 27 December from 20:25 to 20:59, contained 1,120 clips covering j and p–z, while Session 3, recorded from 21:42 to 21:49, contained 202 clips covering t and pd. Each clip is 1.0 s long and sampled at 100 Hz.

Importantly, each class was recorded within a single continuous session rather than being interleaved across sessions. Within a session, repetitions of the same class were recorded consecutively, with a median interval of 1.535 s between successive clips, leaving approximately 0.5 s of idle time between consecutive recordings of the same letter. This acquisition structure is reported explicitly because it constrains the admissible evaluation protocols and the interpretation of generalization performance, as discussed in the following subsection.

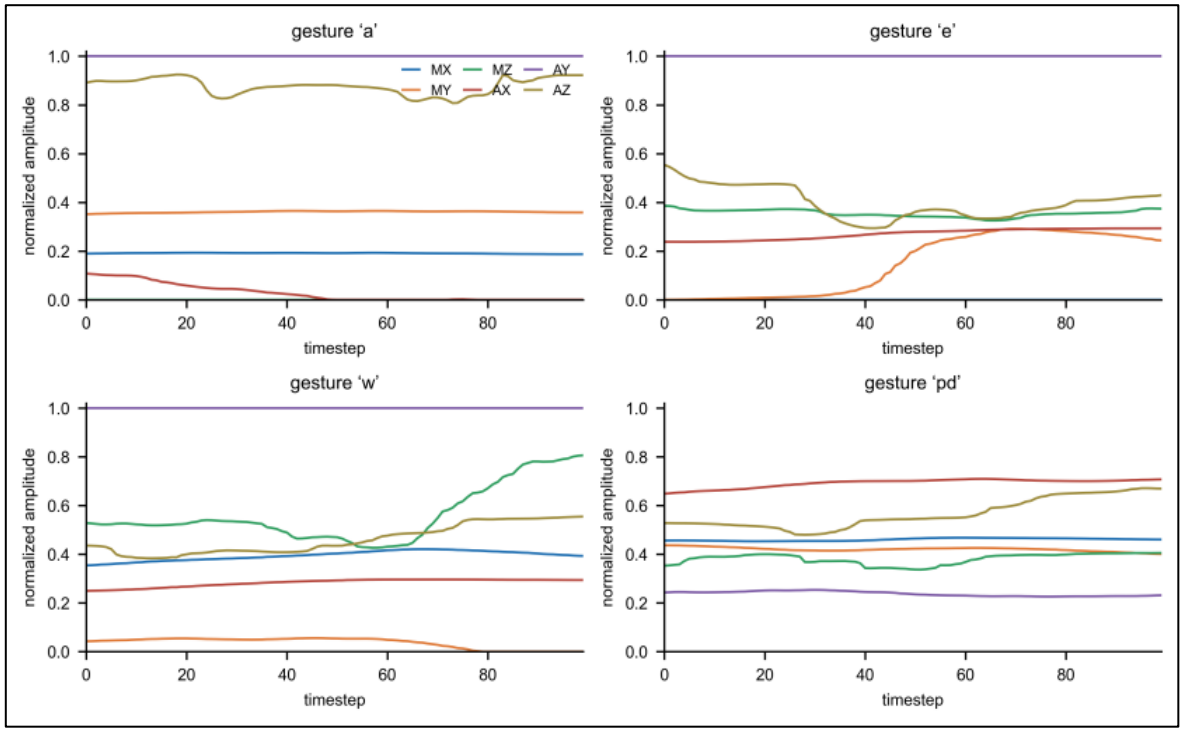

Figure 8. Representative normalized six-channel signals for the gestures a, e, w and the pd end-of-word token, illustrating class-distinct temporal patterns. The A_Y channel saturates at the 16-bit ceiling.

Because each class was recorded as a single continuous block, the choice of train–validation split directly determines what aspect of performance is being measured. We therefore evaluated four split protocols. The random 80/20 split (random_state = 42, without class stratification) reproduces the protocol used in the original evaluation. Stratified 5-fold cross-validation removes the class-balance irregularity of the random split while retaining random sample assignment. The time-blocked 80/20 split assigns, within each class, the earliest 80% of clips by acquisition timestamp to training and the

latest 20% to validation, preventing temporally adjacent recordings from appearing in different partitions. Time-blocked 5-fold cross-validation extends this approach by dividing each class-specific timeline into five consecutive segments and holding out each segment in turn.

All four protocols use identical preprocessing, model architecture, and hyperparameters. The random and time-blocked 80/20 evaluations are repeated across five random seeds, while the two 5-fold protocols use their respective five folds. Data augmentation is applied only to the training partition, using additive Gaussian noise $N(0,0.01)$ on the normalized 65535 scale and random amplitude scaling drawn from $Uniform(0.9,1.1)$. As a training-free control, a 1-nearest-neighbor (1-NN) classifier operating on flattened, normalized clips is evaluated under every protocol. This control provides a useful test of whether apparent recognition performance is attributable primarily to sample-level similarity and temporal proximity rather than to the learned representation: if the trained model cannot consistently outperform 1-NN, the evaluation may be capturing retrieval-like similarity more than genuine generalization. (See Figure 9)

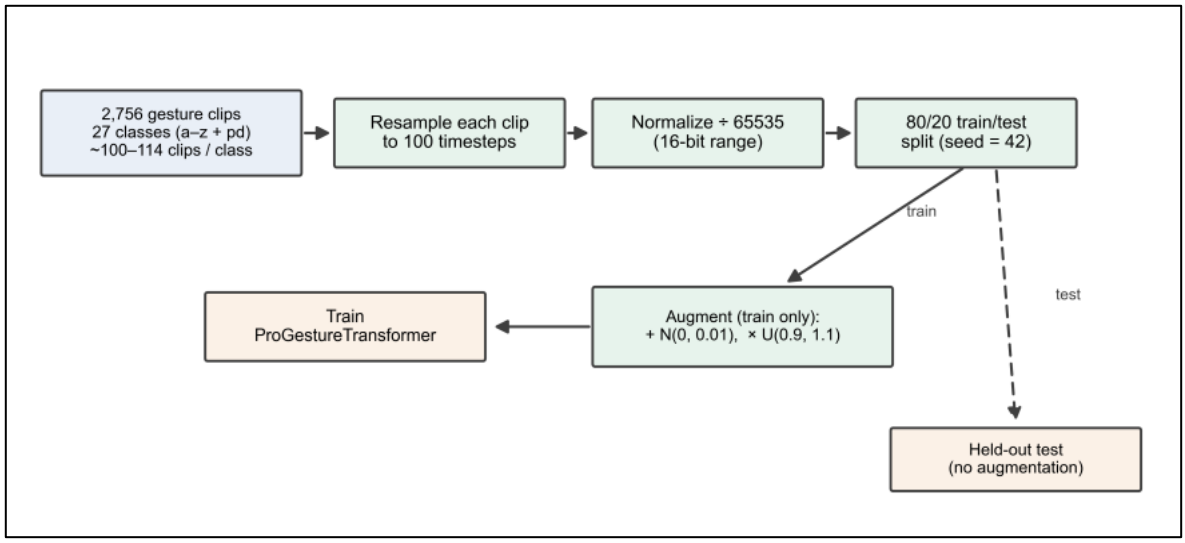

Figure 9. Data acquisition and preprocessing pipeline for the 2,756-clip, 27-class gesture dataset, including train-only augmentation and a held-out test split.

3.4. Gesture Recognition Model

ProGestureTransformer is implemented as a conventional Transformer encoder in PyTorch [5]. Each timestep is linearly projected from the six input channels into a 256-dimensional embedding and combined with a learned positional embedding of dimensions 1 x 100 x 256. The resulting sequence is processed by four Transformer encoder layers with $d_{model} = 256$, four attention heads, a feed-forward dimension of 512, dropout of 0.4, and batch-first tensor ordering. Each encoder layer employs multi-head scaled dot-product attention, defined as

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K, and V denote the query, key, and value matrices, respectively, and d_k is the dimensionality of each attention head.

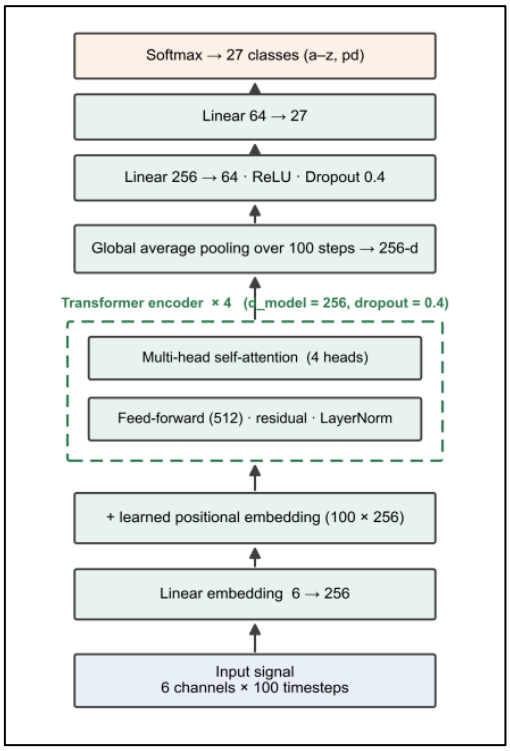

Figure 10. Pro-Gesture-Transformer architecture, from the six-channel input through the four-layer Transformer encoder and global average pooling to the 27-class classifier head.

The encoder output is global-average-pooled across the 100 timesteps to produce a single 256-dimensional representation (Figure 10). The classification head consists of $Linear(256 \rightarrow 64)$, ReLU activation, dropout $(p = 0.4)$, and $Linear(64 \rightarrow 27)$, producing one logit for each of the 27 classes. Training minimizes the cross-entropy loss

$$\mathcal{L}_{CE} = -\sum_{c=1}^{27} y_c log(p_c) \tag{2}$$

where y_c is the one-hot ground-truth label and p_c is the predicted probability for class c. Optimization uses AdamW with a learning rate of 4×10^{-4}, weight decay of 1×10^{-4}, and a batch size of 16.

Each configuration is trained for 200 epochs, with an average training time of approximately 2.6 s per epoch on an NVIDIA RTX 4090. Across the four evaluation protocols Table 1, five runs are performed for each protocol, for a total of 20 training runs. The model contains 2,154,011 trainable parameters (approximately 2.15 million) and requires approximately 2.3×10^8 multiply-accumulate operations per forward pass, corresponding to approximately 0.46 GFLOP when one MAC is counted as two floating-point operations. The trained model is also exported to ONNX for edge deployment [11].

To avoid obscuring the effect of epoch selection, final-epoch accuracy is used as the primary reported metric rather than selecting the epoch with the highest validation accuracy. Selecting the best epoch according to validation performance and then reporting that same validation performance introduces an optimistic bias because the validation set influences both model selection and evaluation. Nevertheless, the best validation accuracy observed during training is also reported in Table 2 to make the magnitude of this optimism explicit. Under the random 80/20 split, final-epoch accuracy is $86.12\% \pm 2.06\%$ across five seeds, compared with $88.99\% \pm 1.77\%$ when the best validation epoch is selected. These results are consistent with the original single run, which achieved a maximum validation accuracy of 89.13% and a final-epoch accuracy of 85.69%.

Table 1. ProGestureTransformer configuration, held constant across all four evaluation protocols.

Parameter	Value
Input shape	6 channels × 100 timesteps
Embedding dimension (d_{model})	256
Number of attention heads	4
Number of encoder layers	4
Feed-forward dimension	512
Dropout	0.4
Optimizer	AdamW
Learning rate	4×10^{-4}
Weight decay	1×10^{-4}
Batch size	16
Training epochs	200
Number of classes	27
Trainable parameters	2,154,011

3.5. On-Device Semantic Expansion

This stage maps a recognized letter sequence, such as "WA," to an intent-level sentence, such as "I need water," entirely on the device. It is presented as a deployment demonstration rather than a research contribution. The objective of this stage is limited to determining whether an instruction-following language model can be compressed sufficiently to run offline on the same Raspberry Pi 5 used by the gesture recognizer.

A Llama-3.1-8B-Instruct teacher model was LoRA fine-tuned in 4-bit precision using rank 16,120 training steps and 1,972 instruction examples. The fine-tuned teacher was then used for offline knowledge distillation into a Llama-3.2-1B-Instruct student model using cached top-200 teacher logits. The student was subsequently quantized to 4-bit GGUF format and executed through llama.cpp [10] on the Raspberry Pi 5. The student model was successfully executed fully offline on the device, establishing the feasibility of local semantic expansion.

Generation quality, however, was not established by a held-out evaluation. The only available scores were computed in-sample on the same instruction examples used during training. In addition, the separately constructed TinyLlama baseline [9] differed in both base model and training corpus, preventing a controlled comparison. These results are therefore not reported. A held-out evaluation using a matched test set and controlled baseline is left for future work. No gesture-recognition result depends on the semantic-expansion stage.

3.6. Edge Deployment and Operating Estimates

Both stages are executed on a single Raspberry Pi 5: the gesture recognizer uses the exported ONNX model [11], while the language model is executed in GGUF format through llama.cpp [10]. The complete demonstration—recognizing W, A, and pd and generating "I need water." entirely offline—is presented as a single qualitative example (n=1), rather than as a measure of end-to-end recognition or translation accuracy.

Operating energy is a design-stage estimate rather than a direct measurement, because power consumption was not instrumented during the experiments. Assuming an active power draw on the order of 6 W and an end-to-end latency of approximately 1.1 s, the energy per demonstrated inference is estimated to be approximately 7 J. This estimate should not be interpreted as a state-of-the-art energy-efficiency claim; lower-power implementations, including wearable glove-based systems, may consume less energy. Similarly, the reported bill-of-materials cost is an estimated prototype cost rather than a measured manufacturing cost.

4. Results

4.1. Contactless Finger-Spelling Recognition

ProGestureTransformer recognized all 27 finger-spelling classes from the contactless quasi-static electric-field signals with accuracy substantially above the 27-class chance level of approximately 3.7% under every evaluation protocol tested, but performance depended strongly on the partitioning strategy (Table 2). Under the random 80/20 split, the final-epoch accuracy was $86.12\% \pm 2.06\%$, whereas time-blocked 5-fold cross-validation yielded $64.48\% \pm 9.18\%$. These results demonstrate that the six-channel signals from the two orthogonal MGC3130 GestIC boards contain substantial discriminative information for the 26 letters plus the pd token without a camera or glove. At the same time, the large difference between random and time-blocked evaluation shows that reported accuracy is strongly dependent on the evaluation protocol and that random splitting can produce a substantially more favorable estimate of generalization. Figure 8 illustrates distinct class-dependent temporal patterns, although the A_Y channel reaches the 16-bit measurement ceiling.

The difference between the best-epoch and final-epoch results in Table 2 quantifies the effect of selecting an epoch using the validation data. Under the random split, the maximum validation accuracy averaged 88.99%, compared with 86.12% at the final epoch. Under time-blocked 5-fold cross-validation, the corresponding values differed by approximately 6.9 percentage points. The original single run exhibited a similar pattern, reaching a maximum validation accuracy of 89.13% near epoch 147 before finishing at 85.69%. Its peak falls within the five-seed best-epoch distribution reported here, indicating that the original result was representative rather than anomalous. Final-epoch accuracy is therefore used as the primary accuracy measure throughout the Results section.

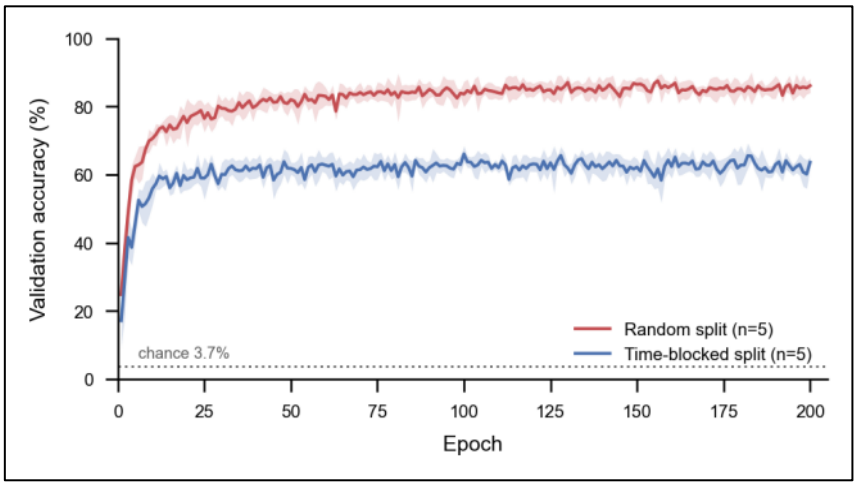

Figure 11. Validation accuracy of the ProGestureTransformer over 200 training epochs under the random and time-blocked protocols. Lines represent the mean over five runs, shaded bands indicate ± 1 standard deviation, and the dotted line denotes the 27-class chance level.

The two protocols separate within the first twenty epochs and never reconverge (Figure 11). Both curves plateau well before epoch 200, so the gap is not an artifact of stopping too early or too late; it is present throughout training. Dropout of 0.4 with train-only Gaussian-noise and amplitude-scaling augmentation limited overfitting in both cases.

4.2. Effect of Evaluation Protocol and Retrieval Baseline

The two evaluation protocols diverge within the first 20 training epochs and never reconverge (Figure 11). Both curves plateau well before epoch 200, indicating that the performance gap is not attributable to premature or excessive training but is present throughout optimization. Train-only Gaussian-noise and amplitude-scaling augmentation, together with dropout of 0.4, limits overfitting under both protocols, but does not eliminate the difference between random and time-blocked evaluation.

The gap follows directly from the structure of the recorded dataset. Each class was captured as a single continuous burst of approximately 100 repetitions, with a median spacing of 1.535 s between consecutive clips. Because each clip spans 1.0 s, successive recordings of the same letter contain only approximately 0.5 s of intervening idle time and therefore capture highly similar hand configurations. The median pairwise cosine similarity between consecutive same-class clips is 0.997, compared with 0.934 for clips of the same letter separated by more than 20 repetitions and 0.783 for arbitrary clip pairs. Across the dataset, the nearest neighbor of a clip is one of its immediate temporal neighbors for 58% of clips, and, among same-class nearest neighbors, the median temporal separation is one repetition. A random split distributes these highly similar neighboring clips across both the training and validation partitions.

A training-free 1-nearest-neighbor (1-NN) control makes the consequence explicit. The classifier operates directly on the flattened normalized clips, storing the training examples and assigning each validation clip the label of its closest training example. Under the random 80/20 split, 1-NN achieves $88.26\% \pm 1.50\%$, with no statistically significant difference from the ProGestureTransformer's $86.12\% \pm 2.06\%$ (paired t-test $p = 0.18$). Under the time-blocked 80/20 protocol, the same baseline falls to 64.35%. Thus, the high random-split accuracy appears to be strongly influenced by the presence of highly similar temporal neighbors across the two partitions, rather than reflecting generalization to temporally separated instances of a gesture.

A second structural confound reinforces this interpretation. Classes $a - o$ were recorded on 26 December, whereas $p - z$ and pd were recorded on 27 December. Consequently, class identity is strongly aligned with acquisition date. Any change in electrode baseline, sensor response, environmental conditions, or other recording characteristics between the two evenings could therefore provide a shortcut for distinguishing large groups of classes independently of gesture shape. Time-blocking within classes does not eliminate this confound because it separates samples temporally but does not separate acquisition days or recording conditions.

Under time-blocked 5-fold cross-validation, the ProGestureTransformer achieves $64.48\% \pm 9.18\%$ final-epoch accuracy, compared with $68.37\% \pm 8.05\%$ for 1-NN. The paired difference is -3.89 percentage points in favor of the 1-NN baseline ($p = 0.027$). At the best validation epoch, the Transformer exceeds 1-NN by 2.96 percentage points, but the difference is not statistically significant ($p = 0.11$). Thus, under the more stringent time-blocked evaluation, the 2.15-million-parameter encoder does not demonstrate a clear advantage over direct template matching. The recognition result therefore primarily establishes that the contactless electric-field modality contains letter-level information, rather than demonstrating that the proposed Transformer architecture provides superior generalization on the present dataset. Establishing an advantage over retrieval-based methods will likely require substantially greater data volume, subject diversity, and session diversity, and may also motivate further architectural investigation.

Table 2. Recognition accuracy under four evaluation protocols, with a training-free 1-nearest-neighbor control on identical splits (mean ± s.d.).

Evaluation protocol	Runs	PGT, final epoch (%)	PGT, best epoch (%)	1-NN control (%)
Random 80/20	5 seeds	86.12 ± 2.06	88.99 ± 1.77	88.26 ± 1.50
Stratified 5-fold	5 folds	86.50 ± 1.70	89.22 ± 1.58	88.32 ± 1.47
Time-blocked 80/20	5 seeds	63.66 ± 1.56	68.88 ± 1.17	64.35
Time-blocked 5-fold	5 folds	64.48 ± 9.18	71.33 ± 5.49	68.37 ± 8.05

Per-class precision, recall, and F1 scores from the time-blocked 5-fold out-of-fold predictions are reported in Table 3; the corresponding confusion matrix and sorted per-class F1 scores are shown in Figure 12. The macro-averaged F1 is 64.10%, close to the 64.40% overall out-of-fold accuracy, indicating that aggregate performance is not dominated by a small number of substantially overrepresented classes. Performance nevertheless varies considerably across classes: z (91.8%), x (85.7%), and o (85.6%) are the most reliably recognized, whereas r (29.3%), t (29.8%), and u

(38.7%) are the weakest. Errors are concentrated in several class pairs rather than being uniformly distributed, with the largest single confusions being $s \rightarrow t$ (30 clips) and $g \rightarrow n$ (25 clips).

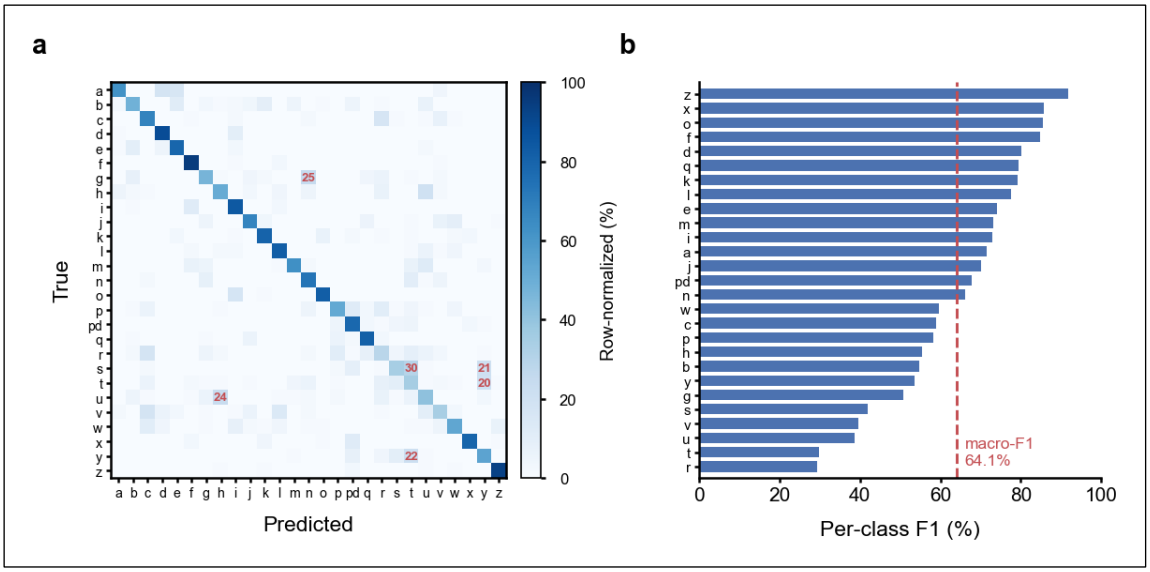

Figure 12. Classifier diagnostics from time-blocked 5-fold out-of-fold predictions at the final epoch: (a) row-normalized 27-class confusion matrix, with the six largest off-diagonal counts annotated; and (b) per-class F1 scores in ascending order, with the macro-average indicated.

Table 3. Per-class precision, recall and F1 scores from the time-blocked 5-fold out-of-fold predictions

Class	P	R	F1	Class	P	R	F1	Class	P	R	F1
a	84.0	62.4	71.6	j	72.9	67.3	70.0	r	29.9	28.7	29.3
b	63.0	48.6	54.8	k	78.1	80.4	79.2	s	53.0	34.6	41.9
c	51.9	68.0	58.8	l	73.5	82.2	77.6	t	25.9	35.0	29.8
d	73.0	89.0	80.2	m	86.8	63.2	73.1	u	36.5	41.2	38.7
e	69.9	79.0	74.2	n	60.0	73.6	66.1	v	46.7	34.3	39.5
f	76.6	95.0	84.8	o	89.2	82.2	85.6	w	67.5	53.5	59.7
g	55.3	47.0	50.8	p	66.2	52.0	58.2	x	93.1	79.4	85.7
h	60.7	51.0	55.4	pd	59.7	78.4	67.8	y	52.9	54.5	53.7
i	64.4	84.2	73.0	q	77.6	81.4	79.4	z	90.5	93.1	91.8

4.3. On-Device Semantic Expansion

The compressed language stage was executed end to end on the Raspberry Pi 5 entirely offline. A roughly one-billion-parameter student model, obtained through knowledge distillation and 4-bit quantization, is invoked by the recognizer when the pd token is detected. This establishes the feasibility of running the semantic-expansion stage locally on the same hardware as the gesture recognizer. Generation quality is not reported because the only available evaluation was in-sample and used an uncontrolled baseline; those results are therefore withheld pending a held-out, matched evaluation.

4.4. End-to-End Offline Demonstration

The recognition and semantic-expansion stages were integrated into a working offline demonstration on the Raspberry Pi 5. A user signs letters in the air and then produces the pd gesture to mark word completion and trigger sentence generation. ProGestureTransformer classifies each gesture clip, each recognized letter is appended to a buffer, and, upon detecting pd, the accumulated sequence is passed to the language model before the buffer is reset. The system then displays and speaks a generated sentence such as “I need water.” Figure 4 illustrates this control flow. This demonstration is a single

qualitative example ($n = 1$ phrase): it confirms that the complete processing path can operate offline on the target hardware but does not quantify end-to-end accuracy, latency variability, or error propagation.

4.5. Estimated Edge Efficiency

The reported efficiency figures are design-stage estimates rather than bench measurements because power consumption was not directly metered. Assuming an active power draw on the order of 6 W, the energy required for one demonstrated translation is estimated at approximately 7 J, with an estimated latency of approximately 1.1 s. The contactless solid-state electrode array contains no moving parts and therefore avoids mechanical wear associated with flex-sensor gloves; because the system does not use a camera, it also does not capture images. These characteristics represent design advantages rather than claims of minimum energy consumption, since lower-power wearable implementations may consume less energy. The bill-of-materials cost is likewise an estimate rather than a measured production cost.

4.6. Discussion

WhisperHands demonstrates the feasibility of a fully contactless, solid-state sensing front end for 27-class finger-spelling recognition in a single-user setting, together with offline sentence expansion on the same Raspberry Pi 5. Recognition accuracy substantially exceeds the 3.7% chance level under all tested protocols, but the magnitude of the result depends strongly on how the data are partitioned: final-epoch accuracy is $86.12\% \pm 2.06\%$ under the random split but $64.48\% \pm 9.18\%$ under time-blocked 5-fold cross-validation. The corresponding 1-NN control closely tracks the Transformer, indicating that the present data do not establish a clear advantage for the learned encoder over direct template matching.

The primary contribution is therefore best interpreted as a modality feasibility result rather than a definitive model-performance result. Contactless electric-field sensing contains sufficient information to distinguish the 26 letters and pd token for one user under the recorded conditions. The current experiments do not establish cross-user generalization, robustness to changes in hand placement or distance, or superiority of the proposed architecture over simpler retrieval-based methods.

The system also avoids camera-based sensing and its associated image-capture requirements [2], while the solid-state electric-field front end contains no mechanically flexing sensing elements [3,4]. These are potentially useful characteristics for privacy-sensitive or mechanically robust interfaces. However, given the single-subject dataset and prototype-level evaluation, WhisperHands should be regarded as a feasibility demonstration rather than a deployable assistive system.

4.7. Interpreting the Recognition Result

The ProGestureTransformer extracts discriminative temporal structure from a six-channel sensing modality with substantially less spatial information than camera-based systems [5]. However, the magnitude of the measured accuracy depends strongly on the evaluation protocol. Accuracy decreases from $86.12\% \pm 2.06\%$ under the random split to $64.48\% \pm 9.18\%$ under time-blocked 5-fold cross-validation, while the 1-NN control remains competitive under both protocols.

The defensible interpretation is therefore that contactless electric-field sensing resolves letter-level distinctions under the present experimental conditions, rather than that the Transformer itself has demonstrated superior generalization. The classifier does not outperform template matching on this dataset, suggesting that substantially larger and more diverse data will be required to determine whether learned representations provide an advantage over direct similarity-based recognition.

Three limitations are particularly important. First, all 2,756 clips were recorded from a single subject across four sessions on two consecutive evenings, so the results provide no evidence of cross-user generalization. Electric-field measurements may also depend on hand size, distance, orientation, and

placement. Second, class identity is strongly confounded with acquisition date: classes $a - o$ were recorded on the first evening, whereas $p - z$ and pd were recorded on the second. Any inter-session sensor or environmental drift could therefore provide a shortcut correlated with class identity. Third, each class was recorded as a single continuous burst, creating the near-duplicate structure described in Section 4.2 and making even the time-blocked estimate dependent on within-session temporal variation rather than on genuinely independent repetitions.

4.8. Interpreting the Semantic-Expansion Stage

The semantic-expansion stage is a deployment demonstration rather than a validated language-model result. Knowledge distillation from a LoRA-fine-tuned Llama-3.1-8B-Instruct teacher [6,8] into a Llama-3.2-1B-Instruct student, followed by 4-bit GGUF quantization, produced a roughly one-billion-parameter model that executes entirely offline on the Raspberry Pi 5 through llama.cpp [7,10]. This establishes that sentence expansion can be accommodated on the same edge device as the gesture recognizer.

It does not establish generation quality. The available language-model scores were obtained in-sample and used an uncontrolled baseline and are therefore not reported. A matched held-out evaluation is required before any claim about translation quality or distillation effectiveness can be made.

Treating pd as an explicit end-of-word class allows the gesture recognizer to perform word segmentation directly and keeps the runtime pipeline fully offline. This provides a simple deterministic handoff between recognition and language generation, but it should not be interpreted as a learned linguistic grammar.

4.9. Limitations

The principal limitations concern data diversity, evaluation independence, semantic validation, and hardware measurement. The recognition dataset contains one subject, one continuous recording burst per class, and two acquisition days, limiting conclusions about cross-user and cross-session generalization. The random split can distribute near-duplicate samples across training and validation sets, while the time-blocked protocol still does not provide independent sessions or subjects. The semantic-expansion stage was demonstrated qualitatively but not evaluated on held-out data. Finally, the approximately 7 J energy consumption and 1.1 s latency are estimates based on an assumed active power of approximately 6 W rather than direct measurements, and the bill-of-materials cost is also estimated.

4.10. Future Work

The highest priority is to expand the dataset rather than further optimize the current evaluation. Each class should be recorded across multiple subjects and multiple days, with repetitions distributed across independent sessions. Such a dataset would enable leave-one-session-out and leave-one-subject-out evaluation while breaking the current alignment between class identity and acquisition day. It would also provide a meaningful test of whether a learned encoder can outperform the nearest-neighbor baseline that remains competitive on the present data.

For semantic expansion, future experiments should use a held-out evaluation set, matched teacher and baseline models, and a controlled training corpus so that generation quality and the benefit of knowledge distillation can be evaluated independently of the recognition stage. On-device power metering should replace the current energy estimate, and repeated latency measurements should be collected under representative workloads. Finally, the letter-to-word and word-to-sentence composition should be formalized as an explicit decoding grammar or finite-state process, allowing segmentation, invalid sequences, and language-model failures to be evaluated systematically.

5. Conclusion

Contactless quasi-static electric-field sensing can support fine-grained finger-spelling recognition under controlled, single-user laboratory conditions and can serve as the front end of a fully on-device assistive communication pipeline. WhisperHands integrates contactless sensing, Transformer-based gesture recognition, and offline language generation on a single Raspberry Pi 5, enabling finger-spelled input to be converted into spoken output without a camera, glove, or network connection. The present work establishes the feasibility of this architecture while also demonstrating why evaluation protocol is critical for interpreting recognition performance.

5.1. Summary of Contributions

The primary contribution is the demonstration of contactless finger-spelling recognition from quasi-static electric-field sensing [3,4]. Two orthogonally mounted three-axis MGC3130 GestIC boards produce a six-channel signal that captures three-dimensional hand motion. A ProGestureTransformer encoder with approximately 2.15 million parameters [5] recognizes 27 classes—the 26 letters plus the pd end-of-word token—with final-epoch accuracy of $86.12\% \pm 2.06\%$ under a random 80/20 split and $64.48\% \pm 9.18\%$ under time-blocked 5-fold cross-validation, compared with a 27-class chance level of approximately 3.7%.

An important secondary contribution is the evaluation-protocol analysis. Because each class was recorded as a single continuous burst, random splitting distributes highly similar temporal neighbors across the training and validation sets and therefore rewards near-duplicate retrieval. A training-free 1-nearest-neighbor control closely tracks the Transformer across the evaluation protocols and does not establish a consistent advantage for the learned encoder. The defensible conclusion is therefore that the sensing modality contains letter-level information for a single user under fixed experimental conditions, rather than that the present architecture has demonstrated superior generalization. The trained model was exported to ONNX for edge deployment [11].

The semantic-expansion component is a deployment demonstration rather than a validated modeling contribution. Knowledge distillation from a LoRA-fine-tuned Llama-3.1-8B-Instruct teacher [6,8] produced a Llama-3.2-1B-Instruct student, which was quantized to 4-bit GGUF format (Q4_K_M) and executed through llama.cpp [10] on the Raspberry Pi 5. This demonstrates that the sentence-expansion stage can run on the same edge device as the gesture recognizer. Its generation quality, however, remains unestablished and is therefore not claimed.

5.2. Significance and Future Work

The contactless and offline design offers potentially useful properties for assistive interfaces: it captures no images and avoids the mechanical sensing elements used in flex-sensor gloves. These characteristics may reduce privacy concerns associated with camera-based systems and mechanical wear associated with flex-based sensing. However, the present work is a feasibility demonstration rather than a deployable assistive system.

Several limitations define the scope of the conclusions. The dataset contains 2,756 clips from a single subject, recorded across four sessions on two consecutive evenings, with class identity strongly confounded with acquisition day. Consequently, cross-user and independent cross-session generalization remain untested. The trained encoder also does not consistently outperform the 1-NN control, so the present results primarily establish the information content of the sensing modality rather than a definitive advantage of the Transformer architecture. Finally, the approximately 7 J per-translation energy consumption and 1.1 s latency are design-stage estimates based on an assumed active power of approximately 6 W, rather than direct measurements; prototype cost is likewise estimated.

The next stage of development should therefore focus on data diversity and independent evaluation. Recording multiple subjects across multiple days and independent sessions would enable leave-one-

subject-out and leave-one-session-out evaluation while eliminating the current alignment between class identity and acquisition day. Such data would also provide a meaningful test of whether a learned temporal encoder can outperform nearest-neighbor retrieval. For semantic expansion, a held-out evaluation with matched baselines and controlled training data is needed to establish generation quality. Finally, direct power metering and repeated latency measurements should replace the current operating estimates, and the letter-to-sentence composition should be formalized and evaluated as a complete decoding pipeline.

6. Data and Code Availability

The 2,756-clip, six-channel dataset, trained model weights, exported ONNX model, and training, evaluation, and analysis code are available from the authors upon reasonable request and will be made publicly available in a repository. Subject and session metadata are documented in the Methods section. The per-clip acquisition timestamps required to reproduce the time-blocked evaluation splits are preserved in the released filenames.

References

[1] World Health Organization. (2021). World report on hearing. Geneva: World Health Organization. https://www.who.int/publications/i/item/9789240020481 (IRIS: https://apps.who.int/iris/handle/10665/341229)

[2] Rastgoo, R., Kiani, K., & Escalera, S. (2021). Sign language recognition: A deep survey. Expert Systems with Applications, 164, 113794. https://doi.org/10.1016/j.eswa.2020.113794

[3] Tang, X., & Yan, S. (2021). Ultra-low power hand gesture sensor using electrostatic induction. Sensors, 21(24), 8268. https://doi.org/10.3390/s21248268

[4] Microchip Technology Inc. (2017). MGC3030/3130 3D tracking and gesture controller data sheet (DS40001667). Chandler, AZ: Microchip Technology. https://ww1.microchip.com/downloads/aemDocuments/documents/OTH/ProductDocuments/DataSheets/MGC3030-3130-3D-Tracking-and-Gesture-Controller-40001667F.pdf

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017). arXiv:1706.03762. https://arxiv.org/abs/1706.03762

[6] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations (ICLR 2022). arXiv:2106.09685. https://arxiv.org/abs/2106.09685

[7] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv:1503.02531 (NIPS 2014 Deep Learning Workshop). https://arxiv.org/abs/1503.02531

[8] Grattafiori, A., et al. (Llama Team, AI @ Meta). (2024). The Llama 3 herd of models. arXiv:2407.21783. https://arxiv.org/abs/2407.21783

[9] Zhang, P., Zeng, G., Wang, T., & Lu, W. (2024). TinyLlama: An open-source small language model. arXiv:2401.02385. https://arxiv.org/abs/2401.02385

[10] Gerganov, G., et al. (2023). llama.cpp: LLM inference in C/C++ [Computer software]. GitHub. https://github.com/ggml-org/llama.cpp

[11] ONNX. (2017). ONNX: Open Neural Network Exchange [Computer software/standard]. https://onnx.ai (GitHub: https://github.com/onnx/onnx)