Mathematical and Computational Applications

MDPI

Article

Terrain-Aware Head Gesture Recognition for Turret Control Using Helmet-Mounted IMUs and Vehicle Vibration Fusion

Lonwabo Rooibaard [1,2,*], Dithoto Modungwa [1], Malusi Sibiya [2] and Thanyani Pandelani [2]

[1] Defence and Security, Landward Sciences, Council for Scientific and Industrial Research (CSIR), Pretoria 0001, South Africa; d.modungwa@csir.co.za
[2] College of Science, Engineering and Technology, University of South Africa (UNISA), Pretoria 0003, South Africa; sibiym@unisa.ac.za (M.S.); epandet@unisa.ac.za (T.P.)
* Correspondence: l.rooibaard@csir.co.za

Abstract

Head gesture-based control using inertial measurement units (IMUs) provides an intuitive alternative to conventional human–machine interfaces for mobile and vehicle-mounted systems. However, gesture recognition reliability degrades significantly under terrain-induced vibration and mechanically dynamic operating conditions. This study investigates terrain-aware head gesture recognition through the integration of helmet-mounted IMU measurements and vehicle vibration sensing to improve discrimination between intentional gestures and non-intentional motion artefacts. Vehicle vibration data were collected from a patrol vehicle traversing the Ndumo Border Patrol route, characterised by variable terrain roughness and dynamic excitation profiles. Triaxial seat-rack acceleration data were acquired at 10 kHz, anti-alias filtered and down sampled to 100 Hz before being integrated with IMU-derived head motion measurements using both vibration-aware data augmentation and early sensor fusion strategies. FFT-based spectral processing and classification using a lightweight fully connected neural network (FCNN) were implemented using the Edge Impulse framework. Experimental evaluation was performed using temporally independent training and testing segments to reduce overlap leakage and ensure realistic generalisation assessment. Results demonstrate that terrain-informed sensing substantially improves operational robustness under mobile conditions. The early-fusion approach achieved 98.45% independent-test accuracy under float32 inference and maintained 90.02% accuracy after int8 quantization, corresponding to an accuracy reduction of 8.43 percentage points. In comparison, the Clean IMU and vibration-augmented models exhibited reductions of 20.13 and 27.02 percentage points, respectively, demonstrating greater sensitivity to quantization. The evaluated models also exhibited low inference latency and compact memory requirements, supporting their suitability for real-time edge implementation. These findings demonstrate that treating terrain vibration as contextual information, rather than solely as environmental noise, can improve the quantization robustness and deployment characteristics of IMU-based gesture-recognition systems intended for mechanically dynamic platforms.

Keywords: head gesture recognition; IMU sensor fusion; vibration compensation; embedded machine learning; mobile platforms; fully connected neural networks; human–machine interaction

Academic Editors: Saurav Mallik, Priyanka Roy and Zhongming Zhao

Received: 31 May 2026
Revised: 15 July 2026
Accepted: 24 July 2026
Published: 18 September 2026

1. Introduction

Human–machine interaction (HMI) systems based on wearable motion sensing are receiving increasing attention as alternatives to conventional manual interfaces, particularly in applications where operator fatigue, constrained ergonomics, and persistent cognitive workload reduce operational effectiveness [1,2]. Among these approaches, head motion and head gesture interfaces based on inertial measurement units (IMUs) are attractive because they are low-cost, portable, non-invasive, and capable of directly capturing operator motion at the point of interaction. Recent studies have demonstrated the effectiveness of head-worn inertial sensing in rehabilitation, assistive technologies, activity monitoring, and wearable HMI applications, highlighting the potential of IMU-based interfaces for intuitive control and interaction tasks [1].

Most contemporary head gesture recognition systems rely on combinations of inertial sensing, signal processing, and machine-learning classification techniques. Prior work includes IMU-based head tracking for rehabilitation systems, activity-aware head gesture recognition using dynamic time warping, and multimodal head motion tracking approaches incorporating inertial and acoustic sensing [3–6]. These studies generally report high recognition accuracy under controlled or semi-controlled conditions, demonstrating the feasibility of IMU-based interaction for gesture-driven applications. However, many existing approaches implicitly assume relatively stable operating environments in which sensor measurements are dominated primarily by intentional user motion.

These assumptions become increasingly problematic when IMU-based interfaces are deployed on mechanically dynamic mobile platforms. In such environments, inertial measurements are influenced not only by intentional operator movement, but also by terrain-induced vibration, structural oscillation, suspension dynamics, and vibration rectification effects, which may significantly degrade measurement accuracy [6–8]. Previous investigations into microelectromechanical system (MEMS) inertial sensors operating in high-vibration conditions have shown that external mechanical excitation can significantly degrade measurement reliability and classification stability, often necessitating both mechanical isolation and signal-level compensation strategies. At the same time, recent reviews of wearable gesture-recognition systems indicate that most studies primarily focus on recognition accuracy, sensor modality selection, and classifier optimisation, while the explicit modelling of environmental excitation as contextual information remains comparatively underexplored [6].

In parallel, recent advances in TinyML and Edge Artificial Intelligence (Edge AI) have enabled compact neural-network models to execute directly on resource-constrained embedded hardware platforms [9–11]. These developments have made real-time local inference increasingly feasible for wearable sensing and mobile control applications, where low latency, reduced communication overhead, low power consumption, and independence from cloud connectivity are operationally important. Nevertheless, deploying robust gesture-recognition systems on embedded platforms remains challenging under mechanically dynamic conditions, particularly when numerical precision reduction and model quantization are required for lightweight deployment [6].

Despite considerable advances in wearable gesture recognition, the influence of terrain-induced vibration on embedded gesture-recognition systems remains insufficiently investigated, particularly under quantized edge-AI deployment conditions. Accordingly, this study addresses the following research questions: (RQ1) Can intentional head gestures be reliably distinguished from terrain-induced motion artefacts under mobile operating conditions? (RQ2) Does vibration-aware data augmentation improve gesture-recognition robustness? (RQ3) Does early multichannel fusion improve embedded deployment robustness compared with conventional IMU-only approaches?

https://doi.org/10.3390/mca31050194

Corresponding to these research questions, the following hypotheses were formulated: (H1) Intentional head gestures can be reliably distinguished from terrain-induced motion artefacts under mechanically dynamic operating conditions; (H2) vibration-aware data augmentation improves the robustness of IMU-based gesture recognition compared with the clean IMU baseline; and (H3) early channel-level fusion of IMU and terrain-vibration signals improves embedded deployment robustness compared with conventional IMU-only approaches. The independent variables investigated include sensing modality, vibration condition, model architecture, and quantization type, while the dependent variables comprise classification accuracy, precision, recall, F1-score, inference latency, RAM utilization, and flash memory requirements.

This study investigates a terrain-aware head gesture recognition framework that integrates helmet-mounted IMU measurements with vehicle-mounted vibration sensing to improve discrimination between intentional operator gestures and non-intentional motion artefacts. Real-world terrain vibration data were collected from a patrol vehicle traversing the Ndumo Border Patrol route, characterised by variable terrain roughness and dynamic mechanical excitation. Rather than treating terrain-induced vibration solely as noise to be suppressed, this work explores the use of vehicle vibration measurements as contextual information within the recognition process. Two complementary approaches are investigated: (i) vibration-aware data augmentation and (ii) early multimodal sensor fusion between head motion and terrain-vibration measurements.

To evaluate these approaches, embedded machine-learning models were implemented using the Edge Impulse framework and assessed under both full-precision and 8-bit quantized deployment conditions. Experimental evaluation was performed using temporally independent training and testing segments to reduce overlap leakage and provide a more realistic assessment of generalisation performance under unseen terrain conditions. The results demonstrate that incorporating terrain-induced vibration as contextual information improves operational robustness for head gesture recognition on mechanically dynamic platforms while maintaining computational characteristics suitable for embedded edge deployment. These findings highlight the potential of context-aware multimodal sensing for next-generation wearable HMI systems operating in mobile and vibration-intensive environments.

The principal contributions of this work include vibration-aware data augmentation, early multichannel fusion of IMU and terrain-vibration measurements, and evaluation of embedded deployment robustness under quantized inference conditions.

2. Materials and Methods

2.1. System Overview

The proposed system implements a terrain-aware head gesture recognition framework that integrates operator head motion sensing with environmental vibration measurements to improve gesture classification robustness under mobile operating conditions. The framework combines a helmet-mounted inertial measurement unit (IMU) with a vehicle-mounted triaxial accelerometer to enable contextual interpretation of operator motion in the presence of terrain-induced vibration. Figure 1 illustrates overall processing pipeline consists of signal acquisition, preprocessing, segmentation, feature extraction, feature fusion, and lightweight embedded classification. The system architecture comprises the following functional stages:

1. Signal acquisition;
2. Signal conditioning and preprocessing;
3. Sliding-window segmentation;
4. FFT-based spectral analysis;

5. Early channel-level fusion of IMU and terrain-vibration signals;
6. Lightweight FCNN-based classification;
7. Embedded control output generation.

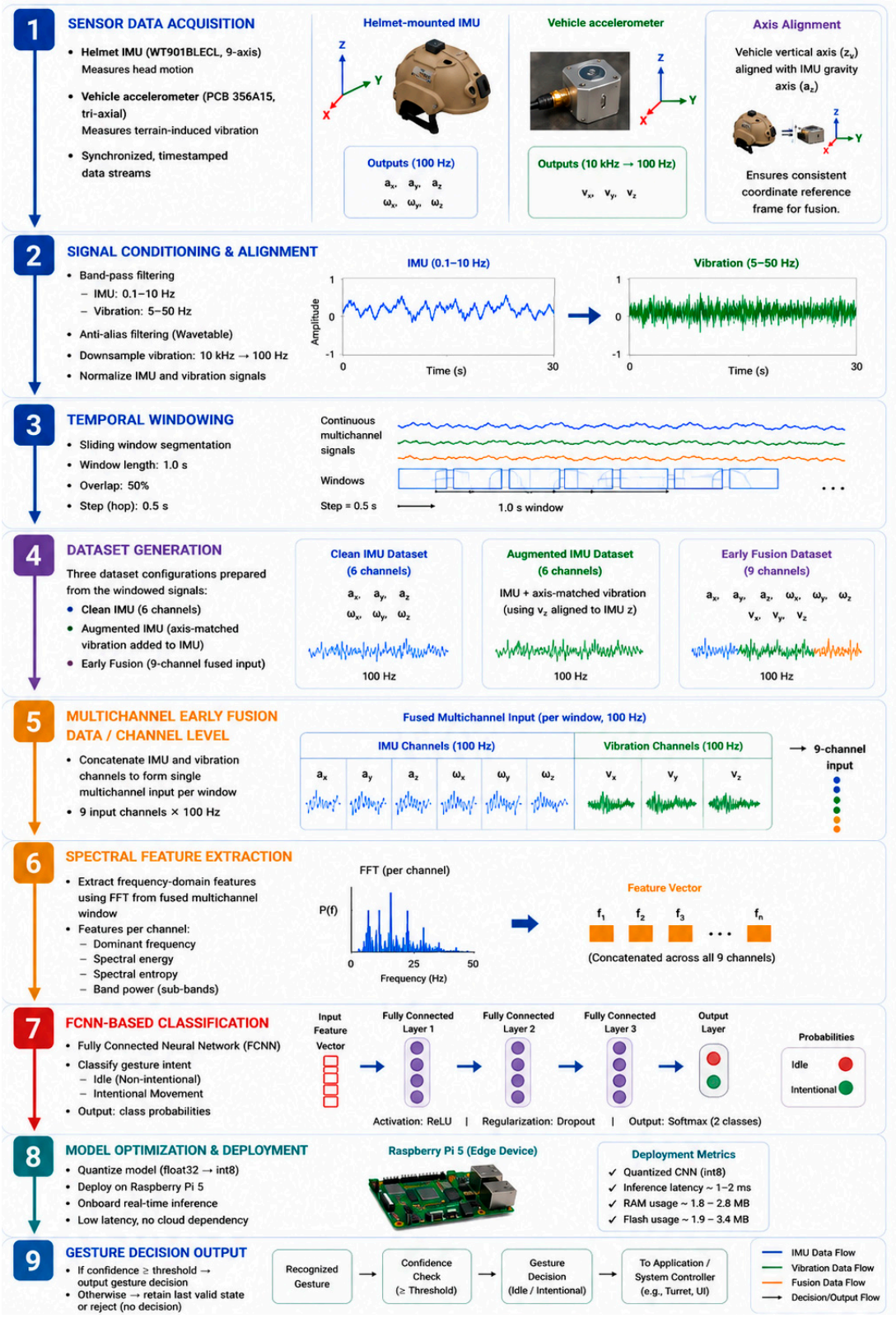

Figure 1. System architecture of the proposed vibration-aware head gesture recognition framework, showing sensor acquisition, signal preprocessing, feature extraction, feature fusion, and FCNN-based classification for robust gesture detection under terrain-induced vibration.

Unlike conventional gesture-recognition approaches that treat vibration solely as unwanted noise, the proposed framework incorporates terrain-induced vibration as contextual information within the classification process. This enables the classifier to distinguish between intentional operator motion and mechanically induced motion artefacts generated by vehicle–terrain interaction.

The operational sequence of the proposed framework is summarised as follows: sensor data streams are synchronised and segmented into fixed-length windows; features are extracted from each sensing modality; IMU and vibration features are concatenated into a fused representation; and the resulting feature vector is processed by a trained Fully Connected Neural Network (FCNN) to produce a gesture classification output.

2.2. Hardware Components

2.2.1. Head Gesture Sensing Unit

Head motion was measured using a WT901BLECL wireless 9-axis IMU (WitMotion Shenzhen Co., Ltd., Shenzhen, China) incorporating a tri-axial accelerometer, gyroscope, and magnetometer. The sensor was mounted centrally on the upper rear section of the operator helmet using hook-and-loop fastening to maintain a stable reference frame relative to head motion while minimising sensor displacement during operation. The physical sensor placement and mounting configuration are shown in Figure 2a–c.

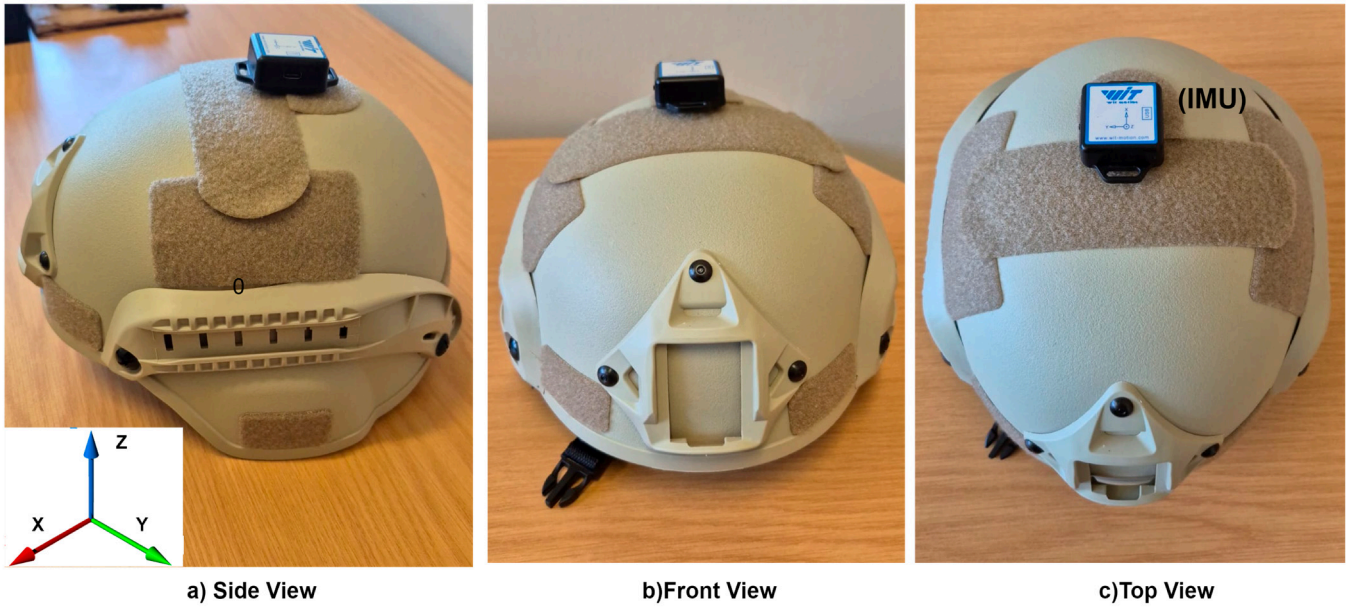

Figure 2. Helmet-mounted WT901BLECL IMU placement showing sensor positioning and coordinate alignment relative to the operator head frame.

The IMU coordinate frame was aligned such that the vertical gravitational axis corresponded to the sensor z-axis. This alignment was selected to maintain consistency with the vehicle-mounted vibration sensor and to simplify axis correspondence during feature fusion. Accelerometer and gyroscope signals were used during model development, while magnetometer measurements were excluded due to susceptibility to environmental magnetic disturbances and vehicle-induced electromagnetic interference.

The IMU was sampled at 100 Hz, selected based on the dominant biomechanical frequency range associated with voluntary human head motion, which typically occurs below 20 Hz. This sampling frequency satisfies Nyquist sampling requirements while maintaining computational efficiency for embedded inference.

The helmet-mounted IMU assembly forms part of the wearable human–machine interaction subsystem and provides direct measurement of intentional operator head gestures under both stationary and dynamically excited conditions.

2.2.2. Vibration Sensing Unit

Terrain-induced vibration was measured using a PCB Piezotronics Model 356A15 triaxial accelerometer mounted to the vehicle seat rack assembly near the operator compartment. The mounting location was selected to capture dominant structural vibration transmitted through the vehicle chassis and suspension system during terrain traversal. The accelerometer mounting configuration is illustrated in Figure 3.

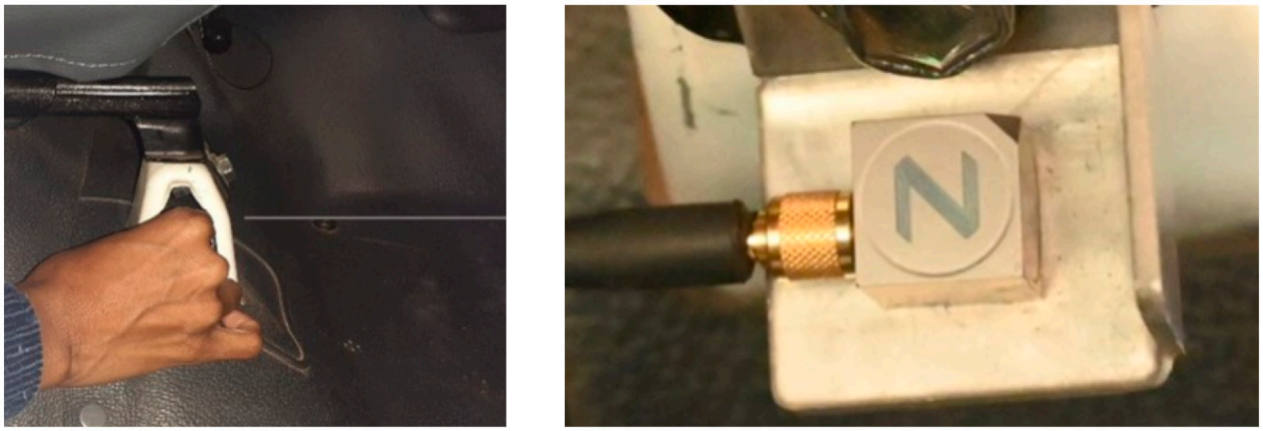

Figure 3. Terrain-vibration measurement setup: (**a**) seat-rack location and placement of the accelerometer sensor; (**b**) seat-rack-mounted accelerometer sensor used for terrain-vibration data acquisition.

The sensor measures acceleration in three orthogonal directions (x, y, z), with the vertical axis used primarily for terrain vibration analysis due to its strong correlation with suspension excitation and terrain roughness. The sensor was rigidly attached to the mounting bracket to ensure reliable mechanical coupling with the vehicle structure and to minimise measurement distortion caused by mounting compliance.

Raw vibration signals were initially sampled at 10 kHz to capture high-frequency mechanical excitation and transient vibration events generated by vehicle–terrain interaction. The acquired signals were subsequently anti-alias filtered and down sampled to 100 Hz to align temporally with the IMU sampling frequency and to reduce computational complexity for embedded processing.

The vehicle-mounted accelerometer forms the environmental context sensing subsystem and provides terrain-vibration information used for both vibration-aware data augmentation and multimodal feature fusion.

2.2.3. Processing Unit

The Raspberry Pi 5 was used as the target embedded processing platform for onboard implementation of the machine-learning pipeline. The processing pipeline comprised signal conditioning, sliding-window segmentation, early channel-level fusion of the IMU and terrain-vibration signals, FFT-based spectral processing, and FCNN inference.

The software architecture consisted of the following:

1. Data acquisition layer;
2. Signal conditioning module;
3. Sliding-window segmentation engine;
4. Early channel-level sensor fusion;

5. FFT-based spectral-analysis processing;
6. Lightweight FCNN inference engine;
7. Output control module.

The system architecture was designed for fully onboard operation without reliance on cloud connectivity, supporting local processing and inference for real-time embedded implementation.

2.3. Data Acquisition Protocol

The present study used an existing dataset generated during a previous internal company research and development project. The dataset was not obtained from a public database and has not previously been reported in a published study. The original project involved controlled head motion recordings performed by two healthy adult principal investigators wearing a helmet-mounted WT901BLECL inertial measurement unit (IMU), together with terrain-vibration measurements collected during operational vehicle field testing. Ethics approval (REF: 3699) was obtained from the Ethics Committee of College of Science, Engineering and Technology, School of Computing, UNISA. Informed consent for participation was obtained from all subjects involved in the study. The data were subsequently anonymised before being used for the present secondary machine-learning analysis. The head motion dataset consisted of two behavioural classes: Idle and Intentional Movement.

Intentional movements comprised five gesture subtypes: Yaw Left, Yaw Right, Pitch Up, Pitch Down, and Mode Nod. For each gesture subtype, each operator performed 20 repetitions under each recording condition. During each repetition, the operator performed the gesture, maintained the final head position for approximately 0.5 s, returned to the neutral position, and rested for 2–5 s before commencing the next repetition. Idle data were obtained by recording each operator maintaining a neutral head position for three continuous 60 s segments under each recording condition.

Terrain-vibration data were collected independently using a PCB 356A15 triaxial accelerometer (PCB Piezotronics, Depew, NY, USA) mounted to the vehicle seat-support structure while a patrol vehicle traversed the Ndumo Border Patrol route in South Africa. For each terrain type, approximately 20 min of continuous vibration data were recorded and subsequently annotated according to the corresponding terrain segment.

The original data-generation activity included both laboratory-controlled recordings and operationally representative vehicle testing. Head motion measurements were acquired using a helmet-mounted WT901BLECL IMU, while terrain-vibration measurements were acquired independently using the vehicle-mounted accelerometer. The present study did not recruit additional participants or conduct new human-participant experiments. Instead, it performed secondary analysis, dataset construction, vibration-aware augmentation, and machine-learning model development using the existing anonymised dataset.

Head motion data were acquired using a helmet-mounted WT901BLECL 9-axis IMU positioned centrally on the top surface of the helmet to minimise rotational bias and maintain alignment with the operator head frame. The IMU recorded tri-axial accelerometer and gyroscope measurements at 100 Hz.

Terrain-vibration measurements were acquired using a PCB 356A15 triaxial seat-rack accelerometer mounted to the vehicle seat support structure. The accelerometer recorded vehicle-induced vibration at 10 kHz to preserve high-frequency terrain excitation components prior to resampling.

The original acquisition procedure consisted of three phases:

1. Static calibration phase (vehicle stationary, engine off)
2. Controlled gesture phase (vehicle stationary)
3. Dynamic operational phase (vehicle moving under terrain excitation)

Helmet IMU and vehicle accelerometer streams were timestamped to support temporal correspondence between operator motion and environmental vibration. To ensure physically meaningful sensor fusion, the vehicle vertical vibration axis was aligned with the IMU gravity axis prior to fusion. The vehicle vertical acceleration component v_z was therefore mapped to the helmet IMU a_z reference frame, ensuring a consistent coordinate representation between operator motion and terrain excitation.

To evaluate classifier robustness under previously unseen vibration conditions, dataset partitioning was performed using temporally independent recording segments rather than random sample-level splitting. This reduced temporal leakage between training and testing data and provided a more realistic evaluation of operational generalisation.

The dataset partition used throughout the study consisted approximately of the following:

- Training set: 85%;
- Independent testing set: 15%.

Independent testing data were constructed from temporally separated vibration segments not used during training. Conventional k-fold cross-validation was not employed because overlapping windows extracted from continuous time-series recordings are temporally correlated and may cause data leakage between training and testing sets. Instead, temporally independent recording segments were used to provide a more realistic evaluation of model generalisation under previously unseen terrain-vibration conditions.

Three dataset configurations were generated for comparative evaluation:

1. Clean IMU dataset;
2. Axis-matched vibration-augmented IMU dataset;
3. Early fusion multichannel IMU + terrain-vibration dataset.

2.4. Signal Preprocessing

Prior to feature extraction and classification, IMU and terrain-vibration signals underwent preprocessing to suppress sensor noise, remove drift, and ensure temporal compatibility between sensing modalities. Similar preprocessing strategies have been widely adopted in wearable inertial sensing and time-series classification systems to improve signal quality and enhance feature extraction performance [1,2].

A fourth-order Butterworth band-pass filter was applied to the IMU signals within the frequency range of 0.1–10 Hz to preserve voluntary head motion dynamics while suppressing low-frequency drift and high-frequency measurement noise. Terrain-vibration signals were filtered within the 5–50 Hz range to preserve dominant terrain-induced excitation associated with suspension dynamics and structural vibration.

Because terrain-vibration measurements were originally sampled at 10 kHz, anti-alias filtering was applied prior to resampling. The vibration signals were subsequently down sampled to 100 Hz to match the IMU sampling frequency and enable synchronous multichannel processing. Synchronisation of heterogeneous sensing modalities is a common requirement in multimodal sensing systems and enables meaningful temporal correlation between sensor channels [5].

Following filtering and resampling, the signals were segmented using a sliding-window approach. Preliminary experiments were conducted to determine an appropriate segmentation strategy for the gesture-recognition task. Window lengths of 1 s and 2 s, combined with overlap values of 25% and 50%, were evaluated during model development. The 1 s window with 50% overlap produced the highest classification performance while maintaining sufficient temporal resolution and computational efficiency for embedded implementation. Larger windows increased computational cost and response latency, whereas lower overlap values reduced classification stability. Similar window durations have been successfully employed in wearable inertial sensing and human activity recognition applica-

tions, where 1 s windows provide an effective balance between recognition performance and real-time responsiveness [12,13].

The selected segmentation parameters were as follows:

- Window length: 1 s;
- Window overlap: 50%;
- Window increment (hop size): 0.5 s.

Each segmented window was treated as an independent classification sample. Window segmentation can be expressed as:

$$X_k = \{x(t) \mid t \in [k\Delta,\ k\Delta + W]\} \tag{1}$$

where

- W represents the window length.
- Δ represents the step size.
- k denotes the window index.

The sliding-window approach enabled continuous temporal analysis while improving classification responsiveness and temporal resolution for real-time deployment.

For the vibration-augmented dataset, axis-matched terrain-vibration signals were superimposed onto the IMU accelerometer channels to simulate mechanically disturbed operating conditions. This augmentation process introduced realistic terrain-induced perturbations into otherwise clean head motion recordings while preserving the underlying intentional gesture structure.

The augmentation process can be expressed as:

$$a_{aug}(t) = a_{imu}(t) + \alpha v(t) \tag{2}$$

where

- $a_{imu}(t)$ represents the original IMU signal.
- $v(t)$ represents the aligned terrain-vibration signal.
- α represents the vibration scaling factor.

Data augmentation has been shown to improve classifier robustness by exposing learning algorithms to a broader range of operating conditions and measurement disturbances during training [14,15]. Multiple vibration scaling levels were evaluated to investigate classifier robustness under varying terrain excitation intensities.

Four vibration scaling factors (0.1, 0.2, 0.3 and 0.4) were initially evaluated to represent increasing levels of terrain-induced excitation. The final models were trained using the 0.2 scaling factor because it provided a balanced augmentation level that introduced measurable vibration disturbance while preserving the underlying intentional head motion structure. Lower scaling (0.1) produced limited perturbation, whereas higher scaling levels (0.3 and 0.4) introduced stronger disturbances that risked masking the original gesture dynamics. The 0.2 scale was therefore selected as a representative moderate terrain-excitation condition for the final comparative experiments.

2.5. Feature Extraction

Spectral analysis was employed to extract discriminative motion characteristics from segmented signal windows. Frequency-domain representations are commonly employed in inertial-sensing applications because they capture periodic and transient motion characteristics that may not be readily distinguishable in the time domain [3]. This approach was

selected because intentional head gestures and terrain-induced vibration exhibit distinct temporal-frequency behaviour.

For each segmented window, spectral features were extracted from the IMU and vibration channels using Edge Impulse spectral analysis blocks. The extracted representation characterised the dominant spectral components, total spectral energy, spectral entropy, and the distribution of signal power across frequency sub-bands. These features are described below [16,17].

The discrete Fourier transform of a segmented signal $x[n]$ of length N can be expressed as:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi kn/N} \tag{3}$$

where $X[k]$ is the complex spectral coefficient at frequency bin k, $x[n]$ is the sampled signal within the analysis window, and N is the number of samples in the window.

The **dominant frequency** represents the frequency associated with the largest spectral magnitude within the analysed window and was determined as:

$$f_d = f_{k^*}, k^* = \arg\max_k \mid X[k] \mid \tag{4}$$

where f_d is the dominant frequency and f_{k^*} is the frequency corresponding to the spectral bin with the maximum magnitude. This feature identifies the strongest oscillatory component present within the analysed motion segment.

The **spectral energy** characterises the overall energy content of the signal in the frequency domain and was computed from the squared spectral magnitude:

$$E = \sum_{k=0}^{N-1} \mid X[k] \mid^2 \tag{5}$$

where E represents the total spectral energy. Higher values indicate stronger motion or vibration activity within the analysed window.

The **spectral entropy** describes how signal energy is distributed across the frequency spectrum. The normalized spectral power distribution was defined as:

$$p_k = \frac{\mid X[k] \mid^2}{\sum_{i=0}^{N-1} \mid X[i] \mid^2} \tag{6}$$

and spectral entropy was calculated as:

$$H = -\sum_{k=0}^{N-1} p_k \log_2(p_k) \tag{7}$$

where H represents the spectral entropy. Lower entropy indicates that the spectral energy is concentrated within a smaller number of frequency components, whereas higher entropy indicates a broader or more distributed spectral-energy pattern.

The **band-power distribution** describes the proportion of spectral energy contained within predefined frequency sub-bands. For a frequency interval bounded by f_l and f_u, the corresponding band power was calculated as:

$$P_b = \sum_{k: f_l \le f_k \le f_u} \mid X[k] \mid^2 \tag{8}$$

Collectively, these spectral descriptors provide a compact representation of the frequency composition, energy magnitude, spectral concentration, and band-wise energy

distribution of each segmented signal window. In the context of this study, deliberate head movements were expected to exhibit structured motion-related spectral patterns, whereas terrain-induced excitation introduced additional vibration components and broader spectral variability. The extracted spectral representations were therefore used as inputs to the FCNN classifier.

Three dataset configurations were investigated:

1. Clean IMU Dataset

This dataset contained only helmet-mounted IMU channels:

$$\left[a_x, a_y, a_z, g_x, g_y, g_z\right] \tag{9}$$

and served as the baseline gesture-recognition model.

2. Axis-Matched Vibration-Augmented IMU Dataset

This dataset was generated by superimposing aligned terrain-vibration signals onto the IMU accelerometer channels to simulate mechanically disturbed operating conditions. The augmented dataset preserved the original six-channel IMU structure while embedding terrain-induced perturbations.

3. Early-Fusion Multichannel Dataset

For the early-fusion configuration, IMU and terrain-vibration signals were combined at the channel level before feature extraction. Each segmented window therefore contained both IMU and terrain-vibration channels:

$$X_{fusion}(t) = \left[a_x, a_y, a_z, g_x, g_y, g_z, v_x, v_y, v_z\right] \tag{10}$$

where a_x, a_y, and a_z represent the tri-axial IMU acceleration signals; g_x, g_y, and g_z represent the tri-axial angular-velocity signals; and v_x, v_y, and v_z represent the corresponding tri-axial terrain-vibration signals. This produced a nine-channel multichannel representation sampled synchronously at 100 Hz.

Unlike feature-level fusion approaches, the proposed method performed early channel-level fusion before spectral feature extraction. Equation (5) represents the channel-level early fusion process. The IMU and terrain-vibration channels were first combined into a unified multichannel signal representation before feature extraction. Spectral features were subsequently extracted from the fused multichannel input according to:

$$F_{fusion} = \phi\left(X_{fusion}\right) \tag{11}$$

where

- X_{fusion} represents the multichannel fused input window.
- $\phi(\cdot)$ represents the spectral feature extraction operation.

Early channel-level fusion was selected because the IMU and terrain-vibration measurements are physically and temporally coupled. Terrain excitation transmitted through the vehicle and operator contributes to the motion observed by the helmet-mounted IMU. Early fusion therefore preserves the relationship between operator motion and environmental excitation within the same observation window, whereas late fusion processes each modality independently and combines their outputs at the decision stage [18].

In the present application, terrain vibration provides contextual information rather than an independent prediction of operator intent. The synchronized IMU and terrain-vibration channels were therefore combined before spectral transformation and FCNN classification. This enabled the classifier to learn nonlinear relationships from spectral

representations derived from both operator head motion and environmental excitation. Late fusion was not experimentally evaluated in the present study and remains an area for future comparative investigation.

Because terrain-induced vibration physically influences inertial measurements during mobile operation, treating vibration as contextual information rather than independent noise enabled improved interpretation of intentional operator movement under mechanically dynamic conditions. This perspective differs from conventional vibration-mitigation approaches, which typically seek to suppress environmental excitation through filtering or isolation. Instead, the proposed framework leverages vibration measurements as an additional source of contextual information describing the operating environment [6,7].

2.6. Machine Learning Model

A lightweight fully connected neural network (FCNN) was employed to classify the spectral representations extracted from the segmented sensor signals [19]. Rather than operating directly on raw time-series data, the classifier received a 222-element feature vector generated by the FFT-based spectral-analysis processing block for each segmented window. The spectral-analysis stage and the FCNN therefore performed complementary functions: the spectral processing stage transformed the windowed sensor signals into a compact frequency-domain representation, while the FCNN learned nonlinear decision boundaries for distinguishing between the Idle and Intentional Movement classes.

The FFT–FCNN processing architecture was selected because the discrimination problem investigated in this study is associated with differences in the spectral characteristics of intentional head motion and terrain-induced vibration. Intentional head movements primarily produce structured low-frequency motion patterns, whereas terrain excitation introduces vibration components with different spectral distributions. Transforming the segmented signals into the frequency domain therefore provides a physically interpretable representation of these differences before classification.

Compared with an end-to-end architecture operating directly on raw multichannel time-series samples, the adopted approach reduces the requirement for the classifier itself to learn frequency-selective transformations directly from the raw signals. This enables the use of a compact FCNN architecture with relatively few trainable parameters, supporting low computational complexity and memory requirements for embedded deployment. Furthermore, the same spectral-processing and classification framework was applied consistently across the Clean IMU, vibration-augmented IMU, and early-fusion multichannel configurations, allowing the influence of the different vibration-aware data representations to be evaluated using a consistent classification architecture.

The FCNN consisted of two fully connected hidden layers containing 20 and 10 neurons, respectively. Both hidden layers employed the rectified linear unit (ReLU) activation function and L1 activity regularisation with a coefficient of 1×10^{-5}. The output layer contained two neurons corresponding to the Idle and Intentional Movement classes and employed a Softmax activation function to generate class probabilities. The network architecture is summarised as follows:

$$222 \text{ input features} \rightarrow \text{Dense (20, ReLU)} \rightarrow \text{Dense (10, ReLU)} \rightarrow \text{Dense (2, Softmax)}$$

The model was trained for 30 epochs using a learning rate of 0.0005 and a batch size of 32. Categorical cross-entropy was used as the loss function, and 20% of the training data was reserved for validation during model development. The model architecture and training parameters are summarised in Table 1.

Table 1. FCNN architecture and training configuration.

Parameter	Configuration
Model architecture	Fully connected neural network (FCNN)
Input	222 spectral features
Hidden layer 1	Dense, 20 neurons
Activation function	ReLU
Hidden layer 2	Dense, 10 neurons
Activation function	ReLU
Activity regularisation	L1, coefficient = (1×10^{-5})
Output layer	Dense, 2 neurons
Output activation	Softmax
Loss function	Categorical cross-entropy
Training epochs	30
Learning rate	0.0005
Batch size	32
Validation split	20%
Deployment profiling	8-bit integer (int8) quantisation

The compact FCNN architecture was selected to provide sufficient nonlinear classification capacity while maintaining low computational and memory requirements for embedded deployment. Because the preceding spectral-analysis stage transformed each signal window into a structured frequency-domain representation, the FCNN operated on spectral representations rather than learning temporal and frequency-selective transformations directly from raw sensor sequences. This separation between signal representation and classification reduced the complexity required at the classifier stage and was considered appropriate for real-time gesture recognition under resource-constrained edge-computing conditions.

Quantization resilience was evaluated empirically rather than assumed to result directly from FFT-based preprocessing. The float32 and int8 implementations of each dataset configuration were therefore compared experimentally in terms of classification performance and deployment resource requirements. Consequently, the observed differences in quantization sensitivity are interpreted as outcomes associated with the different vibration-aware data representations and model configurations, rather than as an inherent property of FFT-based spectral preprocessing.

2.7. Evaluation Metrics

Deployment performance was evaluated using the profiling results generated by the Edge Impulse deployment pipeline and EON Compiler. The reported inference latency, peak RAM usage, and Flash requirements therefore represent deployment profiling estimates for the generated model implementation rather than independent system-level measurements obtained using external Raspberry Pi profiling tools. These metrics were used consistently across all three model configurations and both numerical precision formats to enable comparative assessment of computational footprint and quantization behaviour. The Raspberry Pi 5 served as the target edge deployment platform for system integration; however, the resource values reported in the comparative tables and figures correspond to the Edge Impulse model profiling outputs. System performance was evaluated using the following:

- Classification accuracy;
- Precision;
- Recall;
- F1-score;

- Confusion matrices;
- Embedded inference latency;
- RAM usage;
- Flash memory usage.

Comparative evaluations were conducted between the following:

1. Clean IMU baseline model;
2. Vibration-augmented model;
3. Early fusion multimodal model.

Performance evaluation included both float32 and quantized int8 implementations to assess embedded deployment suitability.

Classification accuracy was calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

The F1-score was calculated as:

$$F_1 = 2 \cdot \frac{Precision \times Recall}{Precision + Recall} \tag{13}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. The evaluation methodology emphasized both classification robustness and embedded deployment feasibility under mechanically dynamic operating conditions. In addition to classification performance, memory utilisation and inference latency were evaluated because resource efficiency is a critical requirement for embedded edge-AI systems operating in real-time environments [10].

3. Results

3.1. Overview of Experimental Evaluation

Three gesture-recognition configurations were evaluated to investigate the influence of terrain-induced vibration and multichannel contextual sensing on head gesture classification performance:

1. Clean IMU baseline model;
2. Axis-matched vibration-augmented IMU model;
3. Early fusion multichannel IMU + terrain-vibration model.

Each configuration was evaluated using both unoptimized float32 and quantized int8 neural-network deployments generated using the Edge Impulse EON compiler Performance evaluation considered classification accuracy, weighted F1-score, confusion-matrix behaviour, memory utilisation, Flash/ROM usage, inference latency, and quantization sensitivity.

The validation performance of the three model configurations under float32 and int8 inference is summarised in Table 2, while their performance on the independent test dataset is presented in Table 3. These evaluations enable comparison of both in-development model performance and generalisation to temporally independent data.

The corresponding embedded deployment requirements, including inference latency, RAM utilisation, and Flash/ROM requirements, are reported in Table 4. Quantization sensitivity is further quantified in Table 5 by comparing the independent-test accuracy of the float32 and int8 implementations and calculating the resulting accuracy reduction.

Table 2. Validation performance comparison.

Model	Precision Type	Accuracy (%)	Weighted F1	Idle Accuracy (%)	Intentional Accuracy (%)	Loss
Model 1: Clean IMU	Float32	99.9	1	99.9	99.9	0
	int8	88.9	0.88	88	88.6	1.82
Model 2: Augmented IMU	Float32	99.9	1.00	99.9	99.9	0.00
	int8	89.9	0.90	90.5	89.0	2.84
Model 3: Early Fusion	Float32	99.9	1.00	99.9	99.9	0.00
	int8	97.5	0.97	95.6	98.8	0.58

Table 3. Independent testing performance.

Model	Precision Type	Accuracy (%)	Weighted F1	Idle Accuracy (%)	Intentional Accuracy (%)	Uncertain (%)
Model 1: Clean IMU	Float32	94.87	0.95	91.8	99.93	4.0/1.0
	int8	74.74	0.8	61	94.7	9.2/3.2
Model 2: Augmented IMU	Float32	99.58	1.00	99.4	99.7	0.2/0.1
	int8	72.56	0.88	45.7	88.2	41.0/5.3
Model 3: Early Fusion	Float32	98.45	0.99	98.4	98.5	0.3/0.1
	int8	90.02	0.91	82.5	97.3	1.2/0.3

Table 4. Deployment requirements.

Model	Precision	Latency (ms)	RAM (KB)	Flash/ROM (KB)
Model 1	Float32	2	3.8	28.9
	int8	2	3.8	18.6
Model 2	Float32	2	3.8	28.9
	int8	2	3.8	18.6
Model 3	Float32	2	5.5	37.6
	int8	2	5.5	20.8

Table 5. Quantization sensitivity.

Model	Float32 Accuracy (%)	int8 Accuracy (%)	Accuracy Drop (%)
Model 1	94.87	74.74	20.13
Model 2	99.58	72.56	27.02
Model 3	98.45	90.02	8.43

Figure 4 provides a graphical comparison of the independent-test classification accuracy across the three model configurations under float32 and int8 inference. The associated confusion matrices are presented in Figure 5, providing class-level insight into the effects of vibration-aware learning, early fusion, and quantization on the Idle and Intentional Movement classes.

Overall, the experiments were designed to evaluate both recognition robustness and embedded deployment feasibility under resource-constrained edge-processing conditions.

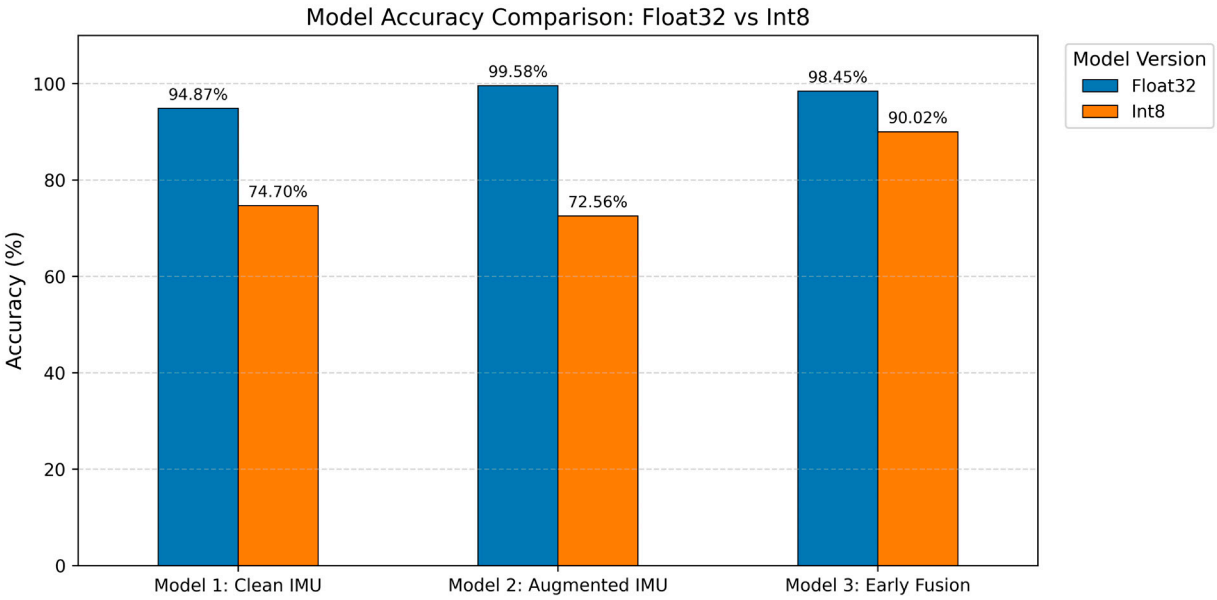

Figure 4. Classification accuracy comparison between the clean IMU baseline, vibration-augmented IMU model, and early fusion multimodal model under float32 and quantized int8 deployment configurations.

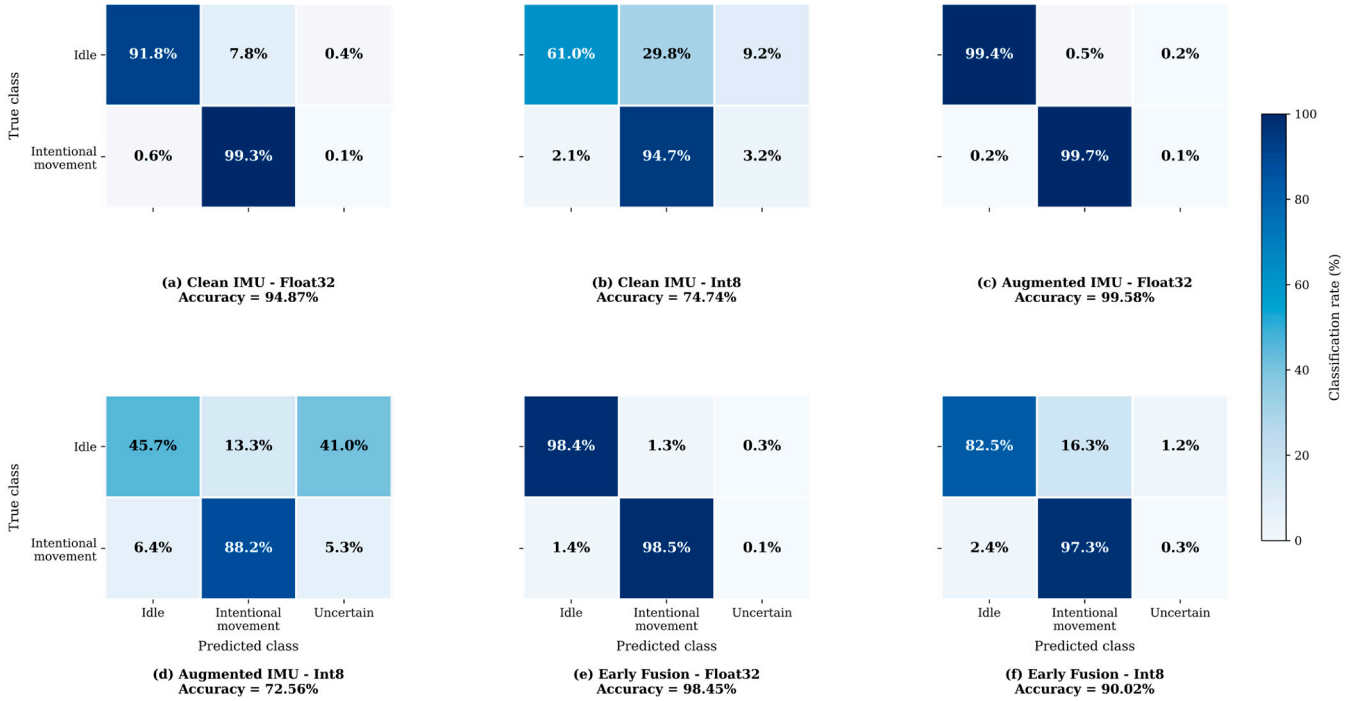

Figure 5. Confusion matrices for the evaluated head gesture recognition configurations: (**a**) clean IMU Float32, (**b**) clean IMU Int8, (**c**) vibration-augmented IMU Float32, (**d**) vibration-augmented IMU Int8, (**e**) multichannel early-fusion Float32, and (**f**) multichannel early-fusion Int8.

3.2. Clean IMU Baseline Performance

The clean IMU baseline model was trained using only helmet-mounted IMU accelerometer and gyroscope channels without terrain-vibration augmentation or sensor fusion. This configuration represented the conventional gesture-recognition approach commonly used in wearable HMI systems [3].

The unoptimized float32 implementation achieved a validation accuracy of 99.9%, with near-perfect class separation between idle and intentional movement classes. The confusion

matrix showed minimal cross-class misclassification, and the feature-space visualization demonstrated highly separable clusters between the two motion categories.

However, after int8 quantization, classification performance degraded substantially. The quantized model achieved 74.74% validation accuracy, with significant increases in idle-state misclassification and uncertain predictions. The quantized feature-space representation showed strong class overlap and reduced cluster separability.

This behaviour indicates that the clean IMU model relied heavily on subtle high-resolution feature representations that were sensitive to reduced numerical precision during quantization. Although the float32 model demonstrated excellent classification capability under controlled conditions, the quantized deployment results suggest reduced robustness for embedded edge deployment.

3.3. Axis-Matched Vibration-Augmented IMU Performance

The vibration-augmented dataset was generated by superimposing axis-aligned terrain-vibration signals onto the IMU accelerometer channels to simulate realistic mechanically disturbed operating conditions [14,15]. This approach preserved the original six-channel IMU structure while exposing the classifier to terrain-induced perturbations during training.

The float32 augmented model achieved 99.9% validation accuracy with complete class separation in the feature-space visualization. Compared with the clean baseline model, the augmented dataset produced broader feature distributions due to the inclusion of terrain-induced perturbations; however, class separability remained strong.

Importantly, the augmented int8 model achieved 88.3% validation accuracy, representing a substantial improvement over the clean quantized baseline. Misclassification rates decreased significantly, and the feature-space distributions showed improved robustness under quantized deployment.

These results indicate that vibration-aware augmentation improved the model's tolerance to quantization effects by exposing the classifier to broader operational variability during training. The augmentation process therefore acted as both a robustness-enhancement mechanism and a regularization strategy for embedded deployment.

3.4. Early-Fusion Multichannel Model Performance

The early-fusion configuration combined synchronized IMU and terrain-vibration channels into a unified nine-channel multichannel representation before spectral feature extraction. This enabled the classifier to learn contextual relationships between intentional head motion and environmental excitation directly from the fused signal representation. Similar multimodal fusion approaches have been shown to improve classification robustness by exploiting complementary information from multiple sensing modalities [1,7].

The float32 early fusion model achieved 94.87% validation accuracy with a weighted F1-score of 0.95. Compared with the previous models, the feature-space distributions exhibited increased overlap between idle and intentional classes, indicating the increased complexity introduced by multichannel contextual sensing.

The quantized int8 implementation achieved 90.02% validation accuracy, outperforming both the clean and augmented quantized models. In contrast to the severe degradation observed in the clean baseline model after quantization, the early-fusion architecture maintained relatively stable performance under reduced numerical precision.

This result suggests that the multichannel fusion strategy produced more robust embedded representations that generalized better under quantized inference conditions. Although the float32 accuracy was lower than the previous models, the early-fusion model demonstrated the best overall balance between classification robustness and deployment resilience.

The feature-space visualization further showed that the early-fusion model maintained structured class boundaries despite increased signal complexity. The model therefore appears to learn contextual vibration-aware representations rather than relying solely on isolated gesture dynamics.

3.5. Comparative Analysis of Model Architectures

The comparative results demonstrate that vibration-aware learning strategies substantially improve embedded robustness under quantized deployment conditions. The quantization-induced accuracy degradation across the three model configurations is illustrated in Figure 6.

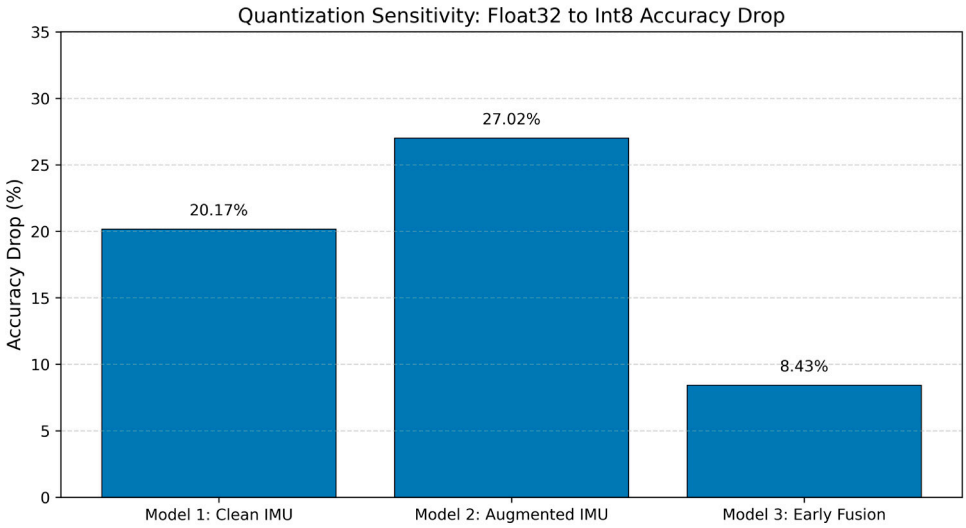

Figure 6. Accuracy degradation observed after int8 quantization for the three investigated model configurations.

The clean IMU baseline achieved the highest float32 accuracy; however, it also exhibited the largest quantization-induced degradation, decreasing from 94.87% to 74.74%. This indicates that the baseline model over-relied on highly precise feature representations that were not robust to reduced numerical precision.

In contrast, both vibration-aware approaches demonstrated improved quantization resilience:

- Augmented IMU model: 99.58% $\rightarrow$ 72.56%;
- Early fusion model: 98.45% $\rightarrow$ 90.02%.

The early-fusion model exhibited the smallest performance degradation after quantization, suggesting that contextual vibration information improved feature robustness under embedded deployment constraints. The results further indicate that vibration-aware learning improves deployment robustness more effectively than purely increasing classification accuracy under ideal conditions. From an embedded systems perspective, the early-fusion model therefore provides the most balanced operational solution.

3.6. Statistical Analysis of Classification Outcomes

A 3 $\times$ 2 Pearson chi-square test of independence was conducted to determine whether classification outcome was associated with model configuration [20]. The analysis compared the numbers of correct and incorrect classifications obtained from the Float32 test results of the Clean IMU, Vibration-Augmented IMU, and Early-Fusion configurations. The analysis revealed a statistically significant association between model configuration and classification outcome, $\chi^2(2) = 3361.39$, $p < 0.001$, with a Cramér's V of 0.134 [20].

The Clean IMU configuration achieved an accuracy of 94.87%, compared with 99.58% for the Vibration-Augmented IMU configuration and 98.45% for the Early-Fusion configuration. These findings indicate that the incorporation of terrain-vibration information was associated with improved classification outcomes relative to the Clean IMU baseline.

Although the Vibration-Augmented IMU configuration achieved the highest Float32 accuracy, the Early-Fusion configuration demonstrated greater stability under int8 quantization. This distinction indicates that maximum floating-point classification accuracy does not necessarily correspond to the strongest embedded deployment robustness. Because consecutive samples were generated using 1 s windows with 50% overlap, some temporal dependence may exist between adjacent classification windows. The statistical results are therefore interpreted as an aggregate comparison of classification outcomes across the evaluated model configurations.

3.7. Embedded Deployment Performance

All evaluated models achieved low-latency edge inference suitable for real-time operation on embedded platforms. Inference latency remained between 1–2 ms across all configurations, demonstrating that lightweight FCNN architectures can support vibration-aware gesture recognition low computational requirements.

The RAM and Flash/ROM requirements of the float32 and int8 implementations are compared in Figure 7. Memory utilisation also remained low, with RAM requirements below 4 KB and flash usage below 30 KB for all quantized models. These results confirm the feasibility of deployment on resource-constrained embedded systems such as the Raspberry Pi 5 and potentially lower-power TinyML platforms.

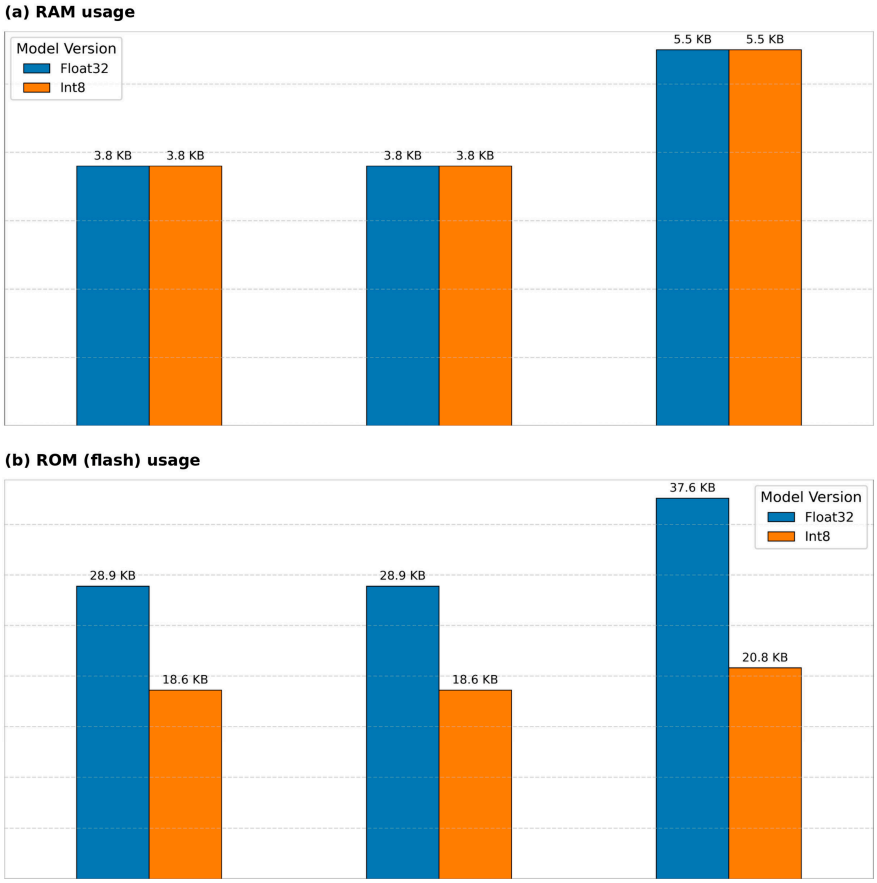

Figure 7. (**a**) Peak RAM usage comparison for float32 and quantized int8 implementations across the evaluated models; (**b**) flash memory usage comparison for float32 and quantized int8 model implementations.

The relationship between classification accuracy and memory requirements is further illustrated in Figure 8, providing a deployment-oriented comparison of predictive performance and computational resource utilisation. The results demonstrate that model selection involves a trade-off between classification performance and embedded resource requirements, particularly when additional terrain-vibration channels are incorporated.

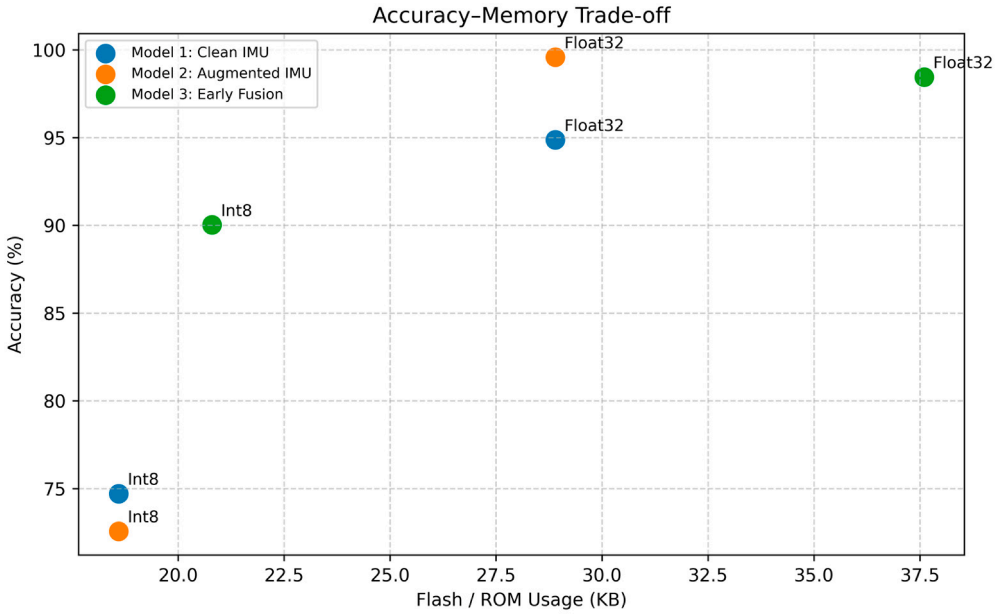

Figure 8. Accuracy versus memory trade-off analysis for the investigated gesture-recognition models under float32 and quantized int8 deployment conditions.

The compact deployment footprint further demonstrates that contextual vibration-aware sensing can be achieved without substantial computational overhead, making the proposed framework suitable for real-time wearable HMI applications operating in mechanically dynamic environments.

4. Discussion

The results demonstrate that terrain-induced vibration influences inertial gesture-recognition performance and that vibration-aware learning strategies can improve robustness under embedded deployment constraints. Rather than treating environmental excitation solely as measurement noise, the findings indicate that terrain vibration can be incorporated into the learning process either as a controlled disturbance through data augmentation or as an independent contextual sensing modality through early fusion. These approaches produced different effects on classification performance, feature-space structure, and quantization robustness.

The clean IMU baseline model demonstrated excellent separability between intentional and idle motion classes under high-precision inference conditions. The corresponding feature-space distributions showed tightly clustered class boundaries, indicating that the extracted spectral features were highly discriminative for controlled gesture recognition tasks. These findings are consistent with previous wearable IMU gesture-recognition studies that report strong classification performance under relatively stable sensing conditions [2–4].

However, the baseline model also exhibited the greatest sensitivity to quantization. This suggests that the classifier relied heavily on highly precision-sensitive feature representations that became unstable after numerical compression during embedded deployment. Similar behaviour has been reported in TinyML systems, where quantization can reduce

feature fidelity and alter decision boundaries when models rely heavily on high-precision feature representations [10,11,21].

From an edge-AI perspective, this behaviour is important because it highlights a common limitation of high-performing floating-point models: excellent laboratory accuracy does not necessarily translate into robust low-precision embedded inference. The results therefore emphasize the importance of evaluating deployment-oriented behaviour rather than relying exclusively on float32 validation performance.

The vibration-augmented approach improved robustness by exposing the classifier to mechanically disturbed motion patterns during training. By superimposing terrain-induced vibration onto the IMU channels, the resulting dataset contained broader operational variability and reduced dependence on tightly constrained feature boundaries. The augmented model consequently demonstrated improved tolerance to quantized inference conditions and reduced sensitivity to environmental perturbation.

This behaviour indicates that augmentation acted as a robustness-enhancement mechanism rather than simply a data-expansion strategy [15]. Exposure to vibration-disturbed motion patterns forced the network to learn more generalized representations that remained stable under reduced numerical precision. These findings align with broader embedded machine-learning research, where controlled perturbation during training often improves deployment resilience and generalization under operational uncertainty.

The early-fusion multichannel architecture demonstrated the strongest overall embedded robustness. This result aligns with multimodal sensing literature, where complementary sensor modalities often improve robustness and contextual interpretation compared with unimodal approaches [5,22]. Unlike the augmentation strategy, which injected vibration directly into the IMU measurements, the fusion architecture preserved terrain vibration as an independent sensing modality. This enabled the classifier to learn contextual relationships between operator motion and environmental excitation rather than interpreting vibration solely as signal corruption.

Although the fusion model produced broader feature-space distributions and increased overlap between motion classes, this behaviour is expected in multichannel contextual sensing systems where environmental variability becomes part of the learned representation. Importantly, the fusion architecture demonstrated substantially improved stability after quantization, indicating that the learned features were less dependent on high numerical precision.

These findings suggest that contextual vibration sensing improves the physical interpretability of the learned representations. Rather than relying exclusively on isolated head motion dynamics [6,7], the classifier could incorporate environmental excitation patterns when evaluating whether motion originated from intentional operator behaviour or terrain-induced disturbance. This supports the hypothesis that mechanically dynamic environments should be treated as part of the sensing context rather than purely as external interference to be suppressed.

An important outcome of the study is that the highest floating-point classification performance did not correspond to the most robust embedded deployment behaviour. The clean IMU model achieved the strongest apparent laboratory performance but exhibited the largest degradation after quantization. In contrast, the fusion architecture produced slightly lower floating-point accuracy while maintaining substantially greater stability under embedded inference conditions.

This observation highlights a critical consideration for TinyML and edge-AI systems operating in real-world environments [10]. In practical deployment scenarios, robustness to quantization, environmental variability, and sensor perturbation may be more operationally valuable than maximizing floating-point validation accuracy under controlled conditions.

The results therefore reinforce the importance of deployment-aware model evaluation for embedded sensing applications.

The low computational requirements observed across all evaluated models further support the feasibility of real-time embedded implementation [23,24]. Lightweight FCNN architectures achieved low inference latency and modest memory utilisation while maintaining robust classification capability under mechanically dynamic conditions. These findings demonstrate that contextual vibration-aware gesture recognition can be implemented on resource-constrained embedded hardware without requiring cloud-based processing or computationally intensive architectures.

The use of temporally independent dataset partitioning also strengthened the reliability of the evaluation methodology. Training and testing subsets were separated using independent recording segments rather than random sample-level partitioning, reducing the likelihood of temporal leakage between datasets. This is particularly important in time-series classification problems involving overlapping sliding windows, where random partitioning can artificially inflate performance estimates. The observed results therefore provide a more realistic indication of generalization under unseen terrain conditions.

Research Questions Revisited

The experimental findings provide direct answers to the three research questions formulated in the Introduction.

RQ1: Can intentional head gestures be reliably distinguished from terrain-induced motion artefacts under mobile operating conditions?
The results indicate that intentional and idle motion states can be distinguished under vibration-aware evaluation conditions. The clean IMU baseline established the underlying discriminability of the gesture classes, while the augmented and early-fusion configurations demonstrated that classification performance could be maintained with improved robustness when terrain-vibration information was incorporated into the learning process. The findings therefore support H1, showing that terrain-induced vibration affects gesture-recognition behaviour and should be explicitly considered when developing IMU-based interfaces for mechanically dynamic platforms.

RQ2: Does vibration-aware data augmentation improve gesture-recognition robustness?
The results indicate that vibration-aware augmentation improved robustness relative to the clean IMU baseline, particularly under quantized inference. The introduction of scaled, axis-matched terrain vibration increased the variability represented during model development and reduced the model's sensitivity to the transition from float32 to int8 inference. The findings therefore support H2, indicating that controlled vibration augmentation can improve the deployment robustness of IMU-based gesture-recognition models.

RQ3: Does early multichannel fusion improve embedded deployment robustness compared with conventional IMU-only approaches?
The early-fusion configuration demonstrated the strongest stability under quantized inference among the evaluated model configurations. By preserving terrain vibration as a separate contextual sensing modality and combining it with the IMU channels before spectral feature extraction, the approach produced a representation that was more resilient to reduced numerical precision. These findings support H3 and indicate that early multichannel fusion provides a practical strategy for improving embedded deployment robustness in mechanically dynamic environments.

Collectively, the answers to the research questions indicate that the principal benefit of vibration-aware modelling is not limited to maximizing floating-point classification accuracy. Its primary value lies in improving robustness to operational variability and reduced-

precision deployment. This distinction is important for wearable HMI systems intended for real-world mobile platforms, where environmental excitation and embedded hardware constraints must be considered jointly during model development and evaluation.

Several limitations should nevertheless be acknowledged. The study focused primarily on binary classification between intentional and idle motion states, and additional gesture vocabularies may introduce increased classification complexity. Furthermore, the terrain-vibration recordings were collected using a specific vehicle platform and operational route, meaning that vibration characteristics may vary under different vehicle dynamics, suspension systems, and terrain profiles. The present proof-of-concept study also utilised data collected from a limited number of healthy adult principal investigators; therefore, the generalisability of the proposed framework across a broader user population remains to be established. Although embedded quantization behaviour was evaluated extensively, long-duration operational field deployment remains an area for future validation.

Future work should evaluate the proposed framework using a larger and more diverse participant cohort to assess generalization across users. Additional investigations should explore cross-platform generalization, larger gesture taxonomies, adaptive vibration-aware learning strategies, and real-time deployment under extended operational conditions. Additional investigation into advanced multimodal fusion approaches, including attention-based architectures and adaptive contextual weighting mechanisms, may further improve robustness in complex dynamic environments.

Overall, the study demonstrates that vibration-aware contextual sensing substantially improves embedded gesture-recognition robustness under mechanically dynamic operating conditions. The findings further indicate that early-fusion multichannel sensing provides superior resilience under quantized deployment constraints, supporting the use of contextual environmental awareness in next-generation wearable human–machine interaction systems.

5. Conclusions

This study investigated terrain-aware head gesture recognition for mechanically dynamic mobile platforms using a vibration-informed embedded machine-learning framework. The work addressed the challenge of distinguishing intentional operator head gestures from non-intentional motion artefacts generated by terrain-induced vibration and structural excitation during vehicle operation.

A multimodal sensing framework was developed using a helmet-mounted inertial measurement unit (IMU) and a vehicle-mounted triaxial accelerometer. Real-world terrain-vibration data collected along the Ndumo Border Patrol route were incorporated into the recognition process using two complementary approaches: vibration-aware data augmentation and early multichannel sensor fusion. The proposed methods were evaluated using lightweight embedded neural-network architectures implemented within the Edge Impulse framework under both float32 and quantized int8 deployment conditions.

The results demonstrated that terrain-induced vibration substantially affects inertial gesture-recognition performance under embedded deployment constraints. Although the clean IMU baseline model achieved 99.9% accuracy under float32 inference, its performance decreased to 74.7% after int8 quantization, corresponding to a 25.2 percentage-point reduction and indicating substantial sensitivity to reduced numerical precision. In contrast, the vibration-aware approaches demonstrated improved robustness under quantized inference conditions. The vibration-augmented IMU model decreased from 99.9% under float32 inference to 88.3% under int8 inference, representing an 11.6 percentage-point reduction, while the early-fusion multichannel model decreased from approximately 94.9% to 90.0%, corresponding to a reduction of only 4.9 percentage points. These results demonstrate that

incorporating terrain-vibration information substantially improved robustness to reduced-precision inference, with the early-fusion configuration exhibiting the greatest quantization stability among the evaluated models.

The vibration-augmented model improved generalization by exposing the classifier to mechanically disturbed motion patterns during training, thereby reducing sensitivity to environmental perturbation and quantization effects. More importantly, the early-fusion multichannel architecture demonstrated the strongest overall deployment robustness by incorporating terrain vibration as contextual information rather than treating it solely as signal corruption. This enabled the classifier to learn relationships between operator motion and environmental excitation directly from synchronized multichannel sensor data.

An important finding of the study is that the highest floating-point accuracy did not correspond to the most robust embedded deployment behaviour. Instead, the contextual early-fusion architecture provided superior resilience under low-precision inference conditions, highlighting the importance of deployment-aware evaluation for embedded AI systems operating in real-world environments.

The study further demonstrated that vibration-aware gesture recognition can be implemented using lightweight embedded neural networks with low latency and modest memory requirements suitable for real-time edge deployment. These findings support the feasibility of terrain-aware wearable human–machine interaction systems for mobile and vibration-intensive operational environments.

From a broader perspective, the work demonstrates that environmental vibration should not necessarily be treated purely as measurement noise in inertial sensing systems. Instead, mechanically induced excitation can provide valuable contextual information that improves interpretation of operator behaviour under dynamic operating conditions. This concept has potential relevance beyond gesture recognition and may benefit future embedded sensing systems operating in robotics, wearable computing, vehicle autonomy, and human–machine interaction applications.

Future work should investigate larger gesture vocabularies, adaptive multimodal fusion strategies, cross-platform terrain generalization, and long-duration operational deployment under diverse environmental conditions. Additional investigation into attention-based fusion mechanisms and adaptive contextual weighting may further improve robustness in increasingly complex dynamic environments.

Overall, the results demonstrate that contextual vibration-aware sensing substantially improves the robustness and deployability of IMU-based gesture recognition systems under mechanically dynamic operating conditions. The proposed framework therefore provides a promising foundation for next-generation embedded human–machine interaction systems capable of operating reliably in real-world mobile environments.

Author Contributions: Conceptualization, L.R.; methodology, L.R.; software, L.R.; validation, L.R.; formal analysis, L.R.; investigation, L.R.; resources, L.R.; data curation, L.R. and D.M.; writing—original draft preparation, L.R.; writing—review and editing, L.R. and D.M.; visualization, L.R.; supervision, L.R., D.M., M.S. and T.P. project administration, L.R. and D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of College of Science, Engineering and Technology, School of Computing, UNISA (Reference number 3699) on 14 June 2024.

Informed Consent Statement: Informed consent for participation was obtained from all subjects involved in the study.

Data Availability Statement: The datasets analysed during the current study are available from the corresponding author upon reasonable request.

Acknowledgments: The author would like to acknowledge the Council for Scientific and Industrial Research (CSIR), Defence and Security, Landward Science, for providing the research environment, equipment, and technical support required for this study. The author also acknowledges the Edge Impulse platform for providing the framework used in this research. During the preparation of this manuscript, OpenAI ChatGPT (5.5) was used for language refinement and editing block diagrams. The author reviewed and edited all generated content and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
BLE	Bluetooth Low Energy
CNN	Fully Connected Neural Network
DSP	Digital Signal Processing
Edge AI	Edge Artificial Intelligence
F1-Score	Harmonic Mean of Precision and Recall
HMI	Human–Machine Interaction
Hz	Hertz
IMU	Inertial Measurement Unit
Int8	8-Bit Integer Quantization
IoT	Internet of Things
MEMS	Micro-Electro-Mechanical Systems
MCA	Mathematical and Computational Applications
ML	Machine Learning
RMS	Root Mean Square
RAM	Random Access Memory
ROM	Read-Only Memory
PCB	Printed Circuit Board
PCB 356A15	PCB Piezotronics 356A15 Triaxial Accelerometer
TinyML	Tiny Machine Learning
WT901BLECL	Bluetooth-Enabled 9-Axis Inertial Measurement Unit
HARA	Hazard Analysis and Risk Assessment
HMI	Human–Machine Interface
Float32	32-Bit Floating-Point Representation
Int8 Model	Quantized 8-Bit Integer Neural Network Model
EON	Embedded Optimization Neural Network (Edge Impulse EON Compiler)
HAR	Human Activity Recognition
DTW	Dynamic Time Warping
MCU	Microcontroller Unit
TRL	Technology Readiness Level
HCI	Human–Computer Interaction
FFT	Fast Fourier Transform

References

1. Song, Z.; Cao, Z.; Li, Z.; Wang, J.; Liu, Y. Inertial Motion Tracking on Mobile and Wearable Devices: Recent Advancements and Challenges. *Tsinghua Sci. Technol.* **2021**, *26*, 692–705. [CrossRef]
2. Al-Azzawi, S.S.; Khaksar, S.; Hadi, E.K.; Agrawal, H.; Murray, I. Headup: A low-cost solution for tracking head movement of children with cerebral palsy using imu. *Sensors* **2021**, *21*, 8148. [CrossRef] [PubMed]

3. Li, H.; Hu, H. Head Gesture Recognition Combining Activity Detection and Dynamic Time Warping. *J. Imaging* **2024**, *10*, 123. [CrossRef] [PubMed]
4. Rahmaniar, W.; Ma'arif, A.; Lin, T.-L. Touchless Head-Control (THC): Head Gesture Recognition for Cursor and Orientation Control. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2022**, *30*, 1817–1828. [CrossRef] [PubMed]
5. Hu, J.; Jiang, H.; Liu, D.; Xiao, Z.; Zhang, Q.; Liu, J.; Dustdar, S. Combining IMU with Acoustics for Head Motion Tracking Leveraging Wireless Earphone. *IEEE Trans. Mob. Comput.* **2024**, *23*, 6835–6847. [CrossRef]
6. Hung, H.H.; Huang, Y.H.; Ma, C.C. An investigation on coupling analysis and vibration rectification of vibration isolation system for strap-down inertia measurement unit. *J. Mech.* **2024**, *40*, 604–624. [CrossRef]
7. Banerjee, S.; Balamurugan, V.; Kumar, R.K. Vibration control of a military vehicle weapon platform under the influence of integrated ride and cornering dynamics. *Proc. Inst. Mech. Eng. Part K J. Multi-Body Dyn.* **2021**, *235*, 197–216. [CrossRef]
8. Jakati, A.; Banerjee, S.; Jebaraj, C. Development of Mathematical Models, Simulating Vibration Control of Tracked Vehicle Weapon Dynamics. *Def. Sci. J.* **2017**, *67*, 465–475. [CrossRef]
9. Saha, S.S.; Sandha, S.S.; Srivastava, M. Machine Learning for Microcontroller-Class Hardware: A Review. *IEEE Sens. J.* **2022**, *22*, 21362–21390. [CrossRef] [PubMed]
10. Heydari, S.; Mahmoud, Q.H. Tiny Machine Learning and On-Device Inference: A Survey of Applications, Challenges, and Future Directions. *Sensors* **2025**, *10*, 3191. [CrossRef] [PubMed]
11. Elhanashi, A.; Dini, P.; Saponara, S.; Zheng, Q. Advancements in TinyML: Applications, Limitations, and Impact on IoT Devices. *Electronics* **2024**, *13*, 3562. [CrossRef]
12. Zhang, S.; Li, Y.; Zhang, S.; Shahabi, F.; Xia, S.; Deng, Y.; Alshurafa, N. Deep Learning in Human Activity Recognition with Wearable Sensors: A Review on Advances. *Sensors* **2022**, *22*, 1476. [CrossRef] [PubMed]
13. Chen, K.; Zhang, D.; Yao, L.; Guo, B.; Yu, Z.; Liu, Y. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 77. [CrossRef]
14. Mekruksavanich, S.; Phaphan, W.; Jitpattanakul, A. Enhancing Sensor-based Human Activity Recognition Using Hybrid Deep Learning and Data Augmentation. In *2024 Research, Invention, and Innovation Congress: Innovative Electricals and Electronics (RI2C)*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2024; pp. 66–71. [CrossRef]
15. Zhang, Y.; Gao, Q.; Hu, R.; Ding, Q.; Li, B.; Guo, Y. Differentiable Prior-Driven Data Augmentation for Sensor-Based Human Activity Recognition. *IEEE Trans. Comput. Soc. Syst.* **2025**, *12*, 3778–3790. [CrossRef]
16. Feng, L.; You, Y.; Liao, W.; Pang, J.; Hu, R.; Feng, L. Multi-scale change monitoring of water environment using cloud computing in optimal resolution remote sensing images. *Energy Rep.* **2022**, *8*, 13610–13620. [CrossRef]
17. Elbaz, S.; Agounad, S.; Yous, M.A.; Moufassih, M. Hand gesture recognition using temporal and spectral electromyography features and machine learning. *Knowl. Based. Syst.* **2026**, *336*, 115256. [CrossRef]
18. Qian, H.; Wang, M.; Zhu, M.; Wang, H. A Review of Multi-Sensor Fusion in Autonomous Driving. *Sensors* **2025**, *25*, 6033. [CrossRef] [PubMed]
19. Sun, X.; Liu, T.; Liu, C.; Dong, W. FCNN: Simple neural networks for complex code tasks. *J. King Saud Univ.—Comput. Inf. Sci.* **2024**, *36*, 101970. [CrossRef]
20. McHugh, M.L. The Chi-square test of independence. *Biochem. Medica* **2013**, *23*, 143–149. [CrossRef] [PubMed]
21. Deng, B.L.; Li, G.; Han, S.; Shi, L.; Xie, Y. Model Compression and Hardware Acceleration for Neural Networks: A Comprehensive Survey. *Proc. IEEE* **2020**, *108*, 485–532. [CrossRef]
22. Doan, H.G.; Nguyen, N.T. Fusion Machine Learning Strategies for Multi-modal Sensor-based Hand Gesture Recognition. *Eng. Technol. Appl. Sci. Res.* **2022**, *12*, 8628–8633. [CrossRef]
23. Di Leo, K.; Biagetti, G.; Falaschetti, L.; Crippa, P. Microcontroller Implementation of LSTM Neural Networks for Dynamic Hand Gesture Recognition. *Sensors* **2025**, *25*, 3831. [CrossRef] [PubMed]
24. Severin, I.C. Head Gesture-Based on IMU Sensors: A Performance Comparison between the Unimodal and Multimodal Approach. In *2021 International Symposium on Signals, Circuits and Systems (ISSCS)*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2021. [CrossRef]