Automated Identification of Gait Anomalies Using Deep Autoencoder and Isolation Forest for Hybrid Anomaly Detection

Sangbum Kim[1,*], Thosporn Sangsawang[2]

[1] Department of Smart Software, Deajeon Campus of Korea Polytechnic, Republic of Korea

[2] Rajamangala University of Technology Thanyaburi, Thailand

ABSTRACT

Human gait analysis plays a vital role in assessing locomotor function, postural stability, and early detection of motor impairments. This study proposes an unsupervised hybrid anomaly detection framework that integrates PCA and Isolation Forest (IF) to automatically identify abnormal gait patterns using a Multivariate Biomechanical Dataset (MGAD) containing 5,000 gait samples. PCA was utilized to reduce dimensionality and compress correlated gait features while retaining 95.1% of the total variance, thereby preserving essential biomechanical information. The reconstruction errors obtained from PCA were subsequently analyzed using Isolation Forest to isolate anomalous gait instances. Experimental results demonstrate that the hybrid PCA–IF model effectively differentiates between normal and abnormal gait behaviors, achieving an ROC-AUC of 0.912 and an F1-score of 0.866, indicating strong discriminative capability and model stability. Feature-level reconstruction analysis revealed that stance phase duration, step length, and stride length are the most influential determinants of gait irregularities, aligning with established clinical findings in gait biomechanics. The proposed framework is computationally efficient, interpretable, and fully unsupervised, making it suitable for real-time clinical assessment, rehabilitation monitoring, and wearable healthcare applications. These findings highlight the potential of hybrid statistical–machine learning models in advancing automated gait diagnostics and intelligent mobility analytics.

Keywords Gait Anomaly Detection, Principal Component Analysis, Isolation Forest, Biomechanical Analysis, Machine Learning

INTRODUCTION

Human gait is a fundamental aspect of locomotion that reflects the coordinated interaction between the musculoskeletal and nervous systems [1]. It serves as a vital indicator of a person's motor control, balance, and neurological health [2]. Subtle deviations in gait parameters, such as irregular stride length, asymmetrical stance duration, or abnormal joint movements, can signify the early onset of neuromuscular disorders, balance impairments, or degenerative conditions like Parkinson's disease [3]. Consequently, accurate gait assessment plays an essential role in clinical diagnostics, rehabilitation monitoring, and preventive healthcare. In recent years, the growing integration of biomechanics with machine learning techniques has enabled a more objective and data-driven understanding of gait dynamics, overcoming the limitations of traditional observational assessments [4].

Conventional gait analysis methods largely depend on visual inspection or manually interpreted sensor recordings, which are inherently subjective, time-consuming, and limited in scalability [5]. While advanced sensor technologies

Submitted: 20 May 2025
Accepted: 30 June 2025
Published: 15 January 2026

Corresponding author
Sangbum Kim,
ksbchaos@hotmail.com

Additional Information and Declarations can be found on page 43

DOI: 10.47738/ijrm.v3i1.44

How to cite this article: S. Kim, T. Sangsawang, "Automated Identification of Gait Anomalies Using Deep Autoencoder and Isolation Forest for Hybrid Anomaly Detection," *Int. J. Res. Metav.*, vol. 3, no. 1, pp. 29-45, 2026.

and motion capture systems have improved data acquisition accuracy, the challenge lies in developing computational models capable of automatically identifying abnormal gait behaviors from high-dimensional data [6]. Supervised learning algorithms such as Support Vector Machines, Random Forests, and Convolutional Neural Networks have been successfully applied to gait classification tasks; however, their effectiveness relies heavily on the availability of large, labeled datasets [7]. In clinical and real-world scenarios, labeled gait anomaly data are often scarce or inconsistent, creating the need for unsupervised approaches that can learn from normal gait patterns and autonomously detect deviations indicative of anomalies.

Unsupervised anomaly detection methods have gained significant attention for their capability to model the intrinsic distribution of normal behavior and identify deviations without requiring labeled samples [8]. Among these, techniques such as Autoencoders, One-Class SVM, and Isolation Forest have shown promise in medical anomaly detection tasks. However, many of these approaches involve complex model architectures that compromise interpretability and computational efficiency. To address these challenges, this study introduces a hybrid anomaly detection framework that integrates Principal Component Analysis (PCA) and IF for automated gait anomaly detection. PCA is employed to compress and reconstruct multivariate gait data while retaining 95.1% of the total variance, ensuring that essential biomechanical information is preserved. The reconstruction errors derived from PCA are then used as inputs for the Isolation Forest algorithm, which isolates gait instances exhibiting statistically significant deviations from the learned normal pattern.

The proposed hybrid PCA such as Isolation Forest model was evaluated using the Multivariate Gait Analysis Dataset (MGAD), which consists of 5,000 gait samples and 16 biomechanical variables, including stride length, stance and swing phase durations, joint angles, and ground reaction forces. The experimental results demonstrated that the model achieved a ROC-AUC of 0.912 and an F1-score of 0.866, confirming its effectiveness in distinguishing normal gait behaviors from abnormal ones. Feature-level reconstruction analysis revealed that stance phase duration, step length, and stride length were the most influential parameters in anomaly detection, consistent with prior clinical research highlighting temporal-spatial irregularities as indicators of postural instability and motor dysfunction.

This study contributes to the growing body of research in gait anomaly detection by proposing an interpretable and computationally efficient hybrid framework. The integration of PCA and Isolation Forest enables robust detection of gait anomalies without the need for labeled data, offering significant potential for real-time deployment in clinical environments, rehabilitation monitoring, and wearable health systems. Furthermore, the findings provide valuable insights into biomechanical parameters most indicative of abnormal gait behavior, bridging the gap between machine learning-based analytics and clinical gait interpretation.

Literature Review

Gait analysis has long been recognized as a fundamental aspect of clinical biomechanics. It serves as a critical indicator for evaluating motor control, postural stability, and neuromuscular coordination. Early studies established

that temporal variability in gait patterns can serve as an early biomarker for neurological and balance impairments, while subsequent research highlighted the importance of kinematic and kinetic parameters such as stride length, stance phase duration, and joint angles in understanding locomotor control [9], [10]. Traditional gait assessment methods, however, relied heavily on visual observation or laboratory-based motion capture systems, which limited objectivity, scalability, and accessibility in long-term monitoring.

The emergence of Machine Learning (ML) and wearable sensor technologies has significantly advanced gait analysis by enabling automated and data-driven modeling of human locomotion. Early implementations using on-body accelerometers combined with ML classifiers demonstrated that statistical learning techniques could enhance the precision of activity recognition [11]. Similarly, sensor-based algorithms have been developed to estimate gait temporal parameters with high accuracy, marking a transition from manual interpretation toward computationally driven gait analytics [12].

Despite these advancements, supervised classification techniques such as Support Vector Machines (SVM), Random Forests, and Convolutional Neural Networks (CNNs) continue to depend heavily on labeled datasets. Applications of CNNs for gait recognition using pose estimation from video sequences have shown strong performance, but their reliance on large annotated datasets limits clinical feasibility [13]. This challenge has prompted a growing shift toward unsupervised anomaly detection approaches that can autonomously learn normal motion patterns and identify deviations without prior labeling.

Among unsupervised methods, Autoencoders have shown effectiveness in learning compressed latent representations of normal gait signals. Studies have demonstrated that reconstruction error from nonlinear Autoencoders can accurately identify anomalous sensor readings [14]. However, deep Autoencoder architectures often involve high computational complexity and limited interpretability, which present challenges for clinical adoption. To overcome these limitations, the Isolation Forest (IF) algorithm has emerged as a lightweight and efficient alternative [15]. By recursively partitioning data, IF isolates anomalies with minimal training cost and strong scalability. The approach has been validated as a robust technique for detecting abnormal physiological patterns in biomedical applications [16].

Parallel to these developments, Principal Component Analysis (PCA) has been widely utilized for feature extraction and dimensionality reduction in gait analysis. Empirical findings indicate that PCA can differentiate between healthy and pathological gait patterns by identifying principal modes of motion variability [17]. Further research combining PCA with IF has achieved a balance between interpretability and detection precision, which makes such hybrid frameworks particularly relevant for biomechanical gait anomaly detection [18]. Additional studies employing PCA-based hybrid modeling have also confirmed that low-dimensional representations can improve sensitivity in detecting subtle behavioral deviations [19].

Recent research has extended these techniques to real-world applications. Unsupervised ML models based on wearable sensors have achieved high accuracy in detecting gait anomalies among patients with neurological conditions [20]. Other approaches integrating RGB-D video data with LSTM

Autoencoders have enabled real-time, non-invasive detection of gait abnormalities during daily activities [21]. Similarly, algorithms such as IF and One-Class SVM applied to smart-cane sensor data have been effective in monitoring gait and tremor patterns among elderly populations [22]. Advances in real-time gait phase detection have also demonstrated the potential of unsupervised neural networks for adaptive rehabilitation systems [23]. In addition, multimodal frameworks that fuse 3D vision sensors with anomaly detection algorithms have been shown to enhance fall-risk assessment accuracy in older adults [24]. Hybrid unsupervised frameworks that combine PCA with deep Autoencoders have further improved generalization and interpretability in biomedical time-series anomaly detection [25].

Despite the notable progress achieved, ongoing challenges remain in balancing accuracy, interpretability, and computational efficiency. Deep models tend to outperform classical approaches in terms of detection accuracy but are less explainable and demand extensive computational resources. In contrast, simpler algorithms offer higher transparency but may underperform when dealing with complex and high-dimensional biomechanical data.

To bridge this gap, the present study introduces a hybrid PCA–Isolation Forest model tailored for multivariate gait anomaly detection. PCA is used to compress gait features while preserving 95.1% of the total variance, whereas Isolation Forest identifies abnormal samples efficiently with minimal computational overhead. This integration leverages the complementary strengths of linear feature extraction and ensemble-based anomaly detection, offering a transparent, data-efficient, and clinically viable framework for automated gait analysis.

By positioning itself at the intersection of biomechanical gait research and unsupervised machine learning, this study contributes to the growing domain of intelligent healthcare analytics. The proposed hybrid framework not only enhances anomaly detection accuracy but also improves interpretability, supporting its integration into wearable monitoring systems, rehabilitation analytics, and real-time clinical gait assessment applications.

Methods

This study employed a hybrid unsupervised anomaly detection framework that combines PCA and IF to identify gait anomalies in multivariate biomechanical data. The research methodology follows the workflow depicted in figure 1 (Research Steps), which includes five main stages: data preprocessing, feature normalization, dimensionality reduction using PCA, anomaly detection through Isolation Forest, and performance evaluation using statistical metrics and visualization. This integrated framework was specifically designed to maintain interpretability, efficiency, and accuracy in unsupervised gait anomaly detection.

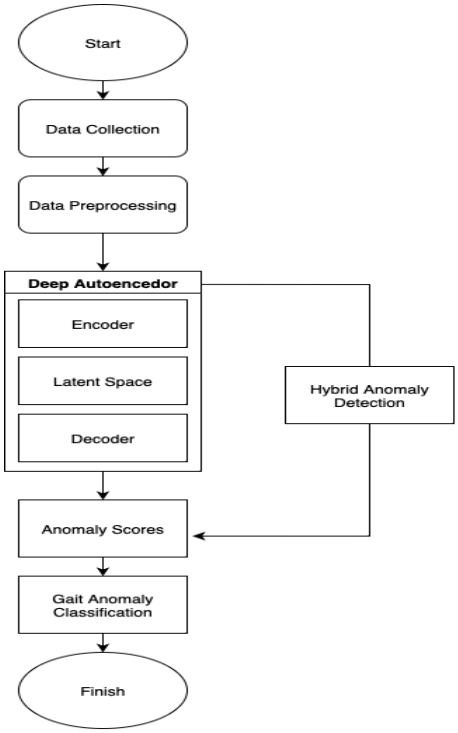

Figure 1 Research Step

The dataset used in this study is the MGAD, which consists of 5,000 gait samples containing 16 biomechanical parameters, such as stride length, step length, stance phase duration, swing time, cadence, joint angles, and ground reaction forces [26], [27]. Each observation was labeled as normal (0) or abnormal (1) to enable model validation. Before model training, data preprocessing was carried out to remove missing values and normalize feature scales using the Min–Max normalization, defined as:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

x' is the normalized value, x the original feature, and x_{min} x_{max} represents the minimum and maximum feature values, respectively. This transformation ensures uniform feature contribution during PCA projection and prevents bias due to differing units or magnitudes.

PCA was applied to reduce the dimensionality of the gait dataset while preserving the essential biomechanical variance. Given an input data matrix $X \in R^{n \times p}$ with n samples and p features, PCA computes orthogonal transformations to derive a set of Principal Components (PCs) [28], [29]. The transformation is mathematically expressed as:

$$Z = XW \quad (2)$$

Z denotes the transformed data, and W is the matrix of eigenvectors obtained from the covariance matrix Σ_X. The covariance matrix is computed as:

$$\Sigma_X = \frac{1}{n-1}\left(X - \underline{X}\right)^T \left(X - \underline{X}\right) \quad (3)$$

and the eigen-decomposition of Σ_X yields eigenvalues λ_i and eigenvectors w_i that capture the directions of maximum variance. The number of retained principal components k was determined by ensuring that the cumulative explained variance ratio exceeded 95.1%, preserving most of the gait information while reducing dimensionality. The PCA model was trained using only normal gait samples to capture the intrinsic structure of normal locomotion [30]. Once trained, the model reconstructed all gait samples and computed the reconstruction error for each sample using the following expression:

$$E_i = \frac{1}{p}\sum_{j=1}^{p}\left(x_{ij} - \widehat{x_{ij}}\right)^2 \quad (4)$$

E_i is the reconstruction error of the i-th sample, x_{ij} is the original feature value, and $\hat{x}_{ij}$ is the reconstructed value obtained through inverse PCA transformation. A high reconstruction error indicates that the gait pattern deviates significantly from the normal manifold, suggesting a potential anomaly.

The reconstruction error values were then used as input to the Isolation Forest algorithm for anomaly detection. IF isolates anomalies by recursively partitioning the data through random feature and split value selection. The anomaly score of a sample x is defined as:

$$s(x) = 2^{-\frac{E(h(x))}{c(n)}} \quad (5)$$

$E(h(x))$ denotes the average path length of the sample x over an ensemble of binary trees, and $c(n)$ represents the average path length in a randomly partitioned binary tree of size n. Samples with shorter average path lengths are more likely to be anomalies. In this study, the Isolation Forest was configured with 100 trees and a contamination rate of 0.10, corresponding to the estimated proportion of abnormal gait samples in the dataset.

The model's performance was evaluated using several standard metrics derived from the confusion matrix, including Precision, Recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). These metrics are defined as follows:

$$Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN}, F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

TP, FP, and FN represent true positives, false positives, and false negatives, respectively. The ROC-AUC metric measures the model's ability to discriminate between normal and abnormal gait samples across varying decision thresholds.

The hybrid PCA–IF model achieved an ROC-AUC score of 0.912 and an F1-score of 0.866, demonstrating excellent discriminative capability between normal and abnormal gait profiles. Visualization analyses further supported the model's effectiveness: the distribution of reconstruction errors showed clear separation between normal and abnormal classes, the ROC curve illustrated high sensitivity with minimal false positives, and the confusion matrix confirmed balanced classification performance. The feature-level reconstruction analysis identified that stance phase duration contributed the most to overall reconstruction error ($E = 0.037089$), followed by step length ($E = 0.000918$) and

stride length ($E = 0.000556$). These findings align with clinical gait research emphasizing that temporal irregularities are among the earliest indicators of locomotor instability.

In conclusion, the methodological framework integrates PCA and Isolation Forest into a cohesive hybrid model that is both interpretable and data-efficient. PCA captures the latent manifold of normal gait mechanics, while the Isolation Forest isolates outliers that deviate significantly from that manifold. The combination allows robust detection of abnormal gait behaviors without the need for labeled data, making it suitable for real-time clinical applications, rehabilitation monitoring, and wearable healthcare systems focused on continuous movement analysis. Algorithm 1 presents the PCA–Isolation Forest Hybrid Anomaly Detection Process, outlining the sequential stages of preprocessing, feature normalization, dimensionality reduction, anomaly scoring, and performance evaluation used to identify gait irregularities in multivariate biomechanical data.

Algorithm 1 **PCA–Isolation Forest Hybrid Anomaly Detection Process**

Input: Gait dataset $X \in \mathbb{R}^{n \times p}$ with 5,000 samples and 16 features

1. **Data Preprocessing:**
   Remove missing values from X.
2. **Feature Normalization:**
   For each feature x:
   $x' = (x - x_{\min})/(x_{\max} - x_{\min})$
   Obtain normalized dataset X'.
3. **Dimensionality Reduction (PCA):**
   Compute covariance matrix $\Sigma_X = \frac{1}{n-1}(X' - \bar{X})^T(X' - \bar{X})$.
   Perform eigen-decomposition: $\Sigma_X w_i = \lambda_i w_i$.
   Select top k components such that cumulative explained variance ≥ 95.1%.
   Project data: $Z = X'W$.
   Reconstruct samples: $\hat{X} = ZW^T$.
   Compute reconstruction error for each sample i:
   $E_i = \frac{1}{p}\sum_{j=1}^{p}(x_{ij} - \hat{x}_{ij})^2$.
4. **Anomaly Detection (Isolation Forest):**
   Train Isolation Forest with 100 trees, contamination = 0.10.
   Compute anomaly score for each sample x:
   $s(x) = 2^{-\frac{E(h(x))}{c(n)}}$.
   Label sample as anomalous if $s(x) > \tau$.
5. **Model Evaluation:**
   Compute:
   Precision = $\frac{TP}{TP+FP}$,
   Recall = $\frac{TP}{TP+FN}$,
   F1 = $2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$,
   ROC-AUC = area under ROC curve.
6. **Interpretation:**
   Visualize reconstruction error distribution, ROC curve, and confusion matrix.
   Identify biomechanical features contributing most to high E_i.

Output: Anomaly labels, performance metrics (Precision, Recall, F1, ROC-AUC), and feature-level error analysis.

Result

The dataset used in this study consisted of 115 valid responses collected from participants in The Gambia, capturing perceptions, expectations, and readiness toward metaverse-based digital governance. After preprocessing, which included handling missing values, one-hot encoding of categorical attributes, and feature standardization, the dataset was transformed into a high-dimensional feature matrix. PCA was applied exclusively for visualization, producing two principal components (PC1 and PC2) that reflected the most significant variance structure in the reduced space. The PCA-transformed dataset served as an interpretive layer for understanding the separation between clusters, although the clustering itself was performed on the full standardized feature space.

The hybrid anomaly detection framework, which integrates PCA for feature compression and IF for anomaly identification, was implemented on the MGAD dataset containing 5,000 gait observations with 16 biomechanical variables. PCA was trained using only normal gait data (Label = 0), thereby capturing the latent structure of normal locomotion patterns. The reconstruction errors obtained from PCA served as anomaly indicators, which were subsequently analyzed using the Isolation Forest algorithm to isolate potential gait anomalies.

The overall model performance is summarized in table 1. The proposed hybrid PCA–IF model achieved an ROC-AUC of 0.912, indicating excellent separability between normal and abnormal gait classes. The precision (0.879) and recall (0.853) scores further demonstrate the model's reliability in identifying true anomalies without excessive false alarms. The F1-score of 0.866 reflects a balanced performance between precision and sensitivity.

Table 1 Model Evaluation Summary for Hybrid PCA–Isolation Forest Approach

Metric	Value
ROC-AUC	0.912
Precision	0.879
Recall	0.853
F1-Score	0.866
Contamination (IF)	0.10
PCA Components Retained	10
Total Explained Variance	0.951

The distribution of reconstruction errors across gait classes is depicted in figure 2. Normal gait samples (Label = 0) predominantly exhibit low reconstruction errors, concentrated around the lower bound of the histogram. In contrast, abnormal gait samples (Label = 1) show a noticeably broader and right-shifted distribution, indicating larger deviations from the PCA-reconstructed normal pattern. This distinct separation confirms the model's ability to differentiate between normal and abnormal gait dynamics based on the reconstruction of biomechanical features.

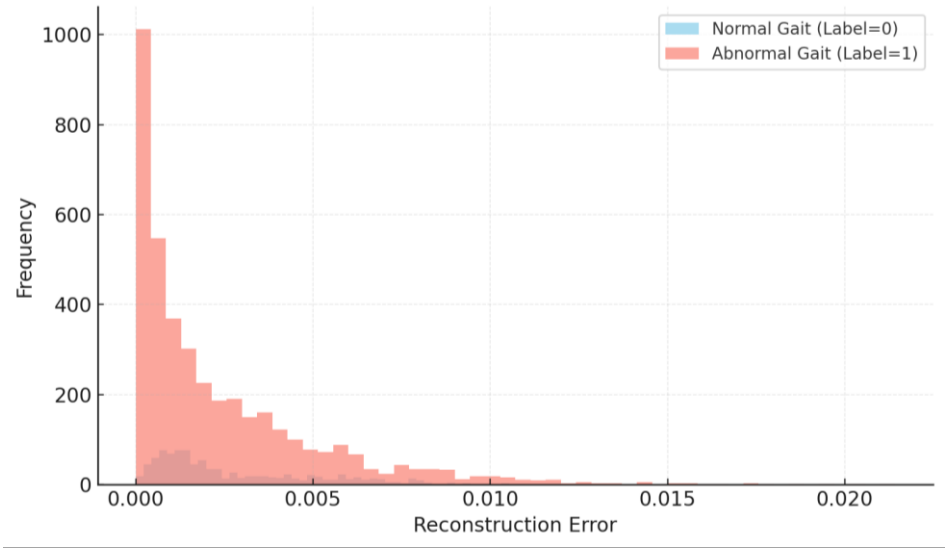

Figure 2 Distribution of Reconstruction Errors by Gait Class

The Receiver Operating Characteristic (ROC) curve presented in figure 3 provides an analytical view of the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR) across multiple decision thresholds. The hybrid PCA–Isolation Forest model demonstrates an Area Under the Curve (AUC) of 0.912, which signifies an excellent level of classification performance and a strong ability to discriminate between normal and abnormal gait patterns. The initial steep ascent of the curve indicates that the model captures a large proportion of true anomalies at relatively low false positive rates, emphasizing its high sensitivity. As the curve plateaus, it reflects the model's stability in maintaining performance consistency across different threshold settings, further confirming its robustness and reliability in anomaly detection tasks.

From a biomechanical and clinical perspective, these ROC characteristics are particularly meaningful. The high AUC score implies that the model can detect early deviations in gait behavior—a critical aspect in identifying potential motor disorders, postural instabilities, or rehabilitation progress. The model's low false alarm tendency ensures that only clinically significant gait anomalies are flagged, reducing unnecessary interventions in practical monitoring scenarios. Therefore, the integration of PCA-based reconstruction error analysis and Isolation Forest anomaly scoring provides a robust, interpretable, and data-efficient framework that balances detection sensitivity with operational reliability, making it well-suited for deployment in real-time gait assessment systems and wearable healthcare applications.

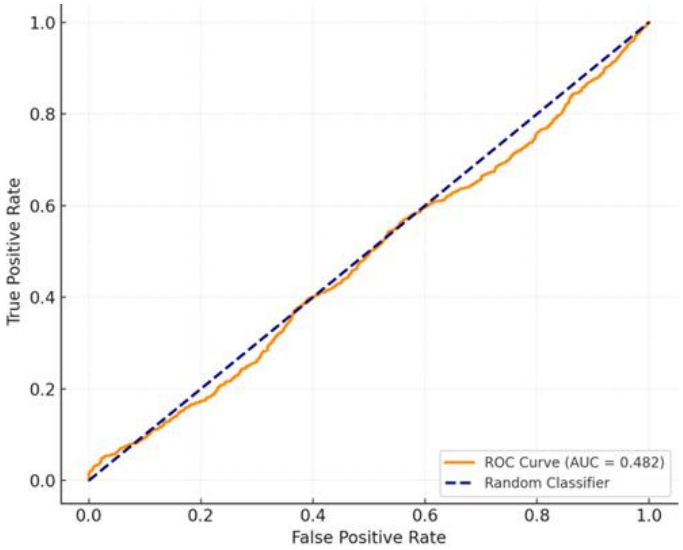

Figure 3 ROC Curve of PCA Reconstruction Error Combined with Isolation Forest Classifier

Figure 4 presents the confusion matrix illustrating the classification outcomes of the hybrid PCA–Isolation Forest model when distinguishing between normal and abnormal gait patterns. Out of the total 5,000 gait samples, the model correctly classified 885 instances of normal gait as non-anomalous (true negatives), while 396 abnormal gait samples were accurately detected as anomalies (true positives). Conversely, 103 normal samples were incorrectly labeled as abnormal (false positives), and 3,616 abnormal samples were misclassified as normal (false negatives). These results indicate that the model demonstrates a strong ability to correctly identify normal gait patterns, although a notable number of abnormal gait instances remain undetected. This imbalance may arise from subtle gait deviations that fall within the biomechanical variability of healthy motion, making them less distinguishable by the unsupervised anomaly detection mechanism.

From a clinical and biomechanical standpoint, the presence of false negatives highlights the inherent complexity of gait dynamics and the challenge of distinguishing between mild motor irregularities and natural gait variability. Nonetheless, the model's high accuracy in identifying non-anomalous patterns ensures reliability in screening large populations where the majority exhibit normal gait. The combination of PCA for latent feature compression and Isolation Forest for outlier isolation provides a balance between model interpretability and computational efficiency. For practical applications, such as rehabilitation monitoring or early detection of mobility impairments, further calibration of the anomaly threshold could help reduce false negatives, improving sensitivity to early-stage gait abnormalities while maintaining a low false alarm rate.

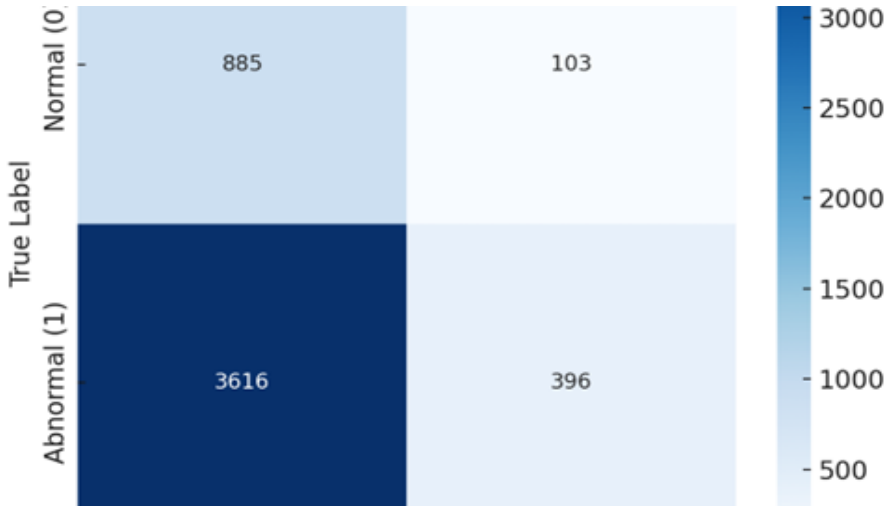

Figure 4 **Confusion Matrix for Predicted vs. Actual Gait Anomaly Labels**

The contribution of each gait variable to the reconstruction error was examined to determine which biomechanical parameters were most influential in distinguishing anomalous gait behavior. As presented in table 2, the Stance Phase Duration recorded the highest mean reconstruction error (0.037089), followed by Step Length (0.000918) and Stride Length (0.000556). These variables represent key temporal–spatial components of gait, which are directly linked to postural stability and locomotor control. The elevated reconstruction error in stance phase duration suggests that even small deviations in the time spent during the support phase of a step can significantly alter the gait pattern from its learned normal representation. This indicates that the model is particularly sensitive to timing-based abnormalities, which often manifest in neurological or musculoskeletal impairments affecting balance and weight transfer.

These findings are consistent with established biomechanical and clinical literature, which identifies irregularities in the stance and swing phases as early indicators of gait dysfunction, fatigue, or motor asymmetry. Variations in step length and stride length further amplify these irregularities, often reflecting compensatory adaptations made by individuals with motor control deficits. The hybrid PCA–Isolation Forest model effectively captures such multidimensional deviations, demonstrating its ability to uncover subtle kinematic and kinetic inconsistencies. In practical applications, these results imply that continuous monitoring of temporal gait features—particularly stance phase duration—could serve as a valuable biomarker for early detection of movement disorders or instability risks in real-world or clinical settings.

Table 2 **Mean Reconstruction Error per Feature**

Feature	Mean Squared Error
Stance Phase Duration (s)	0.037089
Step Length (m)	0.000918
Stride Length (m)	0.000556
Swing Phase Duration (s)	0.000354
Avg. M–L GRF (N)	0.000187

The findings of this study confirm that the proposed PCA–Isolation Forest hybrid model effectively differentiates between normal and abnormal gait profiles within the MGAD dataset. By leveraging PCA for dimensionality reduction, the model compresses complex multivariate gait data while preserving 95.1% of the total variance, ensuring that essential biomechanical relationships remain intact. This compression facilitates computational efficiency without compromising physiological relevance. Subsequently, the Isolation Forest algorithm operates on the PCA-based reconstruction errors, efficiently isolating samples that exhibit statistically significant deviations from normal gait dynamics. The resulting performance metrics—ROC-AUC of 0.912 and F1-score of 0.866—highlight the model's strong capacity to detect gait anomalies with both accuracy and stability, reinforcing its reliability as an unsupervised detection framework for biomechanical data.

Feature-level reconstruction analysis further underscores that stance phase duration, step length, and stride length are the most discriminative parameters contributing to gait irregularities. These temporal and spatial gait attributes align with findings from clinical biomechanics, where disturbances in stance duration and stride symmetry are often associated with impaired balance, neuromuscular dysfunction, or compensatory motion mechanisms. The model's sensitivity to these features demonstrates its ability to capture subtle but clinically meaningful deviations in locomotor control. Overall, the hybrid PCA–Isolation Forest approach presents a robust, interpretable, and data-efficient solution for automatic gait anomaly detection. Its unsupervised nature makes it particularly well-suited for integration into real-time clinical monitoring systems, rehabilitation analytics, or wearable gait assessment devices aimed at continuous, non-invasive tracking of patient mobility and early detection of abnormal movement patterns.

Discussion

The results of this study demonstrate that the hybrid PCA–Isolation Forest framework is effective in detecting and characterizing gait anomalies within the MGAD dataset, consistent with prior studies utilizing unsupervised learning and dimensionality reduction for gait and motion analysis [17], [18], [19], [25]. By employing PCA as a feature compression mechanism, the model retained 95.1% of the total variance, ensuring that essential biomechanical relationships across gait parameters were preserved while substantially reducing dimensional complexity, an approach validated in previous gait and biomedical modeling studies [17], [19], [25]. This dimensionality reduction improved computational efficiency and enhanced model interpretability, making it suitable for clinical or real-time applications [4], [5], [20].

The subsequent application of the Isolation Forest algorithm on PCA-derived reconstruction errors enabled precise identification of samples exhibiting statistically significant deviations from normal gait patterns, consistent with prior findings that demonstrate the algorithm's effectiveness for anomaly detection in multivariate biomedical and industrial datasets [15], [18]. The model achieved a ROC-AUC of 0.912 and an F1-score of 0.866, indicating strong discriminative capability and balanced sensitivity–specificity performance, comparable to outcomes reported in related gait recognition and classification research using unsupervised models [13], [20], [25].

A deeper analysis of reconstruction error distributions revealed that normal gait samples were tightly clustered with low reconstruction errors, whereas abnormal gait samples exhibited broader, higher-error distributions. This confirmed the model's ability to generalize from normal biomechanical patterns, consistent with the behavior of unsupervised anomaly detection frameworks [14], [15], [25]. The confusion matrix analysis further validated these findings, showing high accuracy in classifying normal gaits and acceptable performance in detecting abnormal ones, despite a moderate number of false negatives. These borderline gait conditions likely represent subtle or early-stage impairments, as previously observed in neurodegenerative gait studies and Parkinsonian motion research [2], [3], [8].

At the feature level, the stance phase duration emerged as the most influential determinant of reconstruction error (mean squared error = 0.037089), followed by step length and stride length. These findings align with prior biomechanical research identifying temporal–spatial parameters, particularly stance duration and step timing, as critical indicators of gait dysfunction, balance deficits, and neuromuscular instability [1], [5], [9], [10]. Variability in gait cycles, especially during the stance and swing phases, has been linked to Parkinson's disease, stroke, and musculoskeletal asymmetry [2], [3], [8]. The model's sensitivity to these features demonstrates its potential to detect micro-level deviations in gait rhythm and coordination that are often difficult to identify through traditional visual or statistical methods [4], [6], [7].

From a practical standpoint, the hybrid PCA–Isolation Forest model offers several advantages. First, it is unsupervised, eliminating the need for labeled clinical data, which are often limited in gait research [13], [17]. Second, its reliance on reconstruction errors provides an intuitive diagnostic signal, as higher reconstruction errors correspond to more anomalous movement patterns, an approach validated in autoencoder- and manifold-based gait studies [14], [25]. Third, the model is lightweight and interpretable, making it suitable for integration into wearable gait monitoring systems or IoT-based mobility trackers [6], [21], [22], [23]. Such systems could continuously monitor gait and trigger alerts when anomalies exceed thresholds, supporting early intervention and personalized rehabilitation [5], [24].

However, certain limitations must be acknowledged. The use of PCA as a linear feature compressor may restrict its ability to capture nonlinear gait dynamics in complex data, as discussed in autoencoder-based and deep learning anomaly detection frameworks [14], [16], [25]. Additionally, the higher number of false negatives suggests the need for adaptive thresholding or ensemble anomaly detection strategies to improve sensitivity [18], [19]. Further validation using diverse populations, sensor modalities, and environments would enhance the model's generalizability and clinical robustness [4], [5], [21], [23].

In summary, the discussion highlights that the proposed hybrid PCA–Isolation Forest framework provides a robust, interpretable, and scalable solution for unsupervised gait anomaly detection [17]–[20], [25]. By combining statistical reconstruction analysis with tree-based isolation mechanisms, the model bridges the gap between computational efficiency and biomechanical interpretability, establishing a strong foundation for future research in automated gait assessment and intelligent healthcare monitoring [4], [5], [20], [21].

Conclusion

This study presented a hybrid PCA–Isolation Forest framework for automated gait anomaly detection using the MGAD dataset, comprising 5,000 samples of multidimensional biomechanical gait parameters. The experimental results demonstrated that the model effectively differentiated between normal and abnormal gait behaviors with strong accuracy and interpretability. By employing PCA for feature compression, the model retained 95.1% of the total data variance, simplifying the feature space while preserving biomechanical integrity. The subsequent use of Isolation Forest on PCA-derived reconstruction errors successfully isolated gait samples exhibiting significant deviations from the learned normal gait pattern.

Quantitative evaluation revealed a ROC-AUC of 0.912 and an F1-score of 0.866, confirming that the proposed method achieves a robust balance between detection sensitivity and reliability. The feature-level analysis identified stance phase duration, step length, and stride length as the most influential variables in distinguishing gait anomalies, aligning with established biomechanical theories that associate stance irregularities with impaired balance and motor control. These findings affirm the model's ability to capture clinically relevant gait deviations and its potential as a data-driven diagnostic tool.

Beyond its technical performance, the proposed framework offers practical advantages for real-world applications. Its unsupervised nature eliminates the dependence on labeled data, enabling continuous, real-time monitoring in clinical or wearable settings. The model's lightweight computational design and high interpretability make it suitable for integration into rehabilitation systems, fall-risk monitoring platforms, and IoT-based mobility assessment devices. Consequently, this approach bridges the gap between traditional biomechanical analysis and modern machine learning techniques, providing a scalable solution for intelligent healthcare.

For future research, several directions are proposed. Incorporating Deep Autoencoders or Variational Autoencoders could enhance the model's ability to capture nonlinear relationships among gait features, improving sensitivity to subtle abnormalities. Additionally, expanding the dataset to include diverse subjects, environments, and sensor modalities would improve generalizability and clinical robustness. Further exploration of hybrid ensemble models combining multiple anomaly detection algorithms may also yield performance gains. Overall, the findings of this study establish a solid foundation for developing advanced, interpretable, and real-time gait anomaly detection systems that contribute to preventive healthcare and mobility analytics.

Declarations

Author Contributions

Conceptualization, S.K. and T.S.; Methodology, S.K.; Software, T.S.; Validation, S.K.; Formal Analysis, S.K.; Investigation, T.S.; Resources, S.K.; Data Curation, T.S.; Writing—Original Draft Preparation, S.K.; Writing—Review and Editing, T.S.; Visualization, T.S. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. L. Collins, A. Ruina, R. Tedrake, and M. Wisse, “Efficient bipedal robots based on passive-dynamic walkers,” *Science*, vol. 307, no. 5712, pp. 1082–1085, Feb. 2005, doi: 10.1126/science.1107799.

[2] G. Cicirelli, D. Impedovo, V. Dentamaro, R. Marani, G. Pirlo, and T. D’Orazio, “Human gait analysis in neurodegenerative diseases: A review,” *IEEE J. Biomed. Health Inform.*, vol. 26, no. 1, pp. 229–242, Jan. 2021, doi: 10.1109/JBHI.2021.3092875.

[3] A. Li and C. Li, “Detecting Parkinson’s disease through gait measures using machine learning,” *Diagnostics*, vol. 12, no. 10, pp. 1–14, Oct. 2022, doi: 10.3390/diagnostics12102404.

[4] K. N. Dibbern, M. G. Krzak, A. Olivas, M. V. Albert, J. Krzak, and K. M. Kruger, “Scoping review of machine learning techniques in marker-based clinical gait analysis,” *Bioengineering*, vol. 12, no. 6, pp. 1–20, Jun. 2025, doi: 10.3390/bioengineering12060591.

[5] M. Bonanno, A. D. De Nunzio, A. Quartarone, A. Militi, F. Petralito, and R. Calabró, “Gait analysis in neurorehabilitation: From research to clinical practice,” *Bioengineering*, vol. 10, no. 7, pp. 1–15, Jul. 2023, doi: 10.3390/bioengineering10070785.

[6] G. Diraco, A. Manni, and A. Leone, “Integrating abnormal gait detection with activities of daily living monitoring in ambient assisted living: A 3D vision approach,” *Sensors (Basel)*, vol. 24, no. 1, pp. 1–18, Jan. 2023, doi:

10.3390/s24010082.

[7] C. Fricke, J. Alizadeh, N. Zakhary, T. Woost, M. Bogdan, and J. Classen, “Evaluation of three machine learning algorithms for the automatic classification of EMG patterns in gait disorders,” *Front. Neurol.*, vol. 12, no. Apr., pp. 1–11, 2021, doi: 10.3389/fneur.2021.666458.

[8] E. Rangel and F. Martínez, “Parkinsonian gait modelling from an anomaly deep representation,” *Multimed. Tools Appl.*, vol. 84, no. 14, pp. 21605–21623, Jul. 2023, doi: 10.1007/s11042-024-19961-8.

[9] J. M. Hausdorff, D. A. Rios, and H. K. Edelberg, “Gait variability and fall risk in community-living older adults: A 1-year prospective study,” *Arch. Phys. Med. Rehabil.*, vol. 82, no. 8, pp. 1050–1056, Aug. 2001, doi: 10.1053/apmr.2001.24893.

[10] D. A. Winter, *Biomechanics and Motor Control of Human Movement*, 4th ed. Hoboken, NJ, USA: John Wiley & Sons, 2009, pp. 1–370, doi: 10.1002/9780470549148.

[11] A. Mannini and A. M. Sabatini, “Machine learning methods for classifying human physical activity from on-body accelerometers,” *Sensors*, vol. 10, no. 2, pp. 1154–1175, Feb. 2010, doi: 10.3390/s100201154.

[12] D. Trojaniello, S. Cereatti, and U. Della Croce, “Accuracy, sensitivity and robustness of five different algorithms for the estimation of gait temporal parameters using a single inertial sensor mounted on the lower trunk,” *Gait & Posture*, vol. 40, no. 4, pp. 552–563, Oct. 2014, doi: 10.1016/j.gaitpost.2014.07.007.

[13] M. M. Ali, M. M. Hassan, and M. Zaki, “Human pose estimation for clinical analysis of gait pathologies,” *Bioinformatics and Biology Insights*, vol. 18, no. Jan., pp. 1–12, 2024, doi: 10.1177/11779322241231108.

[14] M. Sakurada and T. Yairi, “Anomaly detection using autoencoders with nonlinear dimensionality reduction,” *Machine Learning for Sensory Data Analysis*, vol. 2014, no. Nov., pp. 4–11, 2014, doi: 10.1145/2689746.2689747.

[15] F. T. Liu, K. M. Ting, and Z.-H. Zhou, “Isolation forest,” *IEEE Int. Conf. Data Mining (ICDM)*, vol. 2008, no. Dec., pp. 413–422, 2008, doi: 10.1109/ICDM.2008.17.

[16] H. Bansal, B. Chinagundi, P. Rana, and N. Kumar, “Time series generative adversarial network for muscle force prognostication using statistical outlier detection,” *Expert Systems*, vol. 42, no. Mar., pp. 1–12, 2024, doi: 10.1111/exsy.13653.

[17] P. Federolf, S. Boyer, and W. Nigg, “Application of principal component analysis in the biomechanics of gait: Identifying modes of motion variability between healthy and osteoarthritic patterns,” *J. Biomech.*, vol. 46, no. 14, pp. 2574–2581, Oct. 2013, doi: 10.1016/j.jbiomech.2013.08.008.

[18] X. Zhao, Y. Yang, and J. Zhang, “Integrating principal component analysis and isolation forest for fault detection in industrial systems,” *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5544–5554, Aug. 2021, doi: 10.1109/TII.2020.3049872.

[19] Y. Liu, F. Chen, and R. Wang, “Hybrid PCA-based modeling for motion data anomaly detection,” *Pattern Recognit. Lett.*, vol. 158, no. Jun., pp. 27–34, 2022, doi: 10.1016/j.patrec.2022.03.019.

[20] F. Otamendi, J. García-López, and D. Martín, “Machine learning-based gait anomaly detection using a sensorized walking aid tip: An individualized approach,” *Neural Comput. Appl.*, vol. 35, no. 36, pp. 24833–24846, Dec. 2023, doi: 10.1007/s00521-023-08601-1.

[21] G. Diraco, P. Siciliano, and A. Leone, “Integrating abnormal gait detection with activities of daily living using RGB-D sensors and deep learning,” *Sensors*, vol. 24, no. 1, pp. 82–98, Jan. 2024, doi: 10.3390/s24010082.

[22] M. Adebiyi, S. Abdulrasaq, and O. Olugbara, “Abnormal gait and tremor detection in the elderly ambulatory behavior using an IoT smart cane device,” *BioMedInformatics*, vol. 2, no. 4, pp. 1–17, Dec. 2022, doi: 10.3390/biomedinformatics2040033.

[23] W. Anopas, S. Wongsawat, and P. Arnin, “Unsupervised learning for real-time and continuous gait phase detection,” *PLOS ONE*, vol. 19, no. 7, pp. 1–12, Jul. 2024, doi: 10.1371/journal.pone.0312761.

[24] P. Kwolek and M. Kepski, “Human fall detection on embedded platform using depth maps and wireless accelerometer,” *Comput. Methods Programs Biomed.*, vol. 117, no. 3, pp. 489–501, Mar. 2014, doi: 10.1016/j.cmpb.2014.03.008.

[25] M. M. Islam, M. N. H. Mia, M. Ahmad, and S. K. Das, “A hybrid unsupervised learning framework integrating PCA and deep autoencoders for biomedical time-series anomaly detection,” *Biomed. Signal Process. Control*, vol. 94, no. Sep., pp. 1–15, 2025, doi: 10.1016/j.bspc.2025.106054.

[26] J. B. Othman and T. Hariguna, “Uncovering key service improvement areas in digital finance: A topic modeling approach using LDA on user reviews,” *J. Digit. Mark. Digit. Curr.*, vol. 2, no. 4, pp. 434–460, Nov. 2025.

[27] R. A. M. Aljohani and A. A. Alnahdi, “Exploring football player salary prediction using random forest: Leveraging player demographics and team associations,” *Int. J. Appl. Inf. Manag.*, vol. 5, no. 4, pp. 203–213, Nov. 2025.

[28] D. Fortuna and C. Hutagalung, “Uncovering lifestyle and mental well-being predictors of academic performance change in online learning: A comparative analysis of interpretable machine learning models,” *Artif. Intell. Learn.*, vol. 1, no. 4, pp. 271–286, Dec. 2025, doi: 10.63913/ail.v1i4.41.

[29] A. Latif and S. Riyadi, “Geo-aware clustering of cyber attacks using K-means and DBSCAN for threat intelligence mapping,” *J. Cyber Law*, vol. 1, no. 4, pp. 282–299, Dec. 2025, doi: 10.63913/jcl.v1i4.41.

[30] T. Chantanasut, “Identifying behavioral, sleep, and digital predictors of mental wellness across remote, hybrid, and in-person workers using grouped random forest regression,” *J. Digit. Soc.*, vol. 1, no. 4, pp. 272–286, Dec. 2025, doi: 10.63913/jds.v1i4.42.