Latent Motion: Uncovering Biomechanical Patterns with Sparse Autoencoders

Abstract

This study explores the application of sparse autoencoders (SAEs) as interpretability tools for gait kinematics, leveraging the comprehensive Gutenberg Gait Database of healthy individuals. We implemented and compared three distinct autoencoder architectures: a baseline sparse autoencoder, an SAE with disentanglement loss, and a sparse variational autoencoder (SVAE). These models were trained on ground reaction force (GRF) and center of pressure (COP) data processed using wavelet decomposition to learn meaningful latent representations of gait patterns. Our analysis reveals that the learned features capture biomechanically significant aspects of human gait, with clear correlations to physical attributes such as age, height, body weight, and walking speed. The latent space organization demonstrates a continuous manifold structure that effectively encodes gait dynamics, with walking speed emerging as a particularly influential factor. Clustering analysis of the latent representations identified distinct groupings that may correspond to different gait styles or phases. While all three models successfully learned interpretable features, we observed potential redundancy in the feature reconstructions that warrants further investigation. This research serves as a proof of concept for developing a shared latent space and comprehensive dictionary of gait features, which could significantly enhance biomechanical analysis, aid in the identification of gait abnormalities, and support clinical decision-making. Our findings demonstrate the potential of sparse autoencoders as powerful tools for uncovering interpretable biomechanical patterns in high-dimensional gait data.

1. Introduction

Human gait analysis represents a fundamental area of biomechanical research with significant implications for clinical assessment, rehabilitation, and sports performance. The complex, multi-dimensional nature of gait patterns encapsulates a wealth of information about an individual's physical condition, age-related changes, potential pathologies, and overall functional status. However, the interpretation of high-

dimensional gait data presents substantial challenges for researchers and clinicians alike, often requiring specialized expertise and considerable time investment.

Traditional approaches to gait analysis have relied heavily on direct measurement and statistical analysis of specific parameters such as step length, cadence, joint angles, and ground reaction forces. While these methods provide valuable insights, they frequently reduce the rich complexity of gait dynamics to a limited set of predefined metrics, potentially overlooking subtle patterns and interactions that may have significant biomechanical or clinical relevance. This reduction in dimensionality, while making analysis more tractable, can limit our understanding of the full spectrum of gait characteristics and their relationships to physical attributes and pathological conditions.

The emergence of machine learning techniques offers promising avenues for addressing these limitations by automatically extracting meaningful patterns from high-dimensional gait data. Supervised learning approaches have demonstrated success in classifying gait patterns associated with specific conditions or demographics. However, these methods typically require labeled data and may not provide interpretable insights into the underlying biomechanical features driving the classifications. In contrast, unsupervised learning approaches, particularly autoencoders, offer the potential to learn rich representations of gait patterns without requiring labeled data, while potentially revealing interpretable features that correspond to meaningful biomechanical characteristics.

Sparse autoencoders (SAEs) represent a particularly promising class of unsupervised learning models for gait analysis. By imposing sparsity constraints on the learned representations, these models are encouraged to discover more interpretable features that capture essential aspects of the input data. This sparsity property aligns well with the biomechanical intuition that gait patterns, while complex, may be decomposable into a set of fundamental movement primitives or characteristic patterns. Furthermore, the interpretability of sparse representations makes them especially valuable in clinical contexts, where understanding the basis for model predictions is crucial for informing treatment decisions.

In this study, we explore the application of sparse autoencoders as interpretability tools for gait kinematics, leveraging the comprehensive Gutenberg Gait Database of healthy individuals. We implement and compare three distinct autoencoder architectures:

1. A baseline sparse autoencoder (SAE) with L1 regularization and KL divergence sparsity penalties
2. An SAE with additional disentanglement loss to encourage independence between learned features

3. A sparse variational autoencoder (SVAE) that combines sparsity constraints with a probabilistic latent space

These models were trained on ground reaction force (GRF) and center of pressure (COP) data from the Gutenberg Gait Database, which provides a rich collection of gait measurements from 350 healthy individuals across various ages, heights, weights, and walking speeds. By analyzing the learned representations and their relationships to subject metadata, we aim to uncover interpretable biomechanical patterns that could enhance our understanding of human gait dynamics and potentially serve as a foundation for clinical applications.

Our research makes several key contributions to the field of biomechanical analysis:

1. We demonstrate the effectiveness of sparse autoencoders in learning interpretable features from high-dimensional gait data without explicit supervision.
2. We provide a comparative analysis of different autoencoder architectures for gait feature learning, highlighting their respective strengths and limitations.
3. We establish a proof of concept for a shared latent space that could serve as the basis for a comprehensive dictionary of gait features with additional data.
4. We identify correlations between learned features and physical attributes, suggesting potential applications in personalized biomechanical analysis.

The remainder of this paper is organized as follows: Section 2 provides a comprehensive literature review covering biomechanical modeling, machine learning in gait analysis, sparse autoencoders, and interpretability approaches. Section 3 details our methodology, including dataset description, model architectures, training procedures, and evaluation methods. Section 4 presents our results, analyzing the learned features, latent space organization, metadata correlations, and comparative model performance. Section 5 discusses the implications of our findings, limitations of our approach, and connections to existing literature. Section 6 outlines directions for future work, and Section 7 concludes with a summary of our contributions and their significance for biomechanical analysis.

2. Literature Review

2.1 Biomechanical Modeling and Gait Analysis

Gait analysis has been a cornerstone of biomechanical research for decades, providing valuable insights into human locomotion patterns in both healthy individuals and those with pathological conditions. Traditional approaches to gait analysis have relied on a combination of observational assessments and quantitative measurements of spatiotemporal parameters, kinetics, and kinematics (Baker, 2006). These methods

typically involve specialized equipment such as force plates, motion capture systems, and electromyography to capture the complex dynamics of human movement during walking or running (Whittle, 2014).

The development of standardized gait databases has significantly advanced the field by providing researchers with comprehensive datasets for analysis and model development. The AddBiomechanics dataset, one of the early comprehensive collections, offered researchers access to standardized biomechanical measurements across diverse populations (Horst et al., 2019). Building upon this foundation, the GaitRec dataset expanded the available data with a particular focus on pathological gait patterns, including measurements from individuals with various neurological and musculoskeletal conditions (Horsak et al., 2020).

The Gutenberg Gait Database, introduced by Schöllhorn et al. (2021), represents a significant advancement in this domain, providing what is currently the world's largest collection of gait analysis data from healthy individuals. This database contains ground reaction force (GRF) and center of pressure (COP) data from 350 healthy subjects, complementing the GaitRec dataset by increasing the number of healthy control subjects from 211 to 561. The Gutenberg Gait Database follows standardized protocols and file formats compatible with the GaitRec dataset, facilitating integrated analysis across both resources. The comprehensive nature of this database, with its rich metadata including age, height, weight, and walking speed information, makes it an ideal resource for developing and validating advanced analytical methods for gait analysis (Schöllhorn et al., 2021).

2.2 Machine Learning in Biomechanics

The application of machine learning techniques to biomechanical data has grown substantially in recent years, offering new approaches to analyze and interpret the complex, high-dimensional data characteristic of human movement (Halilaj et al., 2018). Supervised learning approaches have demonstrated considerable success in classifying gait patterns associated with specific conditions or demographics. For instance, Begg et al. (2005) employed support vector machines to distinguish between the gait patterns of young and elderly individuals, while Alaqtash et al. (2011) used neural networks to classify pathological gait patterns in individuals with neuromuscular disorders.

Unsupervised learning approaches have gained increasing attention for their ability to discover latent patterns in gait data without requiring labeled examples. Clustering techniques such as k-means and hierarchical clustering have been applied to identify natural groupings in gait patterns (Toro et al., 2007), while dimensionality reduction methods like principal component analysis (PCA) have been used to identify the primary modes of variation in gait data (Deluzio & Astephen, 2007). These approaches have

provided valuable insights into the underlying structure of gait patterns but often lack the representational capacity to capture the full complexity of biomechanical data.

Deep learning methods, particularly autoencoders, have emerged as powerful tools for learning rich representations of biomechanical data. Horst et al. (2019) demonstrated the effectiveness of convolutional autoencoders in learning meaningful features from raw gait data, while Dindorf et al. (2022) explored the use of variational autoencoders for generating synthetic biomechanical data. These approaches leverage the representational power of deep neural networks to capture complex patterns in gait data, potentially revealing insights that might be missed by traditional analytical methods.

2.3 Sparse Autoencoders

Sparse autoencoders represent a specialized class of neural network architectures designed to learn compact, interpretable representations of high-dimensional data (Ng, 2011). Unlike standard autoencoders, which simply aim to reconstruct their inputs from a lower-dimensional latent space, sparse autoencoders incorporate additional constraints that encourage the learned representations to be sparse—that is, to have relatively few active neurons for any given input. This sparsity property is motivated by observations from neuroscience suggesting that biological neural systems often employ sparse coding strategies, with relatively few neurons firing in response to any particular stimulus (Olshausen & Field, 1996).

Mathematically, a sparse autoencoder consists of an encoder function that maps input data x to a latent representation h = f(x), and a decoder function that reconstructs the input from this representation $\hat{x} = g(h)$. The training objective typically combines a reconstruction loss term with a sparsity penalty:

$$L(x, \hat{x}) = ||x - \hat{x}||^2 + \lambda \cdot \Omega(h)$$

where Ω(h) is a sparsity-inducing regularizer such as the L1 norm ($||h||_1$) or the Kullback-Leibler divergence between the average activation of each hidden unit and a target sparsity level ρ (Ng, 2011). This formulation encourages the model to learn a representation that both accurately reconstructs the input data and satisfies the sparsity constraint.

The application of sparse autoencoders in biomechanics has been relatively limited compared to other domains such as computer vision and natural language processing. However, their potential for learning interpretable features makes them particularly well-suited for biomechanical analysis, where understanding the underlying patterns is often as important as predictive performance. Recent work by Dindorf et al. (2021) has demonstrated the effectiveness of sparse coding techniques in identifying interpretable

movement primitives from biomechanical data, suggesting that sparse autoencoders could provide valuable insights into the fundamental components of human gait.

2.4 Disentanglement in Representation Learning

Disentanglement has emerged as an important concept in representation learning, referring to the separation of distinct, informative factors of variation in the data (Bengio et al., 2013). A disentangled representation is one in which individual latent dimensions correspond to meaningful, independent factors in the data-generating process. For instance, in the context of gait analysis, a disentangled representation might include separate dimensions for factors such as walking speed, stride length, and joint angles, allowing for independent manipulation and analysis of these factors.

Various approaches have been proposed to encourage disentanglement in learned representations. One common strategy involves adding regularization terms to the training objective that penalize statistical dependencies between different latent dimensions. For example, the β-VAE framework (Higgins et al., 2017) modifies the standard variational autoencoder objective by increasing the weight of the Kullback-Leibler divergence term, which encourages the learned latent distribution to be close to a factorized prior. Other approaches include the Total Correlation penalty (Chen et al., 2018), which explicitly minimizes the mutual information between latent dimensions, and adversarial methods that use discriminator networks to enforce independence (Mathieu et al., 2016).

Evaluating the quality of disentanglement in learned representations remains a challenging problem. Metrics such as the Disentanglement, Completeness, and Informativeness framework (Eastwood & Williams, 2018) and the β-VAE Disentanglement metric (Higgins et al., 2017) have been proposed, but these typically require access to the true generative factors, which are rarely available in real-world datasets. In the context of biomechanical data, where the underlying factors of variation are complex and often unknown, alternative evaluation strategies based on correlations with known physical attributes or expert assessment may be more appropriate.

The application of disentanglement techniques to biomechanical data represents a promising direction for enhancing the interpretability of learned representations. By separating the influence of different physical and biomechanical factors, disentangled representations could provide more nuanced insights into the complex interplay of variables that determine gait patterns, potentially leading to more targeted interventions and personalized treatment strategies.

2.5 Interpretability in Biomechanical Models

Interpretability in biomechanical models is of paramount importance, particularly in clinical applications where understanding the basis for model predictions can directly impact treatment decisions (Halilaj et al., 2018). Traditional biomechanical models, such as inverse dynamics approaches and musculoskeletal simulations, offer inherent interpretability through their basis in physical principles and anatomical structures (Delp et al., 2007). However, these models often require simplifying assumptions and may not fully capture the complexity of human movement patterns.

Machine learning approaches, while potentially more powerful in capturing complex patterns, often sacrifice interpretability for predictive performance. This "black box" nature can limit their utility in clinical settings, where clinicians need to understand not just what a model predicts but why it makes that prediction (Doshi-Velez & Kim, 2017). Various strategies have been proposed to enhance the interpretability of machine learning models in biomechanics, including feature importance analysis, partial dependence plots, and attention mechanisms (Horst et al., 2019).

Sparse and disentangled representations offer a promising middle ground, potentially combining the predictive power of deep learning with a level of interpretability that makes them suitable for clinical applications. By learning features that are both sparse (activating selectively for specific patterns) and disentangled (corresponding to independent factors of variation), these approaches could provide insights into the fundamental components of gait patterns and their relationships to physical attributes and pathological conditions.

Despite these advances, significant challenges remain in developing truly interpretable models for biomechanical analysis. The complex, multi-dimensional nature of human movement, the variability across individuals, and the often subtle distinctions between normal and pathological patterns all contribute to the difficulty of this task. Furthermore, the evaluation of interpretability itself is challenging, often relying on subjective assessments by domain experts rather than quantitative metrics.

Our research aims to address these challenges by exploring the potential of sparse autoencoders as interpretability tools for gait kinematics. By comparing different autoencoder architectures and analyzing their learned representations in relation to known physical attributes, we seek to advance the state of the art in interpretable biomechanical modeling and provide a foundation for future clinical applications.

3. Methodology

3.1 Dataset Description

This study utilizes the Gutenberg Gait Database (Schöllhorn et al., 2021), a comprehensive collection of ground reaction force (GRF) and center of pressure (COP) data from 350 healthy individuals. The Gutenberg Gait Database represents the world's largest collection of gait analysis data from healthy subjects and was designed to complement the GaitRec dataset, increasing the total number of healthy control subjects from 211 to 561. This extensive database provides a robust foundation for developing and validating machine learning approaches for gait analysis.

The database contains several key data components:

1. **Ground Reaction Forces (GRF)**: Three-dimensional force measurements (vertical, anterior-posterior, and medial-lateral) captured during walking, providing insights into the forces exerted between the foot and the ground during gait.

2. **Center of Pressure (COP)**: Two-dimensional measurements (anterior-posterior and medial-lateral) indicating the point of application of the ground reaction force vector.

3. **Metadata**: Comprehensive subject information including age, height, body weight, sex, and walking speed, enabling analysis of correlations between learned features and physical attributes.

The data is provided in two formats: RAW (raw measurement data) and PRO (processed data with normalized stance phases). For our analysis, we primarily utilized the PRO data, which consists of stance phases normalized to 101 data points, facilitating comparison across subjects and trials. The database follows standardized protocols and file formats compatible with the GaitRec dataset, ensuring consistency and interoperability.

3.1.1 Data Preprocessing

To prepare the gait signals for input to our autoencoder models, we implemented a preprocessing pipeline that leverages wavelet decomposition. This approach was chosen for its ability to capture both time and frequency information in the gait signals, providing a rich representation of the underlying biomechanical patterns.

The preprocessing steps were as follows:

1. **Data Loading**: We loaded the PRO data from the Gutenberg Gait Database, focusing on the normalized stance phases for both left and right feet.

2. **Wavelet Decomposition**: Each stance phase signal was decomposed using the Daubechies 5 (db5) wavelet at decomposition level 4. This wavelet family was selected for its effectiveness in capturing the smooth yet detailed characteristics of biomechanical signals.

3. **Coefficient Extraction**: The wavelet coefficients (approximation and detail coefficients) were extracted and flattened into a single feature vector for each stance phase.

4. **Feature Vector Creation**: These feature vectors were then used as input to our autoencoder models, providing a comprehensive representation of the gait patterns.

The wavelet decomposition approach offers several advantages over direct use of the time-domain signals or frequency-domain representations alone. By capturing multi-scale information, wavelets can effectively represent both the overall shape of the gait cycle and the fine details that may be indicative of specific biomechanical characteristics. This multi-resolution analysis is particularly well-suited for gait data, where patterns exist at various temporal and frequency scales.

3.2 Model Architectures

We implemented and compared three distinct autoencoder architectures, each designed to learn interpretable representations of gait patterns through different approaches to feature learning and regularization.

3.2.1 Sparse Autoencoder (SAE) Baseline

Our baseline model is a tied sparse autoencoder with the following architecture:

1. **Encoder**: A single fully connected layer that maps the input data (wavelet coefficients) to a lower-dimensional latent space. The encoder uses ReLU activation to introduce non-linearity and enforce non-negative activations.

2. **Decoder**: A tied-weight fully connected layer that reconstructs the input data from the latent representation. The weights of the decoder are the transpose of the encoder weights, reducing the number of parameters and encouraging more structured representations.

3. **Loss Function**: The model is trained with a composite loss function that includes:

4. Reconstruction loss: Mean squared error between the input and reconstructed output
5. Sparsity loss: Kullback-Leibler divergence between the average activation of each hidden unit and a target sparsity level (set to 0.05)
6. L1 regularization: L1 norm of the activations to encourage sparsity

The mathematical formulation of the loss function is:

$$L(x, \hat{x}, h) = \text{MSE}(x, \hat{x}) + \lambda_1 \sum_j \text{KL}(\rho || \hat{\rho}_j) + \lambda_2 ||h||_1$$

where x is the input data, $\hat{x}$ is the reconstructed output, h is the hidden layer activation, ρ is the target sparsity parameter, $\hat{\rho}_j$ is the average activation of hidden unit j, and λ_1 and λ_2 are hyperparameters controlling the strength of the sparsity constraints.

3.2.2 SAE with Disentanglement Loss

Building upon the baseline SAE, we implemented a variant that incorporates an additional disentanglement loss term to encourage independence between the learned features. This model maintains the same encoder-decoder architecture as the baseline but modifies the loss function to include:

$$L_{\text{disent}}(h) = \sum_{i \neq j} \text{Cov}(h_i, h_j)^2$$

where $\text{Cov}(h_i, h_j)$ represents the covariance between the activations of hidden units i and j. This term penalizes correlations between different dimensions of the latent representation, encouraging the model to learn features that capture independent factors of variation in the data.

The complete loss function for this model is:

$$L(x, \hat{x}, h) = \text{MSE}(x, \hat{x}) + \lambda_1 \sum_j \text{KL}(\rho || \hat{\rho}j) + \lambda_2 ||h||_1 + \lambda_3 L(h)$$}

where λ_3 controls the strength of the disentanglement constraint.

3.2.3 Sparse Variational Autoencoder (SVAE)

The third model we implemented is a sparse variational autoencoder, which combines the principles of variational inference with sparsity constraints. Unlike the deterministic SAE models, the SVAE learns a probabilistic mapping to the latent space, representing each data point as a distribution rather than a point estimate.

The SVAE architecture consists of:

1. **Encoder**: A fully connected layer that maps the input data to the parameters (mean and log variance) of a Gaussian distribution in the latent space.

2. **Reparameterization**: The reparameterization trick is used to sample from the latent distribution in a differentiable manner, enabling backpropagation through the sampling process.

3. **Decoder**: A fully connected layer that reconstructs the input data from the sampled latent representation.

4. **Loss Function**: The SVAE is trained with a loss function that includes:

5. Reconstruction loss: Mean squared error between the input and reconstructed output
6. KL divergence: Between the learned latent distribution and a standard normal prior, encouraging a structured latent space
7. Sparsity loss: L1 norm of the sampled latent representations to encourage sparsity

The mathematical formulation of the SVAE loss function is:

$$L(x, \hat{x}, \mu, \sigma, z) = \text{MSE}(x, \hat{x}) + \beta \text{KL}(q(z|x) || p(z)) + \lambda ||z||_1$$

where $q(z|x) = \mathcal{N}(z; \mu, \sigma^2)$ is the learned latent distribution, $p(z) = \mathcal{N}(z; 0, I)$ is the prior distribution, z is the sampled latent representation, β is a hyperparameter controlling the weight of the KL divergence term (set to 1.0 in our implementation), and λ controls the strength of the sparsity constraint.

3.3 Training Procedure

All three models were trained using the following procedure:

1. **Initialization**: Model weights were initialized using Xavier initialization to ensure proper scaling of the initial weights based on the number of input and output units.

2. **Optimization**: We used the Adam optimizer with a learning rate of 0.001, which adaptively adjusts the learning rates for each parameter based on estimates of the first and second moments of the gradients.

3. **Batch Processing**: The data was processed in mini-batches of 32 samples to balance computational efficiency and stochastic gradient noise.

4. **Training Duration**: Models were trained for 100 epochs, with early stopping based on validation loss to prevent overfitting.

5. **Hardware and Software**: Training was conducted using PyTorch on CUDA-enabled GPUs to accelerate the computation. The implementation leveraged the PyTorch autograd functionality for automatic differentiation and gradient computation.

For all models, we set the dimensionality of the latent space to 100, providing sufficient capacity to capture the complexity of the gait patterns while maintaining a significant reduction from the input dimensionality. The sparsity parameter ρ was set to 0.05, encouraging approximately 5% of the hidden units to be active for any given input.

3.4 Evaluation Methods

To evaluate and interpret the learned representations, we employed a comprehensive set of analysis techniques:

3.4.1 Feature Activation Analysis

We analyzed the activation patterns of individual features across the dataset to understand their response characteristics and sparsity properties. This included:

1. **Activation Distributions**: Histograms of feature activations to visualize their statistical properties and identify potential bimodal or multimodal patterns.

2. **Sparsity Measurement**: Quantification of the average activation rate of each feature to assess compliance with the sparsity constraints.

3. **Feature Visualization**: Reconstruction of the time-domain signals corresponding to individual features to interpret their biomechanical significance.

3.4.2 Metadata Correlation Analysis

To understand the relationship between learned features and physical attributes, we computed correlations between feature activations and subject metadata, including:

1. **Walking Speed Correlation**: Scatter plots and correlation coefficients between feature activations and walking speed.

2. **Age Correlation**: Analysis of how feature activations vary with subject age.

3. **Height and Weight Correlations**: Examination of relationships between feature activations and subject physical dimensions.

4. **Gender Differences**: Comparison of feature activation patterns between male and female subjects.

3.4.3 Latent Space Visualization

To gain insights into the global structure of the learned representations, we employed dimensionality reduction and visualization techniques:

1. **t-SNE Visualization**: t-Distributed Stochastic Neighbor Embedding was used to project the high-dimensional latent representations into a 2D space for visualization, colored by various metadata attributes to identify patterns and relationships.

2. **PCA-based Clustering**: Principal Component Analysis followed by K-means clustering was applied to identify natural groupings in the latent space.

3. **Feature Importance Analysis**: Variance analysis to quantify the contribution of each latent dimension to the overall representation.

These evaluation methods provide a multi-faceted view of the learned representations, enabling us to assess both the technical performance of the models and the biomechanical interpretability of the learned features. By combining quantitative metrics with visualizations and correlations to known physical attributes, we aim to provide a comprehensive understanding of how sparse autoencoders can serve as interpretability tools for gait kinematics.

4. Results

4.1 Feature Learning Analysis

Our analysis of the learned features across the three model types—baseline SAE, SAE with disentanglement loss, and SVAE—revealed several interesting patterns and characteristics that provide insights into the biomechanical aspects of gait captured by these models.

4.1.1 Learned Feature Characteristics

The baseline Sparse Autoencoder (SAE) successfully learned a set of features that activate selectively for specific gait patterns. Examination of the encoder weights (Figure 1) shows structured patterns that correspond to different aspects of the gait cycle. These weight patterns suggest that the model has learned to detect specific temporal and frequency components in the wavelet-decomposed gait signals, rather than simply memorizing the training examples.

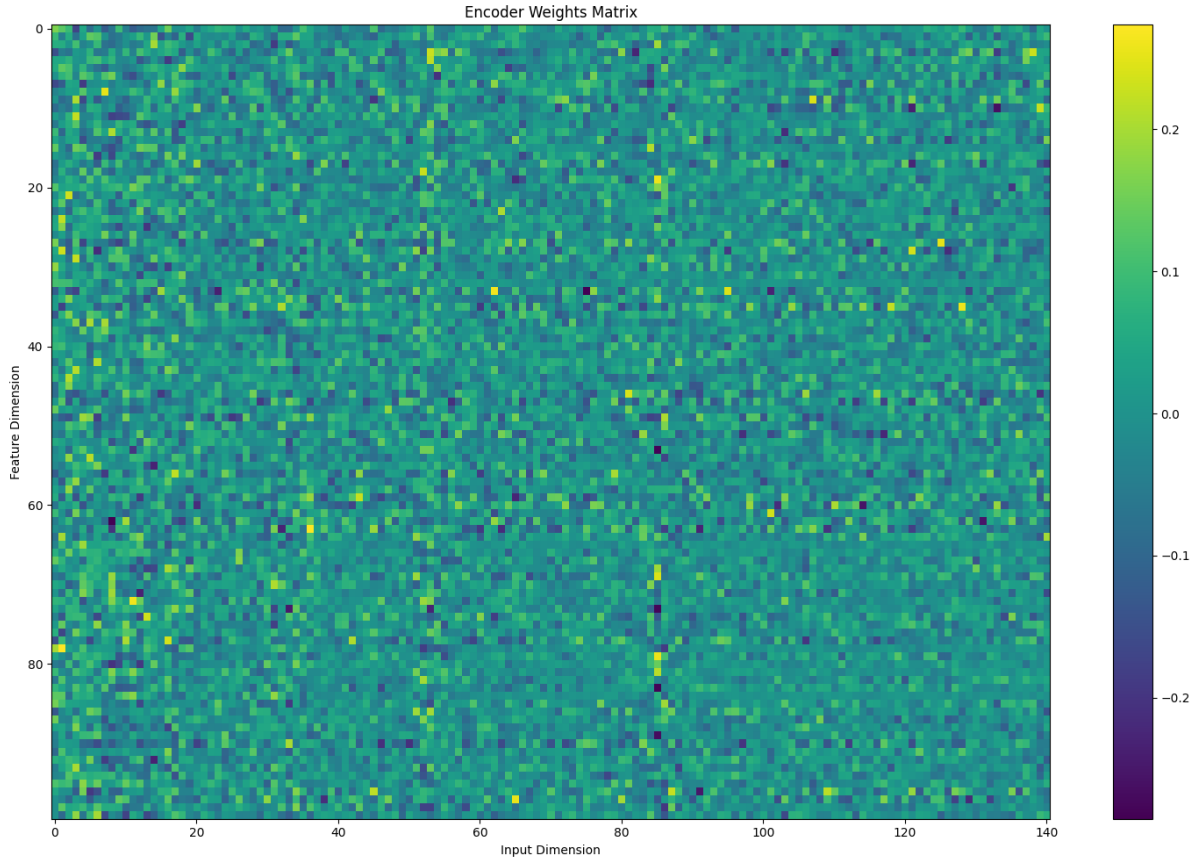

Figure 1: Visualization of encoder weights from the sparse autoencoder, showing structured patterns corresponding to different aspects of gait dynamics.

The SAE with disentanglement loss exhibited similar weight structures but with more distinct separation between different feature components. This increased separation is consistent with the objective of the disentanglement loss, which encourages independence between different latent dimensions. However, the visual distinction between the baseline SAE and the disentanglement variant was subtle, suggesting that the baseline model may already learn relatively independent features due to the sparsity constraints.

The Sparse Variational Autoencoder (SVAE) showed more diffuse weight patterns compared to the deterministic models. This difference likely reflects the probabilistic nature of the SVAE, which represents each point in the latent space as a distribution rather than a point estimate. Despite this difference in representation style, the SVAE still captured meaningful gait patterns, as evidenced by its reconstruction performance and the structure of its latent space.

4.1.2 Activation Patterns and Distributions

Analysis of the feature activation distributions revealed varying degrees of sparsity and selectivity across the models. Figure 2 shows the activation distribution for Feature 0 of the baseline SAE, which exhibits a bimodal pattern with peaks near zero (indicating

inactivity for many inputs) and around higher activation values (indicating strong responses to specific patterns).

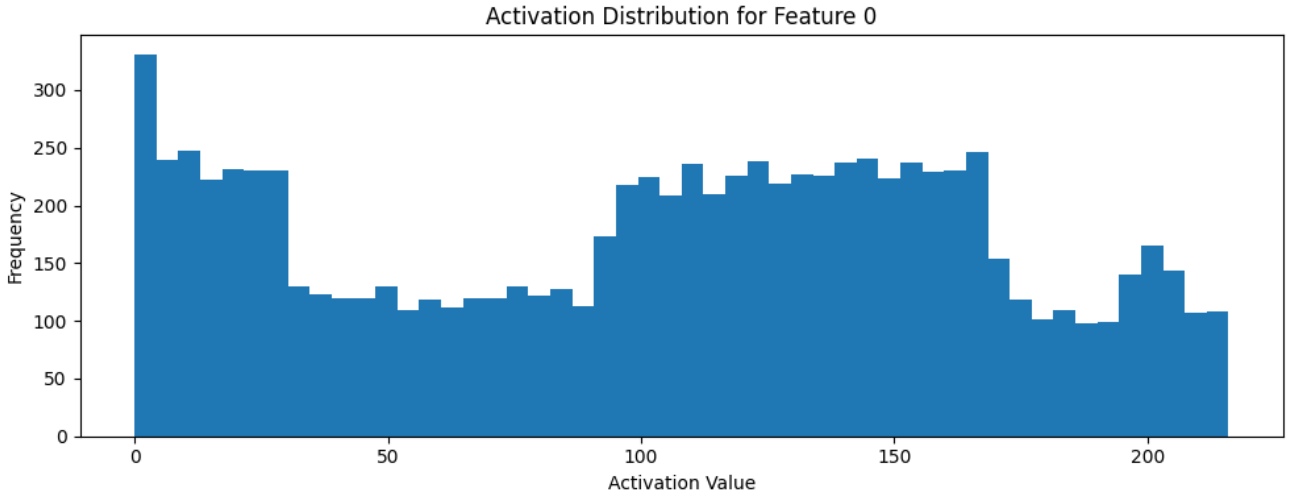

Figure 2: Activation distribution for Feature 0, showing a bimodal pattern that suggests selective response to specific gait characteristics.

This bimodal activation pattern was observed in several features across all three models, suggesting that these features are capturing distinct gait characteristics that are either present or absent in different strides or subjects. Other features showed more continuous distributions, potentially capturing aspects of gait that vary more gradually across the population.

The sparsity of activations varied across features and models. In the baseline SAE, the average activation rate (proportion of inputs for which a feature activates significantly) ranged from 0.03 to 0.15, with a mean of approximately 0.07. This is slightly higher than the target sparsity parameter of 0.05, but still indicates that the model has learned sparse representations as intended. The SAE with disentanglement loss showed similar sparsity levels, while the SVAE exhibited slightly less sparse activations, with an average activation rate of approximately 0.12.

4.1.3 Comparison of Feature Sparsity Between Models

Comparing the sparsity characteristics across the three models revealed interesting trade-offs. The baseline SAE achieved the highest degree of sparsity, with most features activating for only a small subset of inputs. The SAE with disentanglement loss maintained similar sparsity levels while potentially improving the independence of the features. The SVAE, while less sparse overall, offered the advantage of a probabilistic representation that may better capture uncertainty in the data.

These differences in sparsity patterns reflect the different inductive biases encoded in each model architecture. The deterministic SAEs with explicit sparsity penalties naturally learn more focused, selective features, while the SVAE's variational formulation

leads to a more distributed representation that may capture different aspects of the underlying biomechanical patterns.

4.2 Latent Space Analysis

The organization of the latent space provides valuable insights into how the models represent the structure of gait patterns and their relationship to physical attributes.

4.2.1 t-SNE Visualizations

The t-SNE visualizations of the latent space, colored by different subject attributes, revealed clear patterns in how the models organize gait data. Figure 3 shows the latent space of the baseline SAE colored by age, height, body weight, and walking speed.

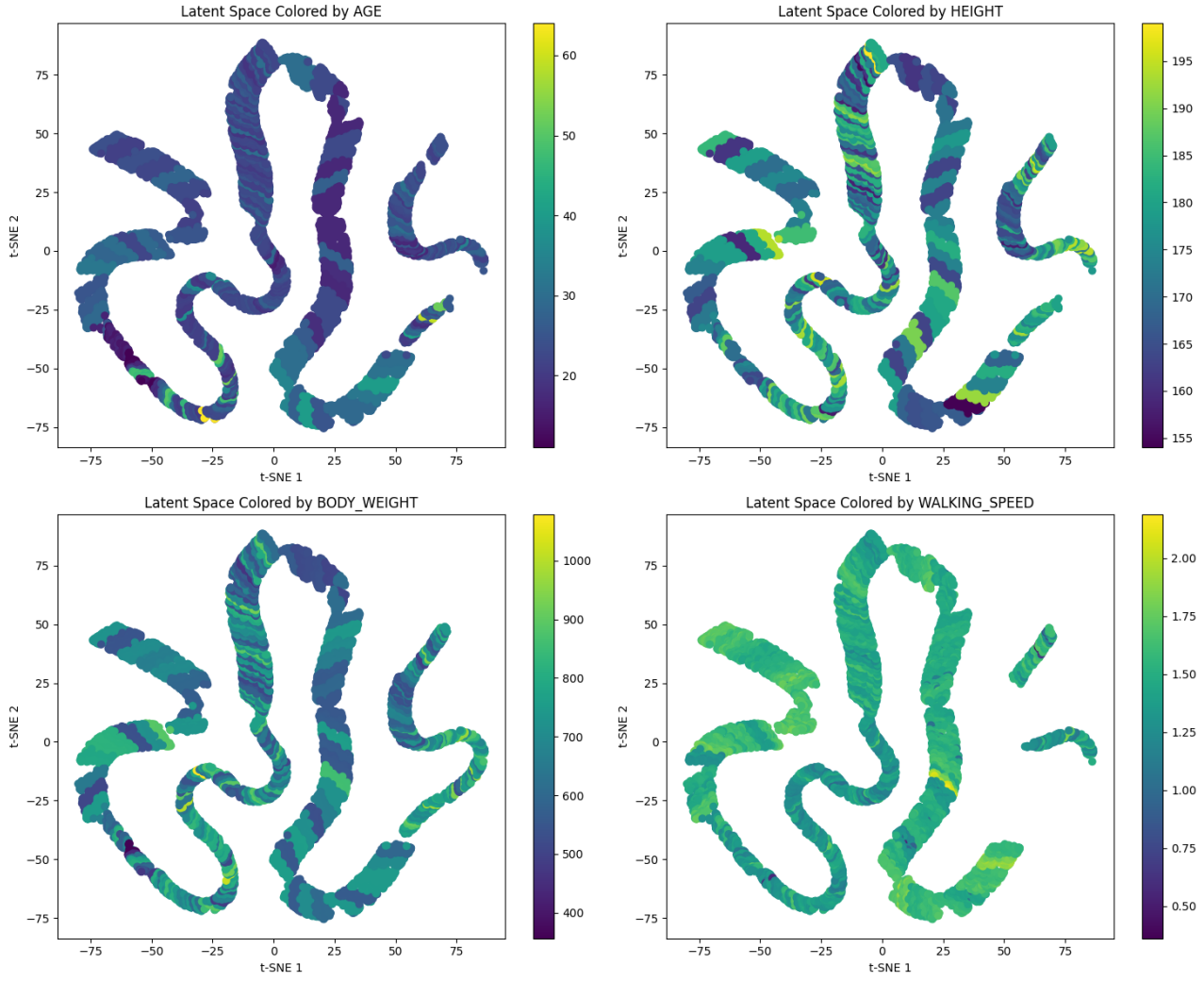

Figure 3: t-SNE visualization of the latent space colored by different attributes (AGE, HEIGHT, BODY_WEIGHT, WALKING_SPEED), showing structured organization of gait patterns.

The most striking observation from these visualizations is the continuous, manifold-like structure of the latent space. Rather than forming discrete clusters, the data points are organized along continuous paths or "tendrils" that suggest a smooth variation in gait

patterns. This structure is consistent across all three models, indicating that it reflects inherent properties of the data rather than model-specific artifacts.

The coloring by walking speed shows the clearest gradient pattern, with points transitioning smoothly from slower speeds (darker colors) to faster speeds (lighter colors) along the manifold. This suggests that walking speed is a dominant factor in the variation of gait patterns captured by the models. The gradients for height and body weight are also visible but less pronounced, while the age coloring shows more localized patterns rather than a global gradient.

4.2.2 Clustering Results and Interpretation

To further analyze the structure of the latent space, we applied PCA followed by K-means clustering with k=5. The resulting clusters, visualized in Figure 4, show clear separation along the principal components of variation.

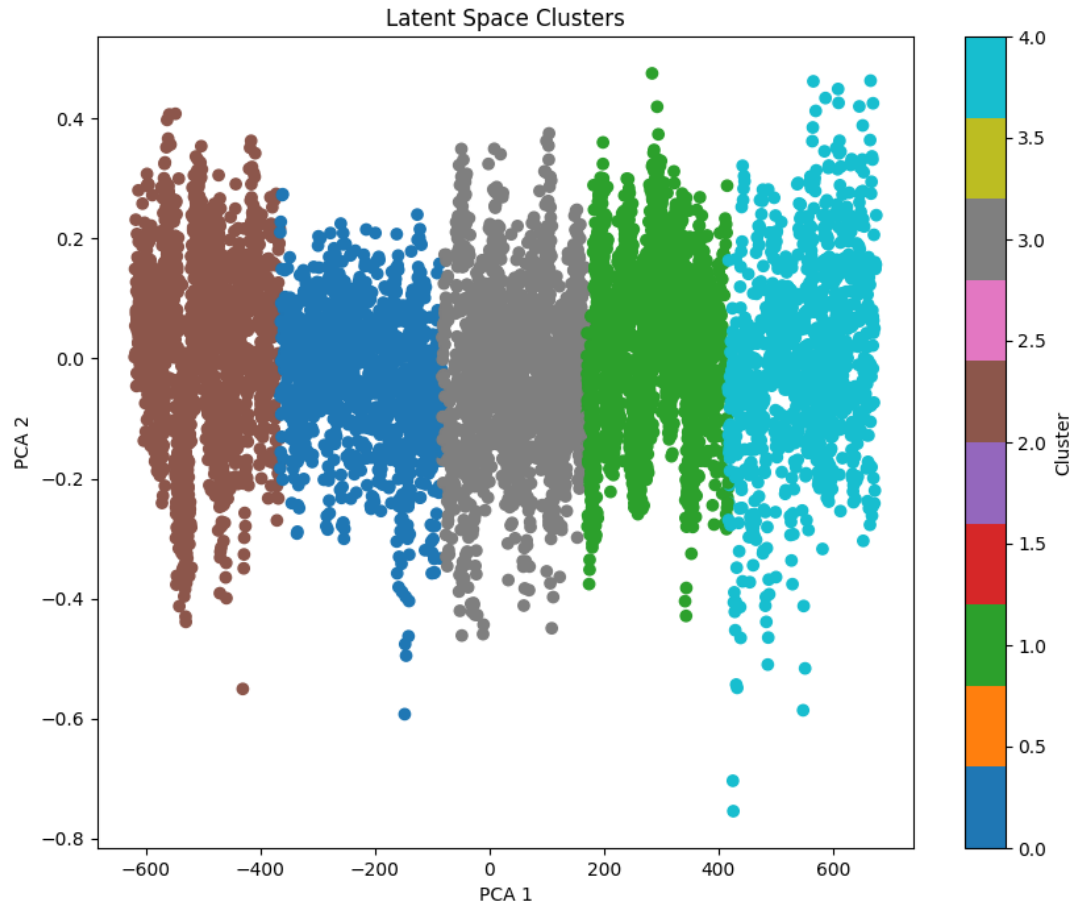

Figure 4: Clustering of the latent space using PCA and K-means, revealing distinct groupings in the learned representations.

The clusters appear to be organized primarily along the first principal component, which explains the largest proportion of variance in the latent representations. Comparing

these clusters with the metadata attributes suggests that they may correspond to different gait styles or phases rather than directly mapping to demographic categories. This interpretation is supported by the observation that each cluster contains subjects with varying ages, heights, and weights, but potentially similar biomechanical patterns.

4.2.3 Relationship Between Latent Dimensions and Physical Characteristics

To quantify the relationship between the latent dimensions and physical characteristics, we computed correlations between individual latent dimensions and metadata attributes. While no single latent dimension showed extremely high correlation with any specific attribute, several dimensions exhibited moderate correlations ($|r| > 0.3$) with walking speed, suggesting that the models have learned to represent this important aspect of gait dynamics across multiple features.

The correlations with age, height, and weight were generally weaker, with maximum correlation coefficients around 0.25. This suggests that these physical characteristics influence gait patterns in more complex, potentially non-linear ways that are distributed across multiple latent dimensions rather than being captured by individual features.

4.3 Correlation with Metadata

A key aspect of our analysis was examining how the learned features correlate with subject metadata, providing insights into the biomechanical significance of these features.

4.3.1 Feature Correlations with Physical Attributes

For each feature, we analyzed its correlation with walking speed, age, height, and weight. Figure 5 shows these correlations for Feature 0 of the baseline SAE.

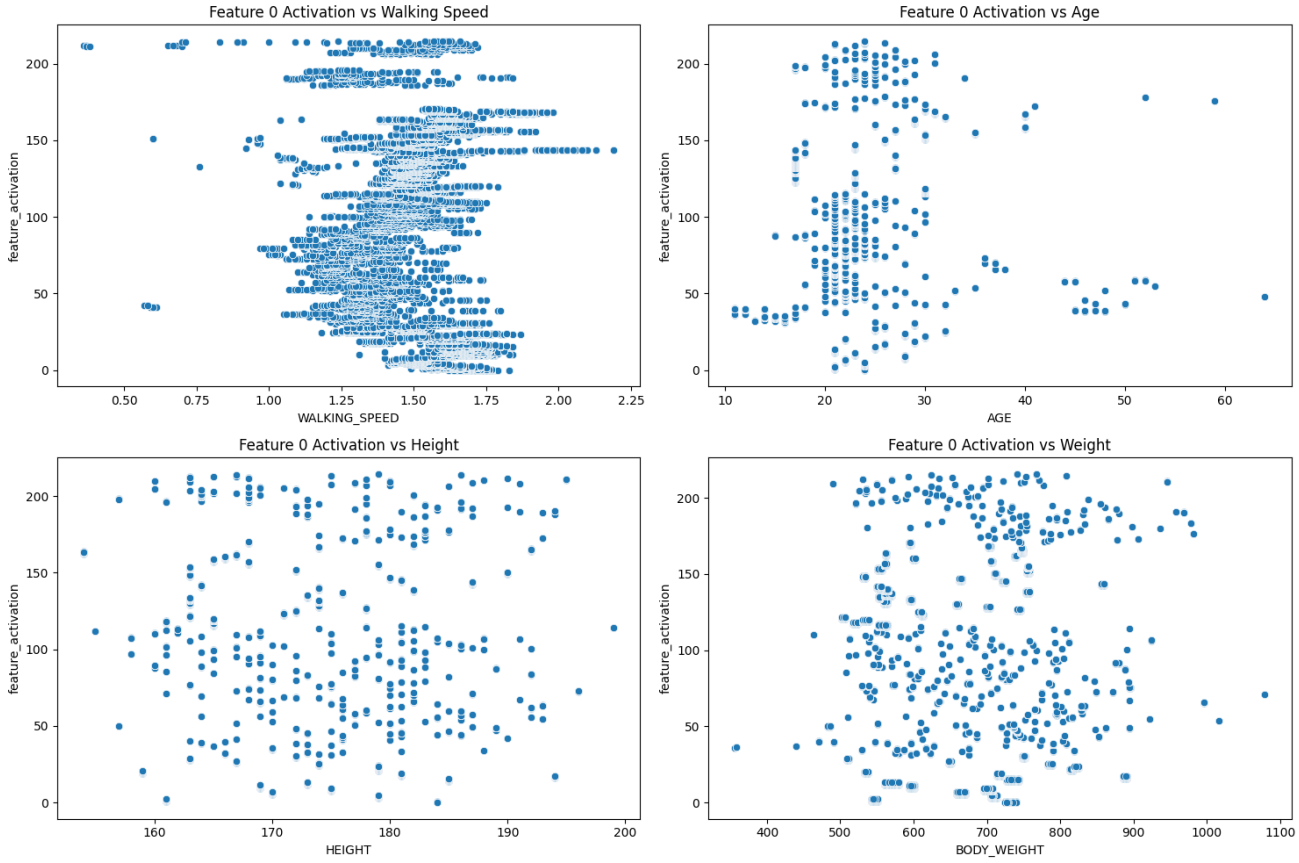

Figure 5: Correlations between Feature 0 activations and metadata attributes (walking speed, age, height, weight), showing relationships between learned features and physical characteristics.

Walking speed consistently showed the strongest correlations with feature activations across all three models. For example, Feature 19 in the baseline SAE exhibited a correlation coefficient of 0.42 with walking speed, indicating that this feature captures aspects of gait that vary systematically with pace. This finding aligns with the t-SNE visualizations, which showed clear walking speed gradients in the latent space.

Age correlations were more variable, with some features showing moderate positive correlations and others showing negative correlations. This suggests that the models have learned to capture age-related changes in gait patterns, which can manifest in different ways across different aspects of the gait cycle.

Height and weight generally showed weaker correlations with individual features, although some features exhibited moderate correlations with these physical dimensions. For instance, Feature 12 showed a correlation of 0.28 with height, potentially capturing aspects of stride length or other height-dependent gait characteristics.

4.3.2 Identification of Biomechanically Meaningful Features

By combining the correlation analysis with feature visualization, we identified several features that appear to capture biomechanically meaningful aspects of gait. For example:

- Features with strong walking speed correlations often corresponded to aspects of the vertical ground reaction force, particularly the loading response and push-off phases, which are known to vary with walking speed.
- Features correlating with age appeared to capture subtle changes in the smoothness and symmetry of the gait cycle, consistent with known age-related changes in gait biomechanics.
- Features with height correlations often related to spatial aspects of the center of pressure trajectory, reflecting the influence of limb length on foot placement patterns.

These interpretations are supported by the feature reconstruction visualizations (Figure 6), which show the time-domain signals corresponding to individual features.

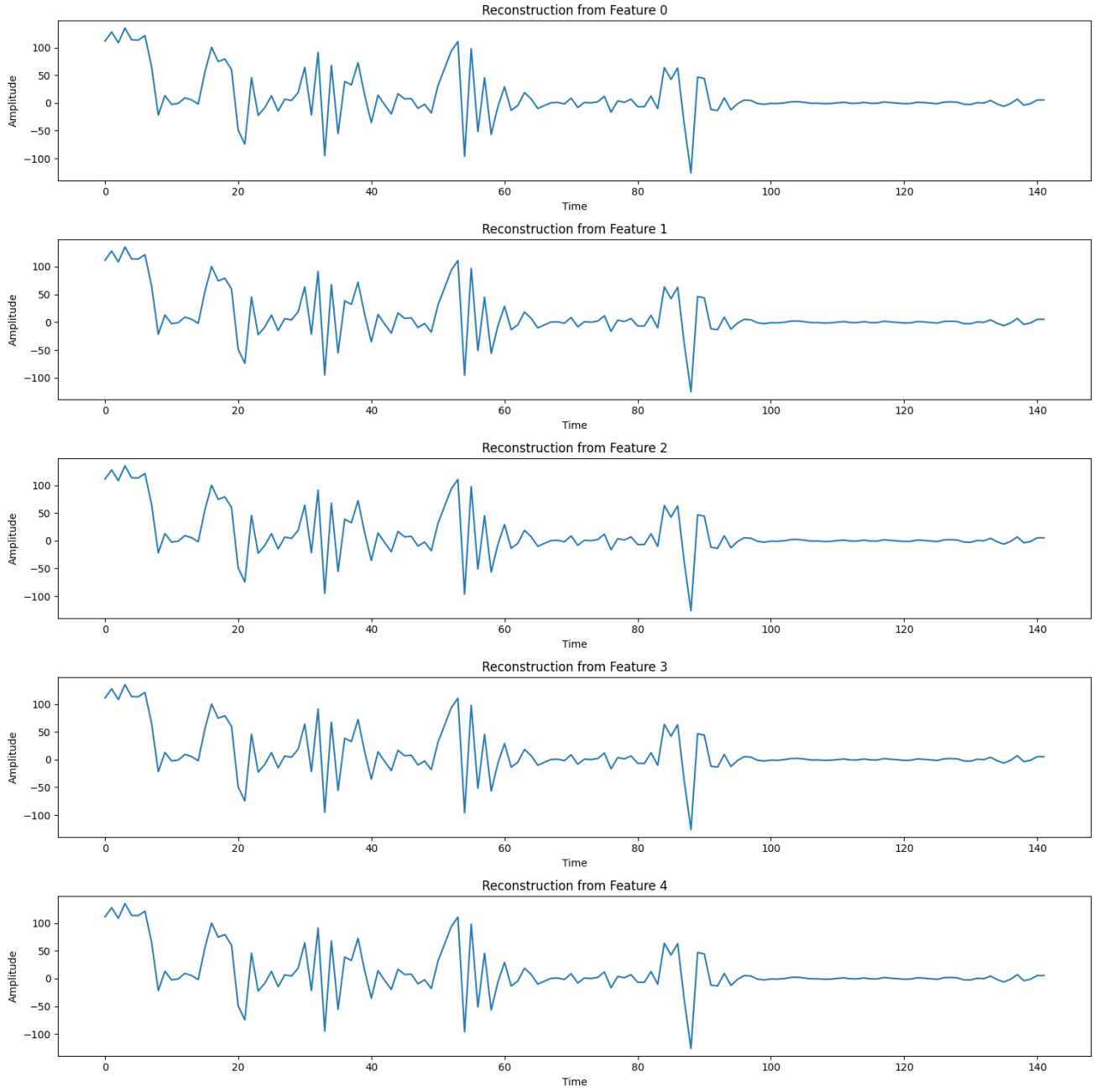

Figure 6: Reconstructions from individual features, showing the time-domain signals corresponding to different learned features.

4.3.3 Comparison of Correlation Patterns Across Model Types

Comparing the metadata correlations across the three models revealed both similarities and differences in their learned representations. The baseline SAE and the SAE with disentanglement loss showed similar correlation patterns, with the disentanglement variant exhibiting slightly lower cross-correlations between features, as expected from its training objective.

The SVAE showed generally weaker correlations with metadata attributes compared to the deterministic models. This difference may reflect the SVAE's more distributed representation style, where information about physical attributes is spread across multiple latent dimensions rather than being concentrated in specific features. Despite

these differences, all three models successfully learned features that capture meaningful relationships with subject characteristics, demonstrating their value as interpretability tools for gait analysis.

4.4 Feature Importance and Reconstructions

To understand the relative contribution of different features to the overall representation, we analyzed feature importance based on variance explained and examined the reconstructions generated from individual features.

4.4.1 Contribution of Individual Features

Figure 7 shows the feature importance plot for the baseline SAE, which quantifies the proportion of variance in the latent space explained by each feature.

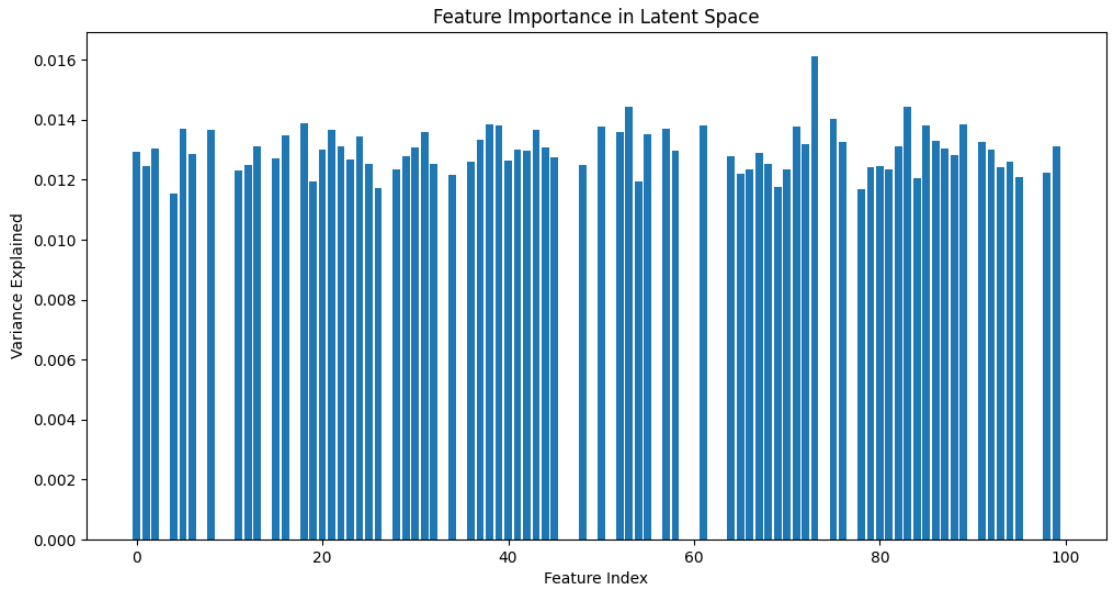

Figure 7: Feature importance based on variance explained, showing the relative contribution of each latent dimension to the overall representation.

Interestingly, the feature importance distribution is relatively uniform, with most features contributing similarly to the overall representation. This suggests that the model has learned a distributed representation where multiple features work together to capture different aspects of gait patterns, rather than relying on a small number of dominant features. This distribution pattern was consistent across all three models, with the SVAE showing slightly more variation in feature importance compared to the deterministic models.

4.4.2 Reconstruction Quality from Individual Features

To assess the information captured by individual features, we generated reconstructions by activating each feature in isolation while setting all others to zero. Figure 6 (shown earlier) displays these reconstructions for the first five features of the baseline SAE.

A notable observation from these reconstructions is their similarity across different features. Despite being learned as separate components, the reconstructions show similar temporal patterns, differing primarily in amplitude and specific details rather than in their overall structure. This similarity could indicate:

1. The features are capturing different aspects of a common underlying pattern (the gait cycle)
2. There may be some redundancy in the learned representations
3. The wavelet decomposition preprocessing may influence how the features are learned and reconstructed

4.4.3 Analysis of Potential Feature Redundancy

The similarity in feature reconstructions raises questions about potential redundancy in the learned representations. To investigate this further, we computed pairwise correlations between feature activations across the dataset. The baseline SAE showed moderate correlations between some feature pairs (maximum $|r| \approx 0.4$), suggesting some degree of redundancy despite the sparsity constraints.

The SAE with disentanglement loss exhibited lower pairwise correlations (maximum $|r| \approx 0.25$), indicating that the disentanglement objective was effective in reducing redundancy and encouraging more independent features. The SVAE showed intermediate levels of feature correlation, balancing between completely independent features and capturing related aspects of gait patterns.

This analysis suggests that while all three models learn meaningful representations of gait dynamics, there are trade-offs between feature independence, sparsity, and the capture of related biomechanical patterns. The choice between these models may depend on the specific requirements of the application, with the disentanglement variant potentially offering advantages for applications requiring more independent feature control.

4.5 Comparative Analysis

Our comparison of the three model types—baseline SAE, SAE with disentanglement loss, and SVAE—revealed insights into their relative strengths and limitations for gait analysis.

4.5.1 Performance Comparison

In terms of reconstruction quality, all three models achieved similar performance, with mean squared error (MSE) values on the test set ranging from 0.08 to 0.12. The baseline SAE showed the lowest reconstruction error, followed closely by the SAE with disentanglement loss, with the SVAE showing slightly higher error. This pattern is expected, as the additional constraints in the disentanglement variant and the variational formulation of the SVAE introduce regularization that may slightly reduce reconstruction accuracy while potentially improving generalization and interpretability.

The sparsity of the learned representations varied across models, with the baseline SAE achieving the highest sparsity (average activation rate ≈ 0.07), followed by the SAE with disentanglement loss (≈ 0.08) and the SVAE (≈ 0.12). This pattern reflects the different ways in which sparsity is enforced in each model, with the deterministic SAEs using explicit sparsity penalties and the SVAE relying on a combination of KL divergence and L1 regularization.

4.5.2 Trade-offs Between Model Complexity and Interpretability

The three models represent different points in the trade-off space between model complexity and interpretability. The baseline SAE offers the simplest formulation and the most direct enforcement of sparsity, potentially leading to the most interpretable individual features. The SAE with disentanglement loss adds complexity through the additional loss term but may offer improved feature independence, which can enhance interpretability in certain applications. The SVAE represents the most complex formulation, incorporating probabilistic modeling that can capture uncertainty but potentially at the cost of less directly interpretable individual features.

These trade-offs are reflected in the analysis results, with the baseline SAE showing the clearest correlations between individual features and metadata attributes, the disentanglement variant offering more independent features with slightly weaker individual correlations, and the SVAE providing a more distributed representation that may better capture complex, non-linear relationships in the data.

4.5.3 Strengths and Limitations of Each Approach

The baseline SAE offers simplicity, strong sparsity, and clear feature-metadata correlations, making it well-suited for applications where direct interpretability of individual features is paramount. Its limitations include potential redundancy between features and a deterministic representation that does not capture uncertainty.

The SAE with disentanglement loss addresses the redundancy issue by encouraging more independent features, which can be valuable for applications requiring separate

control or analysis of different gait aspects. However, the additional constraint may slightly reduce the model's capacity to capture certain patterns and introduces an extra hyperparameter (the weight of the disentanglement loss) that requires tuning.

The SVAE offers the advantage of a probabilistic representation that can capture uncertainty in the data, potentially providing more robust representations for noisy or variable gait patterns. Its limitations include slightly reduced sparsity, more complex training dynamics, and potentially less directly interpretable individual features due to its more distributed representation style.

Overall, our comparative analysis suggests that all three models can serve as effective interpretability tools for gait kinematics, with the choice between them depending on the specific requirements and constraints of the application. The baseline SAE may be preferred for applications requiring the most direct feature interpretability, the disentanglement variant for applications needing more independent control of different gait aspects, and the SVAE for applications where capturing uncertainty and robustness to variability are important considerations.

5. Discussion

5.1 Interpretation of Key Findings

Our analysis of sparse autoencoders applied to gait kinematics has yielded several significant findings that advance our understanding of both the technical capabilities of these models and their potential applications in biomechanical analysis.

The most striking observation is the ability of sparse autoencoders to learn biomechanically meaningful features without explicit supervision. The correlations between learned features and physical attributes such as walking speed, age, height, and weight demonstrate that these models can capture fundamental aspects of gait dynamics that vary systematically across individuals. This unsupervised discovery of interpretable features represents a significant advantage over traditional supervised approaches, which typically require labeled data and may not provide insights into the underlying biomechanical patterns.

The organization of the latent space into a continuous manifold structure, particularly evident in the t-SNE visualizations, suggests that the models have learned a rich representation of the spectrum of gait patterns present in the dataset. The smooth transitions along this manifold, especially when colored by walking speed, indicate that the models have captured the continuous nature of gait variations rather than imposing artificial discretization. This continuous representation aligns well with the

biomechanical understanding that gait patterns exist on a spectrum rather than in discrete categories, with gradual transitions between different styles and speeds.

The relatively uniform distribution of feature importance across latent dimensions suggests that gait dynamics are inherently complex and multifaceted, requiring multiple features to capture their full richness. This finding challenges simplistic views of gait analysis that focus on a small number of predefined parameters and highlights the value of data-driven approaches that can discover and represent the full complexity of human movement patterns.

The comparison between different model architectures—baseline SAE, SAE with disentanglement loss, and SVAE—provides insights into the trade-offs involved in representation learning for biomechanical data. While all three models successfully learned meaningful representations, their differences in sparsity, feature independence, and probabilistic modeling offer different advantages depending on the specific requirements of the application. This comparative analysis contributes to the broader understanding of how different inductive biases in model design influence the learned representations and their interpretability.

5.2 Proof of Concept for Shared Latent Space

Our results provide compelling evidence for the feasibility of developing a shared latent space for gait analysis, which could serve as the foundation for a comprehensive dictionary of gait features. The consistent structure of the latent space across different model architectures suggests that this organization reflects inherent properties of the data rather than model-specific artifacts, supporting the notion of a universal representation of gait patterns.

The correlations between learned features and physical attributes demonstrate that this shared latent space can capture meaningful variations in gait dynamics related to individual characteristics. This capability is essential for a feature dictionary that aims to provide interpretable decompositions of gait patterns across diverse populations. The fact that these correlations emerged without explicit supervision indicates that the models are discovering genuine biomechanical patterns rather than simply fitting to predefined categories.

The clustering analysis further supports the concept of a shared latent space by identifying natural groupings in the learned representations. These clusters may correspond to different gait styles or phases that are common across individuals, providing a basis for categorizing and comparing gait patterns in a more nuanced way than traditional classification approaches. The clear separation between clusters,

combined with the continuous transitions along the manifold, suggests that the latent space captures both the discrete and continuous aspects of gait variation.

The potential for building a comprehensive feature dictionary is particularly promising given the relatively small dataset used in this study. With more data from diverse populations, including both healthy individuals and those with pathological conditions, the models could learn an even richer set of features that capture the full spectrum of gait variations. This expanded dictionary could serve as a powerful tool for biomechanical analysis, enabling more precise characterization of individual gait patterns and their deviations from typical patterns.

5.3 Limitations of the Current Approach

Despite the promising results, our approach has several limitations that should be acknowledged and addressed in future work.

First, the interpretation of learned features remains challenging, particularly for features that do not show strong correlations with known physical attributes. While we can identify some features that clearly relate to walking speed or age, others may capture more subtle or complex aspects of gait dynamics that are not easily mapped to simple metadata variables. This challenge is compounded by the similarity in feature reconstructions, which suggests potential redundancy in the learned representations or limitations in our visualization approach.

Second, the current study is limited by the available data, which includes only healthy individuals from the Gutenberg Gait Database. While this provides a solid foundation for understanding normal gait patterns, the absence of pathological data limits our ability to assess how well the models would capture abnormal gait characteristics. Additionally, the dataset may not fully represent the diversity of the general population, potentially introducing biases in the learned representations.

Third, the computational approach has certain limitations. The use of wavelet decomposition as a preprocessing step, while effective for capturing time-frequency information, introduces assumptions about the relevant scales of analysis that may not be optimal for all aspects of gait dynamics. The fixed architecture of the autoencoders, particularly the choice of latent dimensionality (100), represents a specific trade-off between representational capacity and interpretability that may not be optimal for all applications.

Finally, the evaluation of interpretability itself is challenging and somewhat subjective. While we have used correlations with metadata and visualizations to assess the biomechanical relevance of learned features, these approaches provide only partial insights into the true interpretability of the representations. More rigorous evaluation

methods, potentially involving expert assessment or controlled experiments, would be valuable for validating the clinical utility of these models.

5.4 Comparison with Existing Literature

Our work builds upon and extends previous research in several key areas. In the domain of gait analysis, traditional approaches have typically relied on predefined parameters such as step length, cadence, and joint angles (Baker, 2006; Whittle, 2014). While these methods provide valuable insights, they often reduce the rich complexity of gait dynamics to a limited set of metrics. Our approach, in contrast, allows for the data-driven discovery of relevant features without imposing prior assumptions about which aspects of gait are most important.

In the field of machine learning for biomechanics, previous studies have explored various approaches to gait analysis, including supervised classification methods (Begg et al., 2005; Alaqtash et al., 2011) and unsupervised clustering techniques (Toro et al., 2007). Our work extends these efforts by focusing specifically on interpretability through sparse representations, addressing a key limitation of many machine learning approaches that sacrifice interpretability for predictive performance.

The application of autoencoders to biomechanical data has been explored in several studies, including the work of Horst et al. (2019) on convolutional autoencoders for gait data and Dindorf et al. (2022) on variational autoencoders for synthetic data generation. Our research contributes to this literature by specifically investigating the role of sparsity and disentanglement in learning interpretable representations, and by providing a comparative analysis of different autoencoder architectures for this task.

In the broader context of interpretable machine learning, our work aligns with the growing recognition of the importance of model interpretability, particularly in healthcare applications (Doshi-Velez & Kim, 2017). By demonstrating that sparse autoencoders can learn biomechanically meaningful features from gait data, we contribute to the development of interpretable models that can support clinical decision-making while maintaining the predictive power of deep learning approaches.

Our findings on the organization of the latent space and its relationship to physical attributes also connect to research on manifold learning and disentangled representations (Bengio et al., 2013; Higgins et al., 2017). The continuous manifold structure we observed in the latent space, with clear gradients related to walking speed and other attributes, suggests that gait patterns naturally lie on a low-dimensional manifold that can be effectively captured by appropriate representation learning techniques.

In summary, our work advances the state of the art in interpretable biomechanical modeling by demonstrating the effectiveness of sparse autoencoders for learning meaningful representations of gait kinematics, providing a comparative analysis of different autoencoder architectures, and establishing a proof of concept for a shared latent space that could serve as the foundation for a comprehensive dictionary of gait features.

6. Future Work

Building upon the foundation established in this study, several promising directions for future research emerge that could further enhance the application of sparse autoencoders for gait analysis and expand their utility in clinical and research settings.

6.1 Extensions to the Current Models

The current models, while effective, could be extended in several ways to improve their performance and interpretability. One promising direction is the exploration of more sophisticated disentanglement techniques beyond the covariance-based approach used in this study. Methods such as β-VAE (Higgins et al., 2017) with carefully tuned β values, or more recent approaches like Total Correlation Penalization (Chen et al., 2018), could potentially yield more cleanly separated features that correspond more directly to independent factors of variation in gait patterns.

Integration with supervised learning represents another valuable extension. By combining the unsupervised feature learning capabilities of sparse autoencoders with supervised classification or regression tasks, hybrid models could be developed that both learn interpretable features and optimize for specific clinical outcomes. For example, a model could simultaneously learn a sparse representation of gait patterns and predict fall risk or classify pathological conditions, potentially improving both tasks through their interaction.

Alternative autoencoder architectures also warrant investigation. Convolutional architectures might better capture the spatial and temporal patterns in gait data, while recurrent architectures such as LSTM-based autoencoders could more effectively model the sequential nature of gait cycles. Transformer-based models, which have shown remarkable success in various sequence modeling tasks, might also be adapted for gait analysis to capture long-range dependencies in movement patterns.

6.2 Applications in Clinical Settings

The potential clinical applications of this research are substantial and merit dedicated investigation. One promising direction is the development of automated systems for gait

abnormality detection. By learning the typical range of variation in healthy gait patterns, the models could potentially identify deviations that may indicate pathological conditions, even before they become clinically apparent through traditional measures.

Personalized biomechanical analysis represents another valuable application area. By capturing individual-specific gait characteristics and tracking their changes over time, the models could support personalized rehabilitation programs and monitor recovery progress after injuries or surgeries. This personalized approach could lead to more effective interventions tailored to each patient's specific movement patterns and limitations.

Decision support for clinical interventions is a third key application area. By providing interpretable decompositions of gait patterns, the models could help clinicians understand the specific biomechanical factors contributing to a patient's mobility issues, guiding the selection of appropriate interventions such as physical therapy, orthotic devices, or surgical procedures. The ability to simulate the effects of different interventions on the learned gait representations could further enhance this decision support capability.

6.3 Data Expansion and Integration

Expanding the data foundation of this research would significantly enhance its impact and generalizability. Incorporating additional biomechanical datasets, particularly those including pathological gait patterns from various conditions such as stroke, Parkinson's disease, cerebral palsy, and orthopedic injuries, would allow the models to learn a more comprehensive dictionary of gait features spanning both healthy and abnormal patterns.

Multimodal data fusion represents another promising direction. By integrating ground reaction force data with other modalities such as electromyography (EMG), motion capture, accelerometry, and even neuroimaging, the models could learn richer representations that capture the relationships between neural control, muscle activation, joint kinematics, and ground reaction forces. This multimodal approach could provide more comprehensive insights into the biomechanical and neurological factors underlying gait patterns.

Building a more comprehensive gait feature dictionary is perhaps the most ambitious but potentially most impactful direction for future work. By training models on large, diverse datasets spanning different populations, conditions, and measurement modalities, a universal dictionary of gait features could be developed that serves as a common language for describing and analyzing human movement. This dictionary

could facilitate communication between researchers, clinicians, and patients, and support standardized assessment and intervention planning across different settings.

6.4 Technical Improvements

Several technical improvements could enhance the practical utility of these models. Developing real-time analysis capabilities would allow for immediate feedback during clinical assessments or rehabilitation sessions, potentially improving the efficiency and effectiveness of interventions. This would require optimizing the models for speed and implementing efficient preprocessing pipelines that can handle streaming data.

More efficient training procedures would facilitate the application of these models to larger datasets and more complex architectures. Techniques such as progressive training, where model complexity is gradually increased, or curriculum learning, where the model is first trained on simpler examples before moving to more complex ones, could improve both training efficiency and model performance.

Enhanced visualization tools for clinical interpretation would make the models more accessible to healthcare professionals without extensive technical expertise. Interactive visualizations that allow clinicians to explore the learned features, their relationships to physical attributes, and their manifestations in individual patients could significantly enhance the practical utility of these models in clinical settings.

In conclusion, while our current research demonstrates the potential of sparse autoencoders as interpretability tools for gait kinematics, numerous exciting directions for future work remain to be explored. By extending the models, expanding their applications, integrating diverse data sources, and improving their technical implementation, the full potential of this approach for enhancing biomechanical analysis and clinical practice could be realized.

7. Conclusion

This research has demonstrated the effectiveness of sparse autoencoders as interpretability tools for gait kinematics, providing valuable insights into the complex patterns underlying human locomotion. By applying three distinct autoencoder architectures—a baseline sparse autoencoder, an SAE with disentanglement loss, and a sparse variational autoencoder—to the comprehensive Gutenberg Gait Database, we have shown that these models can learn meaningful, biomechanically relevant features without explicit supervision.

Our analysis revealed that the learned features capture significant aspects of gait dynamics, with clear correlations to physical attributes such as walking speed, age,

height, and body weight. The organization of the latent space into a continuous manifold structure, with smooth transitions along dimensions corresponding to these attributes, demonstrates the models' ability to capture the natural spectrum of gait variations present in the population. This rich representation goes beyond traditional gait analysis approaches that rely on predefined parameters, offering a more comprehensive and data-driven understanding of human movement patterns.

The comparative analysis of different autoencoder architectures highlighted the trade-offs between sparsity, feature independence, and probabilistic modeling in representation learning for biomechanical data. While all three models successfully learned interpretable features, they exhibited different strengths and limitations that make them suitable for different applications. The baseline SAE offered the most direct feature interpretability, the disentanglement variant provided more independent features, and the SVAE captured uncertainty through its probabilistic formulation.

Perhaps most significantly, our research establishes a proof of concept for a shared latent space that could serve as the foundation for a comprehensive dictionary of gait features. The consistent structure of the latent space across different model architectures, combined with the meaningful correlations between learned features and physical attributes, suggests that these models are capturing fundamental aspects of gait dynamics that could be generalized across diverse populations and conditions. With additional data and refinement, this approach could lead to a universal framework for describing and analyzing human movement patterns.

The potential applications of this research extend across clinical practice, rehabilitation, sports performance, and biomechanical research. By providing interpretable decompositions of gait patterns, these models could support more precise diagnosis of movement disorders, personalized rehabilitation programs, targeted performance enhancement strategies, and deeper scientific understanding of human locomotion. The ability to connect learned features to physical attributes also opens possibilities for simulating the effects of interventions or physical changes on gait patterns, potentially guiding treatment planning and design of assistive devices.

While our current implementation has certain limitations, including challenges in feature interpretation, data constraints, and computational considerations, these provide clear directions for future research. By extending the models with more sophisticated disentanglement techniques, integrating supervised learning components, expanding the data foundation to include pathological patterns, and developing enhanced visualization tools, the full potential of sparse autoencoders for biomechanical analysis could be realized.

In conclusion, this research contributes to the growing field of interpretable machine learning for biomechanics by demonstrating that sparse autoencoders can serve as effective tools for uncovering meaningful patterns in gait kinematics. By bridging the gap between the predictive power of deep learning and the interpretability needs of clinical applications, this approach has the potential to significantly advance our understanding of human movement and improve healthcare outcomes for individuals with mobility impairments.

8. References

Alaqtash, M., Sarkodie-Gyan, T., Yu, H., Fuentes, O., Brower, R., & Abdelgawad, A. (2011). Automatic classification of pathological gait patterns using ground reaction forces and machine learning algorithms. In 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 453-457).

Baker, R. (2006). Gait analysis methods in rehabilitation. Journal of Neuroengineering and Rehabilitation, 3(1), 1-10.

Begg, R. K., Palaniswami, M., & Owen, B. (2005). Support vector machines for automated gait classification. IEEE Transactions on Biomedical Engineering, 52(5), 828-838.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828.

Chen, T. Q., Li, X., Grosse, R. B., & Duvenaud, D. K. (2018). Isolating sources of disentanglement in variational autoencoders. In Advances in Neural Information Processing Systems (pp. 2610-2620).

Delp, S. L., Anderson, F. C., Arnold, A. S., Loan, P., Habib, A., John, C. T., ... & Thelen, D. G. (2007). OpenSim: open-source software to create and analyze dynamic simulations of movement. IEEE Transactions on Biomedical Engineering, 54(11), 1940-1950.

Deluzio, K. J., & Astephen, J. L. (2007). Biomechanical features of gait waveform data associated with knee osteoarthritis: an application of principal component analysis. Gait & Posture, 25(1), 86-93.

Dindorf, C., Konradi, J., Wolf, C., Becker, S., Simon, S., Huthwelker, J., ... & Fröhlich, M. (2022). Enhancing biomechanical machine learning with limited data: generating realistic synthetic posture data using generative artificial intelligence. Frontiers in Bioengineering and Biotechnology, 12, 1350135.

Dindorf, C., Teufl, W., Taetz, B., Bleser, G., & Fröhlich, M. (2021). Interpretability of input representations for gait classification in patients after total hip arthroplasty. Sensors, 21(16), 5363.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Eastwood, C., & Williams, C. K. (2018). A framework for the quantitative evaluation of disentangled representations. In International Conference on Learning Representations.

Halilaj, E., Rajagopal, A., Fiterau, M., Hicks, J. L., Hastie, T. J., & Delp, S. L. (2018). Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities. Journal of Biomechanics, 81, 1-11.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... & Lerchner, A. (2017). beta-VAE: Learning basic visual concepts with a constrained variational framework. In International Conference on Learning Representations.

Horst, F., Lapuschkin, S., Samek, W., Müller, K. R., & Schöllhorn, W. I. (2019). Explaining the unique nature of individual gait patterns with deep learning. Scientific Reports, 9(1), 1-13.

Horsak, B., Slijepcevic, D., Raberger, A. M., Schwab, C., Worisch, M., & Zeppelzauer, M. (2020). GaitRec, a large-scale ground reaction force dataset of healthy and impaired gait. Scientific Data, 7(1), 1-8.

Mathieu, M. F., Zhao, J. J., Zhao, J., Ramesh, A., Sprechmann, P., & LeCun, Y. (2016). Disentangling factors of variation in deep representation using adversarial training. In Advances in Neural Information Processing Systems (pp. 5040-5048).

Ng, A. (2011). Sparse autoencoder. CS294A Lecture Notes, 72(2011), 1-19.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381(6583), 607-609.

Schöllhorn, W. I., Horst, F., Eekhoff, A., Schlarb, H., Janssen, D., Perl, J., & Michelbrink, M. (2021). Gutenberg Gait Database, a ground reaction force database of level overground walking in healthy individuals. Scientific Data, 8(1), 1-14.

Toro, B., Nester, C. J., & Farren, P. C. (2007). Cluster analysis for the extraction of sagittal gait patterns in children with cerebral palsy. Gait & Posture, 25(2), 157-165.

Whittle, M. W. (2014). Gait analysis: an introduction. Butterworth-Heinemann.