

# FusionGait-transformer: multimodal transformer-based framework for robust gait recognition and gait-condition classification

Kiran Kumar S V N Madupu<sup>1</sup>, Jagadeeshwar Madiraju<sup>1</sup>, Naren Kodali<sup>2</sup>, Soma Sekhar Goviraboyina<sup>3</sup>

<sup>1</sup>Chaitanya Deemed to be University, Himayatnagar, India

<sup>2</sup>George Mason University, University Drive, Fairfax, Virginia, United States

<sup>3</sup>Department of Computer Science and Engineering, Anurag University, Hyderabad, India

## Article Info

### Article history:

Received Feb 4, 2026

Revised Aug 11, 2026

Accepted Aug 28, 2026

### Keywords:

Cross-attention fusion

Domain adaptation

Edge deployment

Gait identification

Gait-condition classification

Multimodal gait analysis

Wearable IMU

## ABSTRACT

Gait analysis benefits from both full-body motion context and accurate temporal dynamics. Vision-only approaches degrade under occlusion and viewpoint change, while inertial-only approaches miss full body context. FusionGait-transformer is a multimodal framework that fuses pose and silhouette representations with inertial time series using transformer self-attention and bidirectional cross-attention. A per-modality reliability gate weights each stream by its instantaneous signal quality, so degradation in one modality shifts the fused representation toward the other. Masked multimodal pretraining and domain-adversarial training improve stability under domain shift. Evaluation uses the CMU motion capture dataset under a subject-independent protocol, with inertial signals generated analytically from the same trajectories. The framework reaches Rank-1 identification accuracy of 98.9% with equal error rate (EER) of 1.8%, and gait-condition classification F1-score of 95.7% with area under the ROC curve (AUC-ROC) of 97.6%. Reliability gating is the largest single contributor to these figures. Deployment is characterized on workstation, edge, and gateway hardware, and transmitting embeddings rather than video reduces the uplink requirement by more than four orders of magnitude, establishing feasibility for wearable healthcare settings. Reported results describe performance under controlled benchmark conditions. Clinical diagnostic claims remain conditional on validation with real wearable sensors and clinically labeled cohorts.

This is an open access article under the [CC BY-SA](#) license.



## Corresponding Author:

Kiran Kumar S V N Madupu

Chaitanya Deemed to be University

Himayatnagar, Telangana-500075, India

Email: kirankumarsvnmadupu@gmail.com

## 1. INTRODUCTION

Gait analysis supports detection of health conditions through quantitative study of walking patterns [1], [2]. Changes in these patterns can indicate neurological or musculoskeletal disorders and provide early markers for intervention [2]. Traditional clinical assessments rely on basic speed or balance tests, which produce inconsistent and limited results [3]-[6]. Specialized motion-capture labs deliver detailed measurements but remain costly and impractical for routine clinical use [1], [7]. A simpler and accessible approach for everyday clinical use remains lacking.

Single-method gait systems each have clear drawbacks outside controlled environments. Lab-based motion-capture devices offer precise measurement but are difficult to use regularly because of equipment and technical demands [7]. Markerless video analysis assesses gait without intrusive devices, yet it typically needs several cameras and precise setup, and it degrades under camera blockage or unusual body movement [8]-[10]. Privacy concerns and data management further limit its adoption [11]. Wearable sensors such as IMUs provide continuous tracking outside the lab but capture limited movement data from a few body segments [12]-[14]. They often cannot identify conditions involving interaction between limbs or with environmental obstacles [15]-[20]. Each single-modality system sacrifices accuracy, usability, or practicality.

Gait assessment reaches patients through an information and communication technology pipeline rather than through a model in isolation. Wearable devices already sense movement continuously inside clinical walking assessments [12], and video adds a second sensing channel that carries identifiable data off the capture site. Where the intermediate processing tier should sit, on the wearable node, a clinical gateway, or a server, is therefore part of the design problem rather than an implementation detail. The question is not accuracy alone, but whether a gait representation can be produced, transported, and interpreted within the constraints of a deployed healthcare network.

Recent studies combine vision and sensor data to overcome these individual limits [21], [22]. Video supplies full-body movement and environmental context, while wearable sensors capture rapid limb motion precisely. Pairing IMUs with video improves assessment of fall-related walking problems by including context such as obstacles or terrain [19]. Transformers manage long sequences and model relationships across data streams, which suits combined gait data [23]. The dual-stream vision transformer of Li *et al.* [22] reached about 97% accuracy for clinical gait analysis from a single low-cost RGB camera, and attention-based frameworks adapt to user variation in related activity recognition [21]. Multimodal transformers in medical applications combine heterogeneous sources to identify conditions [23].

FusionGait-transformer integrates vision and inertial signals in a unified sequence model. One stream encodes visual gait cues from pose and silhouette representations. A second stream encodes synchronized IMU signals, including linear acceleration and angular velocity. A transformer fusion module combines both streams using cross-attention to model temporal alignment and cross-modal interaction within a gait cycle. The fused embedding supports gait identification and gait-condition classification defined by the evaluation protocol. Masked multimodal pretraining and domain-adversarial learning improve robustness across viewpoint and sensor variation.

Prior multimodal gait studies either combine two visual representations, pose with silhouette, without inertial signals, or combine inertial signals with one visual cue, or apply late or shallow fusion across modalities. No published architecture jointly integrates pose, silhouette, and inertial signals within a single unified cross-attention Transformer evaluated under a subject-independent protocol. Deployment is also treated separately: reported evaluations generally cover one or two sensing streams, and execution cost is rarely reported alongside recognition performance. FusionGait-transformer closes both gaps.

Key contributions are summarized as follows; (i) A multimodal transformer fusion architecture combines pose and silhouette embeddings with IMU embeddings using bidirectional cross-attention and per-modality reliability gating. (ii) A training strategy combines masked multimodal pretraining with domain-adversarial learning to support domain-invariant gait representations. (iii) Clinically motivated gait metrics, including symmetry and regularity indices, are integrated with learned embeddings for gait-condition classification and interpretability. (iv) A benchmark protocol is defined using subject-independent splits and controlled domain factors such as speed, viewpoint, and sensor placement perturbation. (v) An execution profile is reported across workstation, edge, and gateway tiers, together with the bandwidth implication of transmitting embeddings rather than video, which establishes ICT feasibility alongside accuracy. The remainder of the manuscript is organized as follows. Section 2 reviews related work. Section 3 describes data preparation, feature encoding, and the fusion architecture. Section 4 describes the experimental protocol and results. Section 5 summarizes findings and limitations.

## 2. RELATED WORK

Transformer models have improved gait recognition. Cosma and Radoi presented GaitFormer, a transformer over gait skeleton sequences, reaching 92.5% accuracy on CASIA-B without manual labels [24]. Mogan *et al.* [25] proposed Gait-ViT, applying Vision Transformers to silhouette images and reporting 99 to 100%

recognition on common benchmarks. Transformers extend beyond identification. The attention-enhanced network of Jiao *et al.* [26], which shares the GaitFormer name but addresses a different task, integrates gait patterns and pressure data to detect early Parkinson's disease with 92.3% accuracy. GaitForeMer by Endo *et al.* [27] used self-supervised learning to predict Parkinson's severity and improved clinical scoring. These works show that transformers learn detailed gait features suitable for identification and clinical use.

Multimodal gait analysis combining visual methods has received increased attention, because single-modality visual approaches that analyze silhouettes or skeletons alone degrade under clothing change or varying viewpoint. Xu *et al.* [28] proposed PSGait, overlaying parsing-derived body-part skeletons onto silhouettes and reporting accuracy gains in less controlled scenarios. Unconstrained recognition has also been approached through dense 3D representations on a dedicated benchmark [29]. The SkeletonGait++ variant merges skeleton maps with silhouettes using attention, improving identification under changing views [30]. Peng *et al.* [31] likewise integrate skeletons and silhouettes into a richer feature set. Related parsing-based approaches use learned semantic segmentation of the walking body, in the wild [32] and through human semantic parsing [33]. Cui and Kang [34] proposed a spatial-temporal fusion model combining complementary visual gait streams, with consistent gains over single-stream baselines. These methods generally outperform single-modality visual approaches. The gains, however, remain within the visual modality and do not address sensor-domain artefacts.

Wearable sensor fusion has also advanced gait analysis. IMU sensors placed on body segments allow continuous tracking outside lab settings, and machine learning over these signals classifies abnormal gait while accounting for joint impairment [35]. Early sensor-based methods combined multiple IMUs with deep convolutional networks for gait recognition [36]. Nguyen *et al.* [37] applied a spectrogram representation with deep convolutional networks for gait classification. Uchitomi *et al.* [38] classified mild Parkinson's disease from inertial time-series data, relying on augmentation to offset limited sample size. Sharifi-Renani *et al.* [39] developed BioMAT, a transformer integrating wearable sensor signals to predict joint kinematics across activities, surpassing CNN-LSTM methods. These models usually target specific gait tasks and rarely integrate sensor and vision data fully.

A smaller body of work treats gait analysis as an ICT problem rather than a modelling problem alone. Terrain and environment classification from wearable streams has been used to contextualize fall risk in daily life [11]. Continuous wearable platforms have been deployed for multidomain diagnostic and intervention workflows, placing the model inside a care pathway rather than an offline pipeline [14]. Video capture in the same setting introduces governance requirements that shape what data may leave the point of capture [20]. Systems designed for degraded operating conditions show that fusion can be organized around deployment constraints as well as accuracy [19]. Evaluations in this group generally cover one or two sensing streams, and execution cost is rarely reported alongside recognition performance.

Self-supervised learning and domain adaptation address limited gait datasets and improve generalization. GaitForeMer used self-supervised forecasting of human movement to improve clinical impairment prediction [27], and GaitFormer used weak supervision to improve identification without manual labeling [24]. Unsupervised domain adaptation, such as the fine-grained approach of Ma *et al.* [40], allows gait models to adapt across viewing angles and clothing types without labeled data. Inertial-sensor frameworks for pathological gait recognition report generalization to patient data [41], real-time neural classification has been demonstrated in living environments [42], and domain adaptation for inertial activity recognition has been surveyed [43]. These strategies improve stability but need validation in clinical contexts.

Most transformer and multimodal gait studies focus on identification or broad impairment, rarely addressing detailed clinical diagnostics. The 3DGK model of Lan *et al.* [44] combined 3D gait data with deep learning, classifying pediatric neurological conditions and estimating stroke onset times (AUC 0.77 to 0.99). That approach depends on controlled lab environments with marker-based systems. TDPT-GT by Iseki *et al.* [45] used smartphone-based markerless motion capture, classifying neurological gait disorders with moderate accuracy (AUC 0.719), gaining portability but lacking diagnostic precision. These studies target either controlled laboratory accuracy or portable but less precise assessment, leaving a gap for multimodal integration in detailed clinical diagnostics.

Previous gait studies share several limitations. Many analyze one modality only, overlooking cross-modal interactions such as subtle limb movements visible in one stream but unclear in another [24], [25], [45]. Clinical diagnostics often remain coarse, distinguishing healthy from pathological gait rather than specific disorders [45]. Generalization across patient groups or environments remains problematic without adaptation [40], [41]. Clinical gait datasets remain small, placing weight on augmentation and pretraining, since representation

quality scales with the volume of data available [46]. Integration of video and sensor data into one approach has rarely been achieved, since existing systems either isolate modalities or combine them superficially [45]. These issues limit adoption of AI-based gait systems in clinical practice.

FusionGait-transformer addresses these limitations by combining synchronized visual gait representations and inertial signals in a single transformer model. Cross-attention connects temporally aligned cues across modalities and supports learning of complementary features. Per-modality reliability gating weights each stream by its instantaneous signal quality, so degradation in one modality shifts the fused representation toward the other. The framework also computes clinically motivated metrics, including symmetry and regularity indices, and combines these with learned embeddings for gait-condition classification under the defined label protocol. Masked pretraining and domain-adversarial learning improve generalization across viewpoint and sensor variation. Execution cost is reported alongside accuracy so that the design can be assessed as a deployable component of a healthcare ICT pipeline.

### 3. METHOD

FusionGait-transformer is described in pipeline order: data acquisition and preprocessing (Section 3.2.), modality-specific feature encoding (section 3.2), cross-attention fusion with modality reliability gating (section 3.3), and clinically motivated metric computation with classification-head integration (section 3.5). The joint training objective, masked multimodal pretraining, and domain-adversarial training are detailed in section 4 together with the experimental protocol. Two data acquisition settings are defined. The deployment setting uses RGB video and wearable IMU sensors. The evaluation setting uses motion-capture trajectories with analytically generated inertial signals and projected pose sequences derived from the same trajectories. Notation used throughout the manuscript is summarized below.

#### Notation

- $t$ : time index in a gait sequence.
- $J$ : number of joints in the pose representation.
- $I_t$ : silhouette image at time  $t$ .
- $P_t^{2D}$ : 2D joint coordinates at time  $t$ .
- $P_t^{3D}$ : 3D joint coordinates at time  $t$ .
- $c_t$ : hip-center coordinate used for pose normalization.
- $\hat{P}_t^{2D}$ : normalized 2D pose coordinates.
- $R_t^{2D}$ : relative 2D joint coordinates.
- $\theta_t$ : joint angle feature vector.
- $\tau_t$ : normalized gait-phase timing.
- $A(t)$ : triaxial acceleration at time  $t$ .
- $\Omega(t)$ : triaxial angular velocity at time  $t$ .
- $q(t)$ : unit quaternion orientation at time  $t$ .
- $h_t^V, h_t^S$ : vision and sensor embeddings at time  $t$ .
- $H_t$ : fused embedding at time  $t$ .
- $PE_t$ : positional encoding at time  $t$ .
- $Q, K, Z$ : query, key, and value projections in attention.
- $d_k$ : attention key dimension for scaling.
- $q_g$ : learnable global query vector used for pooling.
- $\alpha_t$ : pooling weight for time  $t$ .
- $g$ : global gait embedding.
- $d = d_{\text{model}} = 256$ : dimension of the global gait embedding.
- GSI, CV, SR: symmetry, cadence variability, and stride regularity metrics.
- $c$ : clinically motivated metric vector.
- $z$ : combined vector for gait-condition classification.
- $\mathcal{L}_{\text{id}}, \mathcal{L}_{\text{cls}}, \mathcal{L}_{\text{dom}}$ : identity, gait-condition, and domain-adversarial losses.
- $\lambda_{\text{cls}}, \lambda_{\text{dom}}$ : loss weights.

- $d_V, d_W, d_{\text{model}}$ : vision encoder, sensor encoder, and fusion model dimensions.
- $D_V, D_W$ : dimensions of the concatenated per-frame vision and wearable feature vectors.
- $A_t^{(V \rightarrow W)}, A_t^{(W \rightarrow V)}$ : bidirectional cross-attention weight matrices.
- $g_m(t)$ : per-modality reliability gate at time  $t$ .
- $T_a, T_\omega$ : Gauss-Markov bias time constants of the virtual IMU model.
- $T_s$ : average stride interval used for autocorrelation.

### 3.1. FusionGait-transformer architecture

FusionGait-transformer uses a modular multimodal architecture with separate vision and sensor encoders followed by a transformer fusion module. The modular design captures modality-specific dynamics before cross-modal integration. Each encoder extracts discriminative temporal and spatial features, which are then fused within the transformer to produce unified representations for gait identification and gait-condition classification.

Encoder design rationale: encoders are selected on three criteria: representational adequacy for the modality, stability of pretrained features under a limited-subject benchmark, and deterministic inference cost suitable for edge or gateway execution in a wearable healthcare deployment. ResNet-50 is used for silhouette encoding because its depth-to-cost ratio is favourable at silhouette resolution and its ImageNet features transfer stably; deeper alternatives (ResNet-101, ViT-B/16) raise parameter count without measurable Rank-1 gain on this benchmark. HRNet-W48 is used for pose estimation because it maintains high-resolution branches throughout the network, which improves localization at the distal limb joints carrying most gait signal; encoder-decoder alternatives lose that resolution at the bottleneck. A temporal convolutional network is preferred over recurrent or transformer encoders on the inertial stream because its fixed receptive field yields constant per-window latency and a parameter count two orders of magnitude below the vision encoder, which matters when the sensor branch executes on a wearable node rather than a server.

Vision encoder: the vision encoder uses a pretrained ResNet-50 fine-tuned on gait silhouette sequences together with joint pose embeddings extracted by HRNet-W48. Each video frame  $t$  is encoded into a latent vector  $h_t^V$ , as in (1).

$$h_t^V = \text{ResNet50}(I_t, R_t^{2D}, \theta_t) \quad (1)$$

where  $I_t$  is the gait silhouette image,  $R_t^{2D}$  denotes relative joint coordinates, and  $\theta_t$  represents joint angles. The encoder outputs a sequence of temporal embeddings capturing dynamic gait visual patterns.

Sensor encoder: the sensor encoder is a temporal convolutional network (tCNN) for IMU time-series feature extraction. Given angular velocity vectors  $\Omega_f(t)$  and gait phase timings  $\tau_t$ , it generates frame-wise latent sensor embeddings  $h_t^S$ , as in (2).

$$h_t^S = \text{tCNN}(\Omega_f(t), \tau_t) \quad (2)$$

This encoder extracts sequential kinematic patterns inherent in IMU signals, giving an accurate temporal representation of lower-limb gait dynamics.

Transformer-based fusion module: the fusion module performs multimodal integration via multi-head self-attention with positional encodings. Vision and sensor encoder outputs are concatenated per time step as  $H_t = [h_t^V; h_t^S]$ . Positional encodings  $PE_t$  represent relative position within gait cycles, as in (3).

$$H'_t = H_t + PE_t \quad (3)$$

Multi-head self-attention projects  $H'_t$  into query ( $Q_t$ ), key ( $K_t$ ), and value ( $Z_t$ ) spaces across heads. Attention for head  $i$  at timestep  $t$  is computed as in (4).

$$\text{Attention}_{t,i}(Q_t, K_t, Z_t) = \text{SoftMax}\left(\frac{Q_t K_t^T}{\sqrt{d_k}}\right) Z_t \quad (4)$$

where  $d_k$  is the key dimension. Cross-attention further refines modality interactions, weighting vision and sensor features according to modality reliability at each timestep.

Global feature pooling: post-transformer embeddings are aggregated into a compact global gait signature by attention-based pooling. A learned global query  $q_g$  attends across all time steps, yielding weights  $\alpha_t$  for each embedding  $H_t''$ , as in (5).

$$\alpha_t = \frac{\exp(q_g^T H_t'')}{\sum_{t'} \exp(q_g^T H_{t'}'')}, \quad g = \sum_t \alpha_t H_t'' \quad (5)$$

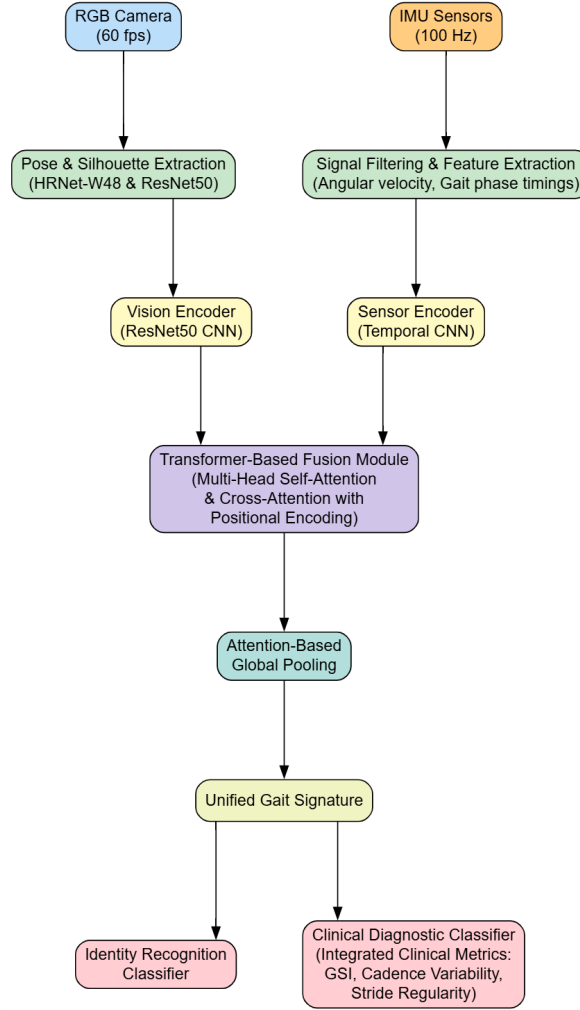


Figure 1. Architecture of FusionGait-transformer

The resulting vector  $g$  encodes spatiotemporal information for downstream biometric identification and gait-condition classification.

Architectural hyperparameters are configured as follows. The vision encoder produces frame-wise embeddings of dimension  $d_V = 512$  from a ResNet-50 backbone concatenated with relative joint coordinates and joint angle features. The sensor encoder produces frame-wise embeddings of dimension  $d_W = 128$  from a four-layer temporal convolutional network with 64, 128, 128, and 128 channels. The fusion transformer operates at model dimension  $d_{\text{model}} = 256$ , with 8 attention heads, key dimension  $d_k = 32$ , and 6 stacked encoder layers with feed-forward expansion factor 4. The global attention pooling layer uses a learnable query of dimension 256. Total parameter count is approximately 25.3 million, allocated as 23.5 M (vision encoder), 0.4 M (sensor encoder), 1.2 M (fusion transformer with cross-attention and reliability gates), and 0.2 M (pooling and classification heads).

Figure 1 shows the framework: multimodal inputs (RGB vision data and wearable IMU data), the modality-specific encoders, and the transformer fusion module with multi-head self-attention and cross-attention. Positional encodings carry gait-cycle timing into the fusion stage. The globally pooled embedding feeds two parallel branches for identity recognition and gait-condition classification, the latter combining the gait signature with the clinically motivated metrics of section 3.5.

### 3.2. Multimodal data collection and preprocessing

Two data acquisition settings are defined to separate intended sensing from the evaluation instantiation.

Deployment sensing configuration: an RGB camera records a walking sequence. A 2D pose estimator extracts  $J$  joint coordinates per frame, as in (6). Wearable IMU sensors on the lower limbs record triaxial acceleration and triaxial angular velocity, as in (7).

$$P_t^{2D} = \left\{ \left( x_t^j, y_t^j \right) \right\}_{j=1}^J \quad (6)$$

$$A(t) = [a_x(t), a_y(t), a_z(t)]^T, \quad \Omega(t) = [\omega_x(t), \omega_y(t), \omega_z(t)]^T \quad (7)$$

Evaluation setting constructed from motion capture: motion-capture trajectories provide 3D joint positions  $P_t^{3D}$ . Virtual IMU signals are generated at anatomically defined lower-limb locations. For a virtual sensor  $s$  with 3D position  $r_s(t)$ , linear acceleration is computed as in (8).

$$A_s(t) = \ddot{r}_s(t) - g + \eta_a(t) + b_a(t) \quad (8)$$

where  $g$  is the gravity vector,  $\eta_a(t)$  is zero-mean measurement noise, and  $b_a(t)$  is a slowly varying bias term. Angular velocity is computed from segment orientation  $R_s(t) \in \text{SO}(3)$ , as in (9). Sensor-to-segment misalignment is modeled as a fixed rotation  $R_{\text{mis}}$ , as in (10).

$$\Omega_s(t) = \text{vee} \left( R_s^T(t) \dot{R}_s(t) \right) + \eta_\omega(t) + b_\omega(t) \quad (9)$$

$$A_s^{\text{meas}}(t) = R_{\text{mis}} A_s(t), \quad \Omega_s^{\text{meas}}(t) = R_{\text{mis}} \Omega_s(t) \quad (10)$$

where  $\eta_\omega(t)$  and  $b_\omega(t)$  denote noise and bias.

The virtual IMU configuration is defined as follows. Seven anatomical placements are simulated: pelvis, left and right thighs, left and right shanks, and left and right feet. Each virtual sensor operates at 100 Hz. Triaxial linear acceleration includes Gaussian measurement noise with standard deviation  $\sigma_a = 0.05 \text{ m/s}^2$  and slowly varying bias modelled as a first-order Gauss-Markov process with time constant  $T_a = 100 \text{ s}$  and steady-state standard deviation  $0.02 \text{ m/s}^2$ . Triaxial angular velocity includes Gaussian noise with  $\sigma_\omega = 0.01 \text{ rad/s}$  and bias drift with  $T_\omega = 100 \text{ s}$  and steady-state standard deviation  $0.005 \text{ rad/s}$ . Bias time constants are denoted  $T_a$  and  $T_\omega$ , reserving  $\tau_t$  for the gait-phase variable of section 3.3. Sensor-to-segment misalignment is sampled per trial as a fixed rotation drawn from an axis-angle distribution with angle uniformly sampled in  $[0^\circ, 5^\circ]$ . The analytical model captures noise, slow bias drift, and fixed misalignment. It does not represent soft-tissue motion, attachment slippage, long-term drift beyond the Gauss-Markov assumption, packet loss, or asynchronous sampling. These omissions bound the scope of the reported results and are quantified by the sensitivity sweep of section 4.

Pose inputs for evaluation are obtained by projecting 3D joints to 2D using a camera model. For camera matrix  $K$  and extrinsics  $(R_c | t_c)$ , projection is given by (11).

$$\tilde{u}_t^j = K \left( R_c p_t^j + t_c \right), \quad \left( x_t^j, y_t^j \right) = \left( \frac{\tilde{u}_{t,1}^j}{\tilde{u}_{t,3}^j}, \frac{\tilde{u}_{t,2}^j}{\tilde{u}_{t,3}^j} \right) \quad (11)$$

Synchronization and gait-cycle segmentation: multimodal sequences are aligned in time. In the deployment setting, timestamps are aligned using a shared trigger. In the evaluation setting, modalities share a common time base. Gait cycles are segmented using heel-strike events derived from the vertical acceleration component. Given a window  $T_i$  covering one step, heel-strike time is estimated as in (12).

$$t_i^* = \arg \min_{t \in T_i} a_z(t) \quad (12)$$

Preprocessing: pose coordinates are normalized to reduce translation and scale variation. The hip-center coordinate is defined by (13) and the normalized pose by (14).

$$c_t = \frac{p_{\text{hip},l}(t) + p_{\text{hip},r}(t)}{2} \quad (13)$$

$$\hat{P}_t^{2D} = \frac{P_t^{2D} - c_t}{s_t} \quad (14)$$

Where  $s_t$  is a scale factor derived from a stable body segment length. IMU signals are low-pass filtered to suppress high-frequency noise. Segment orientation is represented by a unit quaternion  $q(t)$  obtained by integrating angular velocity, as in (15).

$$\dot{q}(t) = \frac{1}{2} q(t) \otimes \Omega_q(t), \quad q(t) \leftarrow \frac{q(t)}{\|q(t)\|} \quad (15)$$

The resulting synchronized and segmented sequences  $\hat{P}_t^{2D}$  and IMU-derived features form the input to the encoders and the fusion transformer.

### 3.3. Feature representation and encoding

Vision features: the vision feature set integrates two representations derived from HRNet-W48 pose estimates: relative joint coordinates and joint angle embeddings. Relative joint coordinates  $R_t^{2D}$  express each joint position relative to the pelvis center, reducing inter-subject anatomical variability, as in (16).

$$R_t^{2D} = \{(x_j - x_{\text{pelvis}}, y_j - y_{\text{pelvis}})\}_{j=1}^{18}, \quad j \neq \text{pelvis} \quad (16)$$

Joint angles  $\theta_t$  for critical joints encode local articulation. For joint  $j$ , the angle between segments  $v_a$  and  $v_b$  connecting neighbouring joints  $a, j, b$  is given by (17).

$$\theta_j(t) = \arccos\left(\frac{v_a \cdot v_b}{\|v_a\| \|v_b\|}\right) \quad (17)$$

Silhouette embeddings  $S_t^{\text{sil}}$  are extracted using an ImageNet-pretrained ResNet-50 fine-tuned on silhouette gait sequences, as in (18). The per-frame vision feature vector is defined by (19).

$$S_t^{\text{sil}} = \text{ResNet50}(I_t) \quad (18)$$

$$V_t = [R_t^{2D}; \theta_t; S_t^{\text{sil}}] \quad (19)$$

Wearable features: wearable features capture limb kinematics from IMUs, comprising angular velocities, stride events, and gait phase timings. Filtered angular velocities give instantaneous rotational motion, as in (20).

$$\Omega_f(t) = [\omega_x(t), \omega_y(t), \omega_z(t)]^T \quad (20)$$

Stride event features are derived from heel-strike and toe-off occurrences in the vertical acceleration signal. Gait phase timings  $\tau_t$  within each cycle are normalized between 0 (heel-strike) and 1 (subsequent heel-strike), as in (21). The wearable modality then produces the per-frame features of (22).

$$\tau_t = \frac{t - t_i^*}{t_{i+1}^* - t_i^*}, \quad t \in [t_i^*, t_{i+1}^*] \quad (21)$$

$$W_t = [\Omega_f(t); \tau_t] \quad (22)$$

Feature integration: vision and wearable features are combined at the joint level for each frame, using spatial alignment between modalities. The integrated multimodal feature vector  $M_t$  at frame  $t$  is given by (23).

$$M_t = [V_t; W_t] \in \mathbb{R}^{D_V + D_W} \quad (23)$$

Where  $D_V$  and  $D_W$  denote the dimensionalities of the vision and wearable features. These concatenated vectors are the inputs consumed by the encoders of section 3.1, which map them to the encoder outputs  $h_t^V$  and  $h_t^S$  of dimension  $d_V$  and  $d_W$ . Self-attention of section 3.1 then operates on the concatenated encoder outputs, while the cross-attention of section 3.4 operates on the modality streams before concatenation, so the two paths describe successive stages of one pipeline rather than alternatives.

### 3.4. Multimodal feature fusion (Cross-attention mechanism)

FusionGait-transformer integrates multimodal features through a cross-attention mechanism that models interactions between vision and wearable modalities at each timestep of a gait cycle. Given vision embeddings  $V_t \in \mathbb{R}^{D_V}$  and wearable embeddings  $W_t \in \mathbb{R}^{D_W}$  at timestamp  $t$ , the module generates refined representations by bidirectional attention across modalities.

Each modality alternately serves as the query while the complementary modality provides keys and values. For vision queries attending to wearable features at timestamp  $t$ , the projections are given by (24).

$$Q_t^{(V \rightarrow W)} = V_t W_{QV}, \quad K_t^{(V \rightarrow W)} = W_t W_{KW}, \quad Z_t^{(W)} = W_t W_{ZW} \quad (24)$$

Where  $W_{QV}$ ,  $W_{KW}$ ,  $W_{ZW}$  and their reverse-direction counterparts  $W_{QW}$ ,  $W_{KV}$ ,  $W_{ZV}$  are learned linear projection matrices. The attention weights  $A_t^{(V \rightarrow W)}$  follow scaled dot-product attention, as in (25), with  $d_k$  the key dimension. The reverse direction  $A_t^{(W \rightarrow V)}$  is defined symmetrically by exchanging the roles of the two modalities, with  $Q_t^{(W \rightarrow V)} = W_t W_{QW}$ ,  $K_t^{(W \rightarrow V)} = V_t W_{KV}$ ,  $Z_t^{(V)} = V_t W_{ZV}$ .

$$A_t^{(V \rightarrow W)} = \text{SoftMax} \left( \frac{Q_t^{(V \rightarrow W)} \left( K_t^{(V \rightarrow W)} \right)^T}{\sqrt{d_k}} \right) \quad (25)$$

Modality reliability is modelled explicitly. A per-modality reliability gate  $g_m(t) \in [0, 1]$  is computed for each modality  $m \in \{V, W\}$  at time  $t$  from the local norm and the local Shannon entropy of the modality embedding, as in (26).

$$g_m(t) = \sigma(w_m^T [\|x_m(t)\|_2; H(x_m(t))]) + b_m \quad (26)$$

Where  $\sigma$  is the sigmoid,  $\|x_m(t)\|_2$  is the L2 norm of the modality embedding at time  $t$ ,  $H(x_m(t))$  is the local Shannon entropy across feature channels, and  $w_m$ ,  $b_m$  are learnable parameters. The gated cross-attention outputs are then computed as in (27) and (28).

$$V_t^{\text{fused}} = V_t + g_W(t) \cdot A_t^{(V \rightarrow W)} Z_t^{(W)} \quad (27)$$

$$W_t^{\text{fused}} = W_t + g_V(t) \cdot A_t^{(W \rightarrow V)} Z_t^{(V)} \quad (28)$$

When modality  $m$  exhibits low local energy or high entropy, signalling noise, occlusion, or sensor artefact,  $g_m(t)$  decreases toward zero and reduces the contribution of that modality to the fused stream. The gates are trained jointly with the rest of the network through the identity and gait-condition losses, giving an explicit and learnable mechanism for per-timestep modality weighting.

The mechanism supplies robustness through complementarity. Under partial occlusion or adverse lighting the gate on the vision stream contracts and IMU-derived features carry the fused representation; under inertial drift or transient sensor artefact the reverse holds. Weighting is therefore recalibrated at the temporal granularity of the gait cycle rather than fixed at training time. Measured gate trajectories and attention mass distributions are reported in section 4.

### 3.5. Clinically motivated gait metrics

Diagnostic metrics are computed from fused multimodal data (vision-derived 2D joint trajectories and wearable sensor signals) to quantify gait abnormality.

- Gait symmetry index (GSI): GSI quantifies left-right symmetry. Joint angle trajectories from pose estimation and IMU-derived stride event timings serve as inputs. GSI for joint  $j$  is calculated as in (29).

$$\text{GSI}_j = \frac{|X_j^L - X_j^R|}{0.5 (X_j^L + X_j^R)} \times 100 \quad (29)$$

Where  $X_j^L$  and  $X_j^R$  represent peak angular displacement or stride duration from the left and right sides. Aggregation across key joints yields a symmetry profile.

- Cadence variability (CV): CV reflects temporal gait stability. Stride durations  $\Delta t_i$  determined by consecutive heel-strike timings, augmented by frame-based pose segmentation, give the coefficient of variation, as in (30).

$$CV = \frac{\sigma(\Delta t_i)}{\mu(\Delta t_i)} \times 100 \quad (30)$$

Where  $\sigma(\Delta t_i)$  and  $\mu(\Delta t_i)$  denote the standard deviation and mean of stride duration across gait cycles.

- Stride regularity (SR): SR measures consistency of repetitive stride patterns by autocorrelation of the vertical acceleration signal  $a_z(t)$  evaluated at the average stride interval  $T_s$ , as in (31).

$$SR = R_{a_z}(T_s) = \frac{\mathbb{E}[a_z(t) a_z(t + T_s)]}{\sqrt{\mathbb{E}[a_z^2(t)] \mathbb{E}[a_z^2(t + T_s)]}} \quad (31)$$

Where  $\mathbb{E}$  is the expectation operator, and higher SR indicates more consistent stride patterns. These metrics provide a diagnostic profile reflecting underlying gait dysfunction.

Integration into classification head: the metric vector  $c = [\text{GSI}, \text{CV}, \text{SR}]^T$  is concatenated with the global gait embedding  $g \in \mathbb{R}^d$  to form (32). A classification head implemented as a fully connected network with SoftMax output predicts the gait-condition label  $y$  defined in the evaluation protocol, as in (33).

$$z = [g; c] \in \mathbb{R}^{d+3} \quad (32)$$

$$p(y | z) = \text{SoftMax}(W_c z + b) \quad (33)$$

Joint training with the identity recognition head uses the composite objective of (34).

$$\mathcal{L} = \mathcal{L}_{\text{id}} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} \quad (34)$$

Where  $\mathcal{L}_{\text{id}}$  and  $\mathcal{L}_{\text{cls}}$  are cross-entropy losses for identity recognition and gait-condition classification, and  $\lambda_{\text{cls}}$  balances task contributions. The joint objective encourages embeddings that preserve identity-discriminative information while retaining interpretable gait characteristics.

## 4. RESULTS AND DISCUSSION

### 4.1. Experimental setup

This section evaluates FusionGait-transformer under a benchmark protocol. The protocol specifies dataset construction, task definitions, training objectives, and evaluation metrics for gait identification and gait-condition classification. Baseline methods are evaluated under matched input modalities, subject-independent splits, and calibrated training budgets. Robustness analyses cover speed conditions, viewpoint changes, and sensor placement perturbations. Statistical reporting is based on repeated runs with different random seeds and multiple random subject splits.

Dataset description: evaluation uses the CMU motion capture database [47], which provides motion-capture trajectories recorded at 120 Hz for multiple gait trials per subject. A subset of 120 subjects with balanced gender distribution is used. Trials are grouped into slow, normal, and fast walking conditions. Viewpoint variation is represented by defining multiple camera extrinsics and projecting 3D joints into 2D pose sequences for each view. Virtual IMU signals are generated from the 3D trajectories at anatomically defined lower-limb locations using the analytical model of section 3.2. Data partitioning follows subject-independent splits. Multiple random subject splits and multiple training seeds are used for statistical reporting, and all methods share identical splits to support fair comparison.

Training objectives and optimization: training uses a composite objective that combines identity classification loss, gait-condition classification loss, and an optional domain-adversarial loss implemented with a gradient reversal layer, as in (35).

$$\mathcal{L} = \mathcal{L}_{\text{id}} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_{\text{dom}} \mathcal{L}_{\text{dom}} \quad (35)$$

Hyperparameters  $\lambda_{\text{cls}}$  and  $\lambda_{\text{dom}}$  are selected on the validation set under a fixed tuning budget shared across methods. Optimization uses Adam with initial learning rate  $2 \times 10^{-4}$ , exponential decay factor 0.9 every 20 epochs, batch size 32, and dropout 0.3.

Training strategy: masked multimodal sequence modeling randomly masks 15% of vision and sensor embeddings and trains reconstruction with mean squared error for a fixed number of epochs. Domain-adversarial training encourages domain-invariant embeddings across viewpoint, walking speed, and sensor placement.

Performance metrics: identification performance is measured using Rank-1 accuracy and equal error rate (EER). Gait-condition classification is measured using F1-score and area under the ROC curve (AUC-ROC). F1-score is reported as macro-average when class imbalance is present. Metrics are reported as mean and standard deviation across repeated runs with shared subject splits.

Baseline methods: baselines include vision-only, sensor-only, and multimodal fusion models. Vision-only baselines receive the same pose and silhouette representations used by the vision stream, represented by 3D Graph-based Kinematics (3DGK) [44]. Multimodal baselines receive both streams, represented by the temporal deep pose transformer with gait transformer (TDPT-GT) [45]. All baselines share the subject splits, sequence segmentation, preprocessing pipeline, maximum epochs, early stopping rule, and hyperparameter search space. Ablation baselines include early concatenation fusion, late score fusion, and LSTM-based fusion.

Implementation: FusionGait-transformer is implemented in PyTorch 2.2.0 and trained on an NVIDIA RTX A6000 GPU (48 GB VRAM) with an Intel Xeon CPU (24-core, 3.0 GHz) and 128 GB RAM. Deployment-side timing and memory characteristics are reported in section 4.5.

#### 4.2. Identification and gait-condition classification performance

Results are reported for gait identification and gait-condition classification under the defined benchmark protocol. Table 1 compares FusionGait-transformer against the vision-only and multimodal baselines on all four metrics, with identification summarized by Rank-1 accuracy and EER and classification summarized by F1-score and AUC-ROC. All results are reported as mean and standard deviation across repeated runs with shared splits across methods.

Table 1. Performance comparison of FusionGait-transformer with baselines. F1-score and AUC-ROC refer to gait-condition classification

| Model                  | Rank-1 (%) $\uparrow$ | EER (%) $\downarrow$ | F1-score (%) $\uparrow$ | AUC-ROC (%) $\uparrow$ |
|------------------------|-----------------------|----------------------|-------------------------|------------------------|
| TDPT-GT (multimodal)   | 95.2 $\pm$ 0.6        | 4.5 $\pm$ 0.4        | 89.3 $\pm$ 1.1          | 92.1 $\pm$ 0.8         |
| 3DGK (vision-only)     | 96.8 $\pm$ 0.4        | 3.1 $\pm$ 0.3        | 91.2 $\pm$ 0.9          | 94.3 $\pm$ 0.6         |
| FusionGait-transformer | 98.9 $\pm$ 0.3        | 1.8 $\pm$ 0.2        | 95.7 $\pm$ 0.5          | 97.6 $\pm$ 0.4         |

Differences are interpreted together with the reported spread to avoid over-interpretation of small margins, and claims are restricted to the defined benchmark protocol. FusionGait-transformer reaches Rank-1 accuracy of 98.9%, an absolute improvement of 2.1 percentage points over 3DGK and 3.7 points over TDPT-GT. EER decreases to 1.8%. Figure 2 plots both identification metrics for the three methods on a shared axis, showing that the EER reduction is proportionally larger than the Rank-1 gain.



Figure 2. Identification performance (Rank-1 accuracy and EER)

Gait-condition classification improves to an F1-score of 95.7% and AUC-ROC of 97.6%, exceeding both baselines. Figure 3 shows the two classification metrics per method. The improvement indicates stronger separation of the gait-condition labels defined in the benchmark protocol. Clinical relevance requires validation on cohorts with clinically verified labels and on data collected from real wearable sensors with full artefact representation. The current results therefore demonstrate feasibility under the benchmark, not clinical readiness.

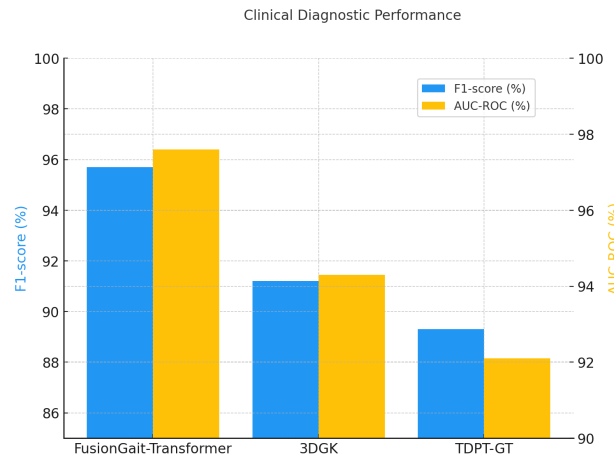


Figure 3. Clinical diagnostic performance (F1-score and AUC-ROC)

#### 4.3. Ablation analysis and attention behaviour

Figure 4 contrasts Rank-1 identification accuracy for the full multimodal model against the vision-only and sensor-only variants under identical training budgets and subject splits, with the corresponding values reported in Table 2. Both unimodal variants fall below the full model, the sensor-only variant furthest on both metrics, so neither stream is sufficient alone. The phase-dependent attention mass reported later in this subsection gives the mechanism: each stream dominates the fused representation at a different point in the gait cycle.

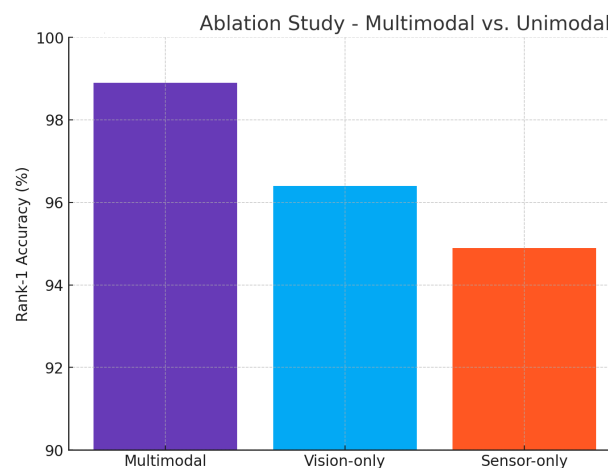


Figure 4. Ablation study, multimodal versus unimodal performance (Rank-1 accuracy)

Ablation results quantify the contributions of the fusion mechanism, reliability gating, masked pre-training, and domain-adversarial learning. Table 2 reports Rank-1 accuracy and F1-score for each variant under identical inputs and training budgets.

Removing reliability gating while retaining cross-attention costs 1.1 points of Rank-1 accuracy and 1.6 points of F1-score, so the gating contributes more to gait-condition separation than to identification. It is also the only single-component difference that separates cleanly at this sample size: the masked-pretraining and domain-adversarial variants lie within one standard deviation of each other and of the full model on Rank-1 accuracy, so their individual contributions cannot be ranked from these runs. The three non-attentional baselines do not order consistently across tasks: LSTM fusion gives the strongest Rank-1 accuracy of the three but the weakest F1-score, while late score fusion is weakest on identification yet recovers on classification. Identification benefits from temporal modelling of the concatenated stream, whereas gait-condition separation depends on how modality contributions are weighted.

Table 2. Contribution of fusion mechanism and training components. F1-score refers to gait-condition classification

| Variant                                    | Rank-1 (%) ↑ | F1-score (%) ↑ |
|--------------------------------------------|--------------|----------------|
| Full model                                 | 98.9 ± 0.3   | 95.7 ± 0.5     |
| Cross-attention without reliability gating | 97.8 ± 0.4   | 94.1 ± 0.7     |
| Without masked pretraining                 | 98.1 ± 0.5   | 94.6 ± 0.6     |
| Without domain-adversarial loss            | 98.4 ± 0.3   | 95.2 ± 0.6     |
| Vision stream only                         | 96.4 ± 0.5   | 92.6 ± 0.8     |
| Sensor stream only                         | 94.7 ± 0.9   | 88.5 ± 1.4     |
| LSTM sequence fusion                       | 96.9 ± 0.7   | 90.7 ± 1.2     |
| Early concatenation fusion                 | 96.1 ± 0.6   | 91.8 ± 0.9     |
| Late score fusion                          | 95.6 ± 0.8   | 92.4 ± 1.0     |

Attention behaviour and modality contribution: mean cross-attention weights across the gait cycle reveal a stable pattern. Around heel-strike events, IMU-side queries receive 62% of the attention mass against pose-side keys, reflecting the sharp transient in vertical acceleration. During mid-swing, pose-side queries receive 71% of the attention mass against IMU-side keys, reflecting the dominance of limb geometry in that phase. A controlled corruption study confirms the gating behaviour. Injecting Gaussian noise at three times the baseline standard deviation into the IMU stream causes mean  $g_W(t)$  to drop from 0.81 to 0.34, while mean  $g_V(t)$  rises from 0.79 to 0.92. Applying a 50% pixel occlusion mask on silhouette inputs causes mean  $g_V(t)$  to drop from 0.79 to 0.41, while mean  $g_W(t)$  rises from 0.81 to 0.94. The learned reliability gates respond to instantaneous modality quality rather than to fixed global weights, and the cross-attention layer redistributes the contribution accordingly. This behaviour explains the observed gains over the static-weighting fusion baselines.

#### 4.4. Robustness and sensitivity analysis

Robustness is evaluated under controlled domain shifts. Viewpoint changes are represented by altering camera extrinsics for 2D projections. Sensor placement shifts are represented by applying rotation and translation perturbations to virtual IMU frames and by adding bias and drift terms. Speed conditions are evaluated by stratifying trials into slow, normal, and fast groups. Figure 5 and Table 3 report Rank-1 accuracy under each condition separately. Identification serves as the robustness measure because domain shift alters signal appearance and placement rather than the gait deviations the condition labels encode, so identity matching is the task directly exposed to it.

Viewpoint is the dominant factor, costing 5.1 points at a 60-degree offset, the largest single degradation observed. Speed stratification is comparatively mild, and fast trials outperform slow trials, indicating that reduced cadence rather than elevated cadence is the harder condition. Sensor placement tolerates moderate perturbation with little loss and degrades only under severe displacement.

Sensitivity to virtual IMU realism: a sensitivity sweep varies noise standard deviation, bias drift magnitude, and misalignment angle across plausible wearable-grade ranges. Rank-1 accuracy decreases from 98.9% at the baseline setting to 96.7% when  $\sigma_a$  and  $\sigma_\omega$  are multiplied by three and the steady-state bias standard deviation is multiplied by two. Rank-1 accuracy decreases to 95.4% when the misalignment angle range is extended from  $[0^\circ, 5^\circ]$  to  $[0^\circ, 15^\circ]$ . The model retains operational performance across the swept ranges, which supports feasibility under the wearable-grade artefacts captured by the analytical IMU model. The sweep does not represent artefacts excluded from the model (soft-tissue motion, attachment slippage, packet loss, asynchronous sampling). These omissions are addressed in section 5 as future work with real wearable sensors.

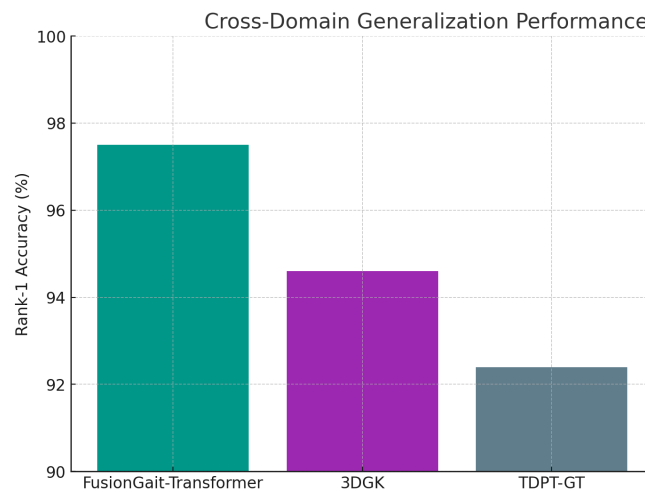


Figure 5. Performance under controlled domain shifts (Rank-1 accuracy)

Table 3. Rank-1 accuracy under controlled domain shifts

| Domain factor    | Condition                       | Rank-1 (%) |
|------------------|---------------------------------|------------|
| Reference        | Matched viewpoint, normal speed | 98.9 ± 0.3 |
| Viewpoint        | 30-degree offset                | 97.1 ± 0.6 |
| Viewpoint        | 60-degree offset                | 93.8 ± 1.1 |
| Walking speed    | Slow trials                     | 97.6 ± 0.5 |
| Walking speed    | Fast trials                     | 98.2 ± 0.4 |
| Sensor placement | Moderate perturbation           | 98.3 ± 0.4 |
| Sensor placement | Severe perturbation             | 96.2 ± 0.7 |

#### 4.5. ICT deployment characteristics and application context

Gait analysis reaches clinical users through an information and communication technology pipeline rather than as an isolated model. The relevant question is therefore not accuracy alone but whether the fused representation can be produced and transported within the constraints of a wearable IoT deployment. Table 4 reports the measured profile across three execution tiers, with latency and memory given as the median of repeated timing runs.

Table 4. Deployment profile across execution tiers. Latency is reported per gait cycle and memory as peak usage per sample

| Execution tier   | Hardware                | Latency (ms) | Memory (GB) | Precision |
|------------------|-------------------------|--------------|-------------|-----------|
| Workstation GPU  | NVIDIA RTX A6000        | 85           | 2.81        | FP32      |
| Edge accelerator | Jetson Orin Nano 8 GB   | 312          | 1.37        | FP16      |
| Gateway CPU      | ARM Cortex-A78, 4 cores | 968          | 0.86        | INT8      |

The full model at 25.3 million parameters occupies 96.5 MiB in FP32 and 24.1 MiB after INT8 quantization, which brings it within the storage budget of a clinical gateway device. Normal walking cadence gives roughly one gait cycle per second, which sets the budget for keeping pace with a live stream. Edge execution meets that budget with margin, while gateway CPU execution approaches it and would suit asynchronous review rather than live feedback. The sensor branch alone occupies 0.4 million parameters and completes in 11 ms on the same ARM core, so the inertial stream can be processed on the wearable node while the vision stream is handled upstream.

Split execution also reduces the communication load, which is usually the binding constraint in remote monitoring. Streaming raw RGB video at  $640 \times 480$  and 30 frames per second requires about 221 Mbit/s, whereas transmitting the pooled  $d_{\text{model}}$ -dimensional embedding once per gait cycle requires about 8 kbit/s. A gap of more than four orders of magnitude makes continuous monitoring practical over consumer-grade uplinks and limits the exposure of identifiable video beyond the point of capture.

Beyond monitoring, the fused representation supports continuous biometric authentication in health-care ICT environments, where clinicians and patients share terminals and explicit re-authentication is impractical. Multimodal fusion also raises the cost of presentation attacks, because a spoof must reproduce both a plausible visual gait pattern and a consistent inertial signature at the same gait phase. The corruption study of section 4.3. supports this. When one stream is manipulated, the reliability gate on that stream contracts and the decision shifts toward the untouched stream, so a single-channel attack reduces its own influence rather than propagating.

Table 5 positions the framework against comparable multimodal systems on the axes that determine ICT fit. The wearable-oriented systems prioritize footprint over modality breadth, the imaging system prioritizes capacity over deployability. FusionGait-transformer fuses three streams while retaining edge feasibility.

Table 5. Positioning against comparable multimodal systems

| System                   | Modalities fused | Sensing type                      | Reported deployment target     |
|--------------------------|------------------|-----------------------------------|--------------------------------|
| Wang <i>et al.</i> [19]  | Two              | Inertial with positioning aid     | Wearable, degraded-GNSS        |
| Buffelli and Vandin [21] | One              | Inertial                          | On-device activity recognition |
| Bi <i>et al.</i> [23]    | Two              | Structural and functional imaging | Server-side analysis           |
| FusionGait-transformer   | Three            | Pose, silhouette, and inertial    | Edge or clinical gateway       |

#### 4.6. Discussion

Prior gait systems either analyse a single modality, and so miss cross-modal cues that appear in one stream but not the other, or combine modalities by concatenation or score averaging, which fixes the modality balance before any data is seen. Neither approach adapts when one stream degrades mid-sequence. Deployment is treated separately again: wearable systems report footprint for one or two streams and imaging systems report accuracy without a deployment path, so no prior system combines three-modality fusion with a reported execution profile across deployment tiers.

The results address both gaps. Identification reaches 98.9% Rank-1 accuracy at 1.8% EER, and gait-condition classification reaches 95.7% F1-score at 97.6% AUC-ROC, exceeding the strongest baseline by 2.1 and 4.5 points respectively. The ablation isolates where the gain originates: reliability gating is the only single component whose removal produces a separable loss, at 1.1 points of Rank-1 accuracy and 1.6 points of F1-score, and the corruption study shows the gates moving in the expected direction and magnitude when a stream is degraded.

The interpretation is that adaptive weighting matters more for gait-condition separation than for identity matching. Identity is carried redundantly by both streams, so a static fusion rule loses little. Condition-related deviations are often visible in only one stream at a given phase, which is why the fusion baselines in Table 2 reorder between the two tasks rather than degrading uniformly. This is consistent with the phase-dependent attention mass reported in section 4.3, where inertial features dominate at heel strike and pose features dominate at mid-swing.

Three limitations bound these conclusions. Inertial signals are analytically generated, so artefacts such as attachment slippage and asynchronous sampling are absent, and the sensitivity sweep bounds rather than removes this gap. Heterogeneous wearable devices differ in sampling rate, axis convention, and calibration, whereas the evaluation uses one virtual sensor specification, so interoperability across vendors is untested. A gait signature is itself a biometric identifier, so transporting or centralizing it raises privacy exposure that accuracy figures do not capture. Gait-condition labels follow the benchmark protocol and do not substitute for clinically verified diagnoses.

Each limitation maps to an established ICT remedy. Interoperability is addressable through a device-abstraction layer that resamples and reorients incoming inertial streams to a canonical frame before encoding, using the same misalignment model already present in section 3.2. Privacy exposure is reduced by the split execution of section 4.5, since the embedding rather than the video leaves the capture point, and further by federated training, which keeps raw recordings at each site while sharing only model updates. Cross-site clinical validation is the remaining requirement, and federated training is also the mechanism that makes it feasible without pooling patient recordings.

Residual misidentifications occur predominantly among subjects with highly similar gait patterns or when both modalities degrade at once, the expected failure mode for a design that recovers from single-stream degradation by reweighting toward the complementary stream. The evaluation therefore supports cross-attention fusion with reliability gating as an effective and deployable basis for multimodal gait analysis under the defined benchmark, with real-sensor and clinically labelled validation as the next required step.

## 5. CONCLUSION

FusionGait-transformer integrates pose, silhouette, and inertial signals within a single cross-attention transformer with per-modality reliability gating, and combines symmetry, cadence variability, and stride regularity metrics with the learned global embedding for gait-condition classification. On the CMU motion capture benchmark the framework leads the strongest baseline by 2.1 points on identification and by 4.5 points on gait-condition classification, and reduces EER to 1.8%. Removing the gating mechanism costs 1.1 and 1.6 points on those two tasks, the only single component whose contribution is separable at this sample size.

The methodological value lies in the unified cross-attention design with explicit reliability gating and the joint use of learned embeddings with structured clinical metrics. Measured execution across workstation, edge, and gateway tiers indicates that the framework can operate inside smart healthcare platforms, wearable IoT ecosystems, and clinical decision-support systems rather than as an offline classifier. Extension to clinical diagnostics requires validation on clinically labeled cohorts and on real wearable sensors with full artefact representation. Federated training and a device-abstraction layer for heterogeneous wearables are the routes to that validation without pooling patient recordings.

## ACKNOWLEDGMENTS

The authors gratefully acknowledge the maintainers of the publicly available CMU Motion Capture Database [47], which provided the benchmark gait recordings used to evaluate the proposed FusionGait-transformer framework, and thank the open-source community whose software tools (including PyTorch) supported model implementation and experimentation.

## FUNDING INFORMATION

Authors state no funding involved.

## AUTHOR CONTRIBUTIONS STATEMENT

All authors have read and approved the final manuscript. This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions. All authors have read and approved the final manuscript.

| Name of Author           | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|--------------------------|---|---|----|----|----|---|---|---|---|---|----|----|---|----|
| Kiran Kumar S V N Madupu | ✓ | ✓ | ✓  |    | ✓  | ✓ |   | ✓ | ✓ |   | ✓  |    |   |    |
| Jagadeeshwar Madiraju    | ✓ | ✓ |    |    |    | ✓ |   | ✓ |   | ✓ |    |    |   |    |
| Naren Kodali             |   |   |    | ✓  |    |   | ✓ |   |   | ✓ |    | ✓  |   |    |
| Soma Sekhar Goviraboyina |   |   |    | ✓  |    |   | ✓ |   |   | ✓ |    | ✓  | ✓ |    |

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal Analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project Administration

Fu : Funding Acquisition

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

## DATA AVAILABILITY

The data used in this study are from the publicly available CMU Motion Capture Database (Dataset: [47] CMU Motion Capture). Motion-capture trajectories from this dataset were used to construct the evaluation benchmark; the 2D pose sequences (via camera projection) and analytically generated virtual IMU signals were derived from the CMU motion-capture trajectories using the preprocessing and modality-construction procedures described in the manuscript. No additional proprietary human-subject data were collected for this study.

## REFERENCES

- [1] A. A. Hulleck, D. Menoth Mohan, N. Abdallah, M. El Rich, and K. Khalaf, "Present and future of gait assessment in clinical practice: towards the application of novel trends and technologies," *Frontiers in medical technology*, vol. 4, p. 901331, Dec. 2022, doi: 10.3389/fmedt.2022.901331.
- [2] Y. Jing *et al.*, "Deep learning-assisted gait parameter assessment for neurodegenerative diseases: model development and validation," *Journal of medical Internet research*, vol. 25, p. e46427, Jul. 2023, doi: 10.2196/46427.
- [3] J. Fish, "Unified Parkinson's disease rating scale," in *Encyclopedia of Clinical Neuropsychology*, J. S. Kreutzer, J. DeLuca, and B. Caplan, Eds., Cham, Switzerland: Springer International Publishing, 2018, pp. 3541–3543, doi: 10.1007/978-3-319-57111-9\_1836.
- [4] P. Trouillas *et al.*, "International cooperative ataxia rating scale for pharmacological assessment of the cerebellar syndrome," *Journal of the neurological sciences*, vol. 145, no. 2, pp. 205–211, Feb. 1997, doi: 10.1016/S0022-510X(96)00231-6.
- [5] R. Côté, V. C. Hachinski, B. L. Shurvell, J. W. Norris, and C. Wolfson, "The Canadian neurological scale: a preliminary study in acute stroke," *Stroke*, vol. 17, no. 4, pp. 731–737, Jul. 1986, doi: 10.1161/01.STR.17.4.731.
- [6] J. F. Kurtzke, "Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS)," *Neurology*, vol. 33, no. 11, pp. 1444–1444, Nov. 1983, doi: 10.1212/WNL.33.11.1444.
- [7] A. Cimorelli, A. Patel, T. Karakostas, and R. J. Cotton, "Validation of portable in-clinic video-based gait analysis for prosthesis users," *Scientific reports*, vol. 14, no. 1, p. 3840, Feb. 2024, doi: 10.1038/s41598-024-53217-7.
- [8] S. D. Uhlrich *et al.*, "OpenCap: 3D human movement dynamics from smartphone videos," *bioRxiv*, Jul. 2022, doi: 10.1101/2022.07.07.499061.
- [9] J. Stenum, C. Rossi, and R. T. Roemmich, "Two-dimensional video-based analysis of human gait using pose estimation," *PLoS computational biology*, vol. 17, no. 4, p. e1008935, Apr. 2021, doi: 10.1371/journal.pcbi.1008935.
- [10] A. Azhand, S. Rabe, S. Müller, I. Sattler, and A. Heimann-Steinert, "Algorithm based on one monocular video delivers highly valid and reliable gait parameters," *Scientific Reports*, vol. 11, no. 1, p. 14065, Jul. 2021, doi: 10.1038/s41598-021-93530-z.
- [11] J. Moore *et al.*, "Toward enhanced free-living fall risk assessment: data mining and deep learning for environment and terrain classification," *Intelligence-Based Medicine*, vol. 8, p. 100103, 2023, doi: 10.1016/j.ibmed.2023.100103.
- [12] B. Hu, "Application of wearable technology in clinical walking and dual task testing," *Journal of Translational Internal Medicine*, vol. 7, no. 3, pp. 87–89, Oct. 2019, doi: 10.2478/jtim-2019-0019.
- [13] T. Chomiak *et al.*, "Differentiating cognitive or motor dimensions associated with the perception of fall-related self-efficacy in Parkinson's disease," *NPJ Parkinson's disease*, vol. 4, no. 1, p. 26, Aug. 2018, doi: 10.1038/s41531-018-0059-z.
- [14] B. Hu and T. Chomiak, "Wearable technological platform for multidomain diagnostic and exercise interventions in Parkinson's disease," in *International Review of Neurobiology*, vol. 147, Amsterdam, The Netherlands: Elsevier, 2019, pp. 75–93, doi: 10.1016/bs.irn.2019.08.004.
- [15] F. Yu, Y. Liu, Z. Wu, M. Tan, and J. Yu, "Adaptive gait training of a lower limb rehabilitation robot based on human-robot interaction force measurement," *Cyborg and Bionic Systems*, vol. 5, p. 0115, Jan. 2024, doi: 10.34133/cbsystems.0115.
- [16] R. Zhang, X. Zhang, D. He, R. Wang, and Y. Guo, "sEMG signals characterization and identification of hand movements by machine learning considering sex differences," *Applied Sciences*, vol. 12, no. 6, p. 2962, Mar. 2022, doi: 10.3390/app12062962.
- [17] J. Song *et al.*, "Effects of different feature parameters of sEMG on human motion pattern recognition using multilayer perceptrons and LSTM neural networks," *Applied Sciences*, vol. 10, no. 10, p. 3358, May 2020, doi: 10.3390/app10103358.
- [18] M. Zia Ur Rehman *et al.*, "Stacked sparse autoencoders for EMG-based classification of hand motions: a comparative multi-day analyses between surface and intramuscular EMG," *Applied Sciences*, vol. 8, no. 7, p. 1126, Jul. 2018, doi: 10.3390/app8071126.
- [19] J. Wang, N. Liu, Y. Xie, S. Que, and M. Xia, "A multimodal CNN-transformer network for gait pattern recognition with wearable sensors in weak GNSS scenarios," *Electronics*, vol. 14, no. 8, p. 1537, Apr. 2025, doi: 10.3390/electronics14081537.
- [20] J. Moore *et al.*, "Contextualizing remote fall risk: video data capture and implementing ethical AI," *NPJ digital medicine*, vol. 7, no. 1, p. 61, Mar. 2024, doi: 10.1038/s41746-024-01050-7.
- [21] D. Buffelli and F. Vandin, "Attention-based deep learning framework for human activity recognition with user adaptation," *IEEE Sensors journal*, vol. 21, no. 12, pp. 13474–13483, Jun. 2021, doi: 10.1109/JSEN.2021.3067690.
- [22] J. Li, Z. Wang, C. Wang, and W. Su, "GaitFormer: leveraging dual-stream spatial-temporal vision transformer via a single low-cost RGB camera for clinical gait analysis," *Knowledge-Based Systems*, vol. 295, p. 111810, Jul. 2024, doi: 10.1016/j.knsys.2024.111810.
- [23] Y. Bi, A. Abrol, Z. Fu, and V. D. Calhoun, "A multimodal vision transformer for interpretable fusion of functional and structural neuroimaging data," *Human brain mapping*, vol. 45, no. 17, p. e26783, Dec. 2024, doi: 10.1002/hbm.26783.
- [24] A. Cosma and E. Radoi, "Learning gait representations with noisy multi-task learning," *Sensors*, vol. 22, no. 18, p. 6803, Sep. 2022, doi: 10.3390/s22186803.
- [25] J. N. Mogan, C. P. Lee, K. M. Lim, and K. S. Muthu, "Gait-ViT: gait recognition with vision transformer," *Sensors*, vol. 22, no. 19, p. 7362, Sep. 2022, doi: 10.3390/s22197362.

- [26] S. Jiao *et al.*, "Gaitformer: a spatial-temporal attention-enhanced network without softmax for Parkinson's disease early detection," *Complex & Intelligent Systems*, vol. 11, no. 5, p. 236, May 2025, doi: 10.1007/s40747-025-01830-y.
- [27] M. Endo, K. L. Poston, E. V. Sullivan, L. Fei-Fei, K. M. Pohl, and E. Adeli, "GaitForeMer: self-supervised pre-training of transformers via human motion forecasting for few-shot gait impairment severity estimation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 13438, Cham, Switzerland: Springer Nature Switzerland, 2022, pp. 130–139, doi: 10.1007/978-3-031-16452-1\_13.
- [28] H. Xu, C. Zhang, Z. Wu, P. Jiao, and H. Wang, "PSGait: gait recognition using parsing skeleton," *arXiv:2503.12047*, 2025, doi: 10.48550/ARXIV.2503.12047.
- [29] J. Zheng, X. Liu, W. Liu, L. He, C. Yan, and T. Mei, "Gait recognition in the wild with dense 3D representations and a benchmark," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 20196–20205, doi: 10.1109/CVPR52688.2022.01959.
- [30] C. Fan, J. Ma, D. Jin, C. Shen, and S. Yu, "SkeletonGait: gait recognition using skeleton maps," *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 2, pp. 1662–1669, Mar. 2024, doi: 10.1609/aaai.v38i2.27933.
- [31] Y. Peng, K. Ma, Y. Zhang, and Z. He, "Learning rich features for gait recognition by integrating skeletons and silhouettes," *Multimedia Tools and Applications*, vol. 83, no. 3, pp. 7273–7294, Jan. 2024, doi: 10.1007/s11042-023-15483-x.
- [32] J. Zheng, X. Liu, S. Wang, L. Wang, C. Yan, and W. Liu, "Parsing is all you need for accurate gait recognition in the wild," in *Proceedings of the 31st ACM International Conference on Multimedia*, Ottawa, ON, Canada, Oct. 2023, pp. 116–124, doi: 10.1145/3581783.3612052.
- [33] Z. Wang, S. Hou, M. Zhang, X. Liu, C. Cao, and Y. Huang, "GaitParsing: human semantic parsing for gait recognition," *IEEE Transactions on Multimedia*, vol. 26, pp. 4736–4748, 2024, doi: 10.1109/TMM.2023.3325962.
- [34] Y. Cui and Y. Kang, "Multi-modal gait recognition via effective spatial-temporal feature fusion," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 17949–17957, doi: 10.1109/CVPR52729.2023.01721.
- [35] S. Hwang *et al.*, "Machine learning based abnormal gait classification with IMU considering joint impairment," *Sensors*, vol. 24, no. 17, p. 5571, Aug. 2024, doi: 10.3390/s24175571.
- [36] O. Dehngani, M. Taherisadr, and R. ChanganVala, "IMU-based gait recognition using convolutional neural networks and multi-sensor fusion," *Sensors*, vol. 17, no. 12, p. 2735, Nov. 2017, doi: 10.3390/s17122735.
- [37] M. D. Nguyen *et al.*, "IMU-based spectrogram approach with deep convolutional neural networks for gait classification," in *2020 IEEE International Conference on Consumer Electronics (ICCE)*, Las Vegas, NV, USA, Jan. 2020, pp. 1–6, doi: 10.1109/ICCE46568.2020.9042999.
- [38] H. Uchitomi, X. Ming, C. Zhao, T. Ogata, and Y. Miyake, "Classification of mild Parkinson's disease: data augmentation of time-series gait data obtained via inertial measurement units," *Scientific reports*, vol. 13, no. 1, p. 12638, Aug. 2023, doi: 10.1038/s41598-023-39862-4.
- [39] M. Sharifi-Renani, M. H. Mahoor, and C. W. Clary, "BioMAT: an open-source biomechanics multi-activity transformer for joint kinematic predictions using wearable sensors," *Sensors*, vol. 23, no. 13, p. 5778, Jun. 2023, doi: 10.3390/s23135778.
- [40] K. Ma, Y. Fu, D. Zheng, Y. Peng, C. Cao, and Y. Huang, "Fine-grained unsupervised domain adaptation for gait recognition," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, Oct. 2023, pp. 11279–11288, doi: 10.1109/ICCV51070.2023.01039.
- [41] L. Palazzo *et al.*, "A deep learning-based framework oriented to pathological gait recognition with inertial sensors," *Sensors*, vol. 25, no. 1, p. 260, Jan. 2025, doi: 10.3390/s25010260.
- [42] D. Robles *et al.*, "Real-time gait pattern classification using artificial neural networks," in *2022 IEEE International Workshop on Metrology for Living Environment (MetroLivEn)*, Cosenza, Italy, May 2022, pp. 76–80, doi: 10.1109/MetroLivEnv54405.2022.9826927.
- [43] A. Chakma, A. Z. M. Faridee, I. Ghosh, and N. Roy, "Domain adaptation for inertial measurement unit-based human activity recognition: a survey," *arXiv:2304.06489*, 2023, doi: 10.48550/ARXIV.2304.06489.
- [44] Z. Lan *et al.*, "Towards a diagnostic tool for neurological gait disorders in childhood combining 3D gait kinematics and deep learning," *Computers in Biology and Medicine*, vol. 171, p. 108095, Mar. 2024, doi: 10.1016/j.compbiomed.2024.108095.
- [45] C. Iseki *et al.*, "Artificial intelligence distinguishes pathological gait: the analysis of markerless motion capture gait data acquired by an iOS application (TDPT-GT)," *Sensors*, vol. 23, no. 13, p. 6217, Jul. 2023, doi: 10.3390/s23136217.
- [46] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE intelligent systems*, vol. 24, no. 2, pp. 8–12, Mar. 2009, doi: 10.1109/MIS.2009.36.
- [47] "CMU motion capture database." [Online]. Available: <https://paperswithcode.com/dataset/cmu-motion-capture>.

## BIOGRAPHIES OF AUTHORS



**Kiran Kumar S V N Madupu**     is a researcher with hands-on experience in IT development and implementation. He is currently working for Guidehouse, USA, as Technical Lead. His research interests include multimodal deep learning, Transformer architectures, and wearable sensing for healthcare applications. He can be contacted at email: [kirankumarsvnmadupu@gmail.com](mailto:kirankumarsvnmadupu@gmail.com).



**Jagadeeshwar Madiraju**    is a professor and dean, Faculty of Engineering and Technology, at Chaitanya Deemed to be University, Hyderabad, India. With 24 years of experience in teaching and research in computer science and engineering, his professional contributions include curriculum design, academic leadership, and research supervision. Five Ph.D. scholars have been awarded degrees under his supervision, and three are currently pursuing doctoral research. His research interests include networking, cloud computing, data science, and sustainable technologies. He has authored three books and published over 30 research papers in peer-reviewed and Scopus-indexed journals. He can be contacted at email: jagadeeshwar07@gmail.com.



**Naren Kodali**    is an academician, researcher, and recognized expert and technology leader in IT security architecture implementation and compliance auditing. He is currently a Professor in Information Security at George Mason University and has served in academic administration as Dean, Computer Information Systems, at Stratford University. His current focus areas of research and teaching are information security policies for large-scale and diverse organizations, and iterative and incremental models of software engineering. He holds a Ph.D. and an M.S. degree, both in computer science, from George Mason University, and two bachelor's degrees in computer science from Andhra University, India. He is a member of ACM, IEEE, IFIP, and SACMAT. He can be contacted at email: nkodali@gmail.com.



**Soma Sekhar Goviraboyina**    is presently working as professor and head of the Department in the CSE-Cyber Security Department at Geethanjali College of Engineering and Technology, Hyderabad, India. He is a senior member of IEEE and a member of the IEEE Standards Committee. He received his Ph.D. in computer science and engineering from Acharya Nagarjuna University, India. He has published 40 research papers in international journals and conferences. His research areas include computer networks and cyber security. He can be contacted at email: somasekharonline@yahoo.co.in.