






Optimizing Gait-Based Biometric Authentication in the Metaverse Using Random Forest and Support Vector Machine Algorithms

Tonni Limbong^{1,*}, Gonti Simanullang²,
Parasian D.P. Silitonga³

^{1,2,3}Department Information System, Faculty of Computer Sciences, Universitas Katolik Santo Thomas, Medan 20135, Indonesia

ABSTRACT

This paper investigates the potential of gait-based authentication for securing virtual environments, specifically within the Metaverse. With the growing need for reliable and secure identity verification in virtual spaces, traditional authentication methods, such as passwords or PINs, have proven insufficient. In contrast, biometric authentication systems, including gait analysis, provide a more secure and user-friendly alternative by leveraging unique physiological and behavioral traits for identity verification. This research applies machine learning algorithms—Random Forest and Support Vector Machine (SVM)—to gait data for distinguishing between authentic users and imposters. The dataset consists of 1,000 simulated gait samples with 16 features, such as stride length, step frequency, joint angles, and ground reaction forces (GRF). After performing exploratory data analysis (EDA), including feature distribution visualization and correlation analysis, two models were trained on the data. The Random Forest model outperformed the SVM model, achieving an accuracy of 56% and a recall of 76%, indicating its effectiveness in identifying authentic users. Despite the promising results, both models showed only marginal improvement over random guessing, highlighting the need for further optimization. This study contributes to the growing body of research on gait-based biometric systems by demonstrating their potential as a viable method for identity verification in virtual environments. It also identifies the most important gait features, such as step frequency, cadence variability, and knee joint angle, that significantly contribute to the classification process. Future research should explore advanced deep learning techniques and the integration of multimodal biometric systems to enhance the performance and reliability of gait-based authentication.

Keywords Gait-Based Authentication, Metaverse Security, Machine Learning, Random Forest, Biometric Systems

INTRODUCTION

The Metaverse is rapidly evolving into a digital frontier that fosters virtual interactions across various sectors, offering immersive user experiences while simultaneously raising significant concerns about security and identity verification. As the virtual world grows, ensuring that users can interact securely becomes a critical challenge. Biometric authentication systems, particularly those leveraging gait analysis, are increasingly seen as essential tools to address these concerns. By providing a means of verifying identity through unique behavioral characteristics, gait analysis offers a promising solution to safeguard interactions within virtual environments.

How to cite this article: T. Limbong, G. Simanullang, P. D.P. Silitonga, "Optimizing Gait-Based Biometric Authentication in the Metaverse Using Random Forest and Support Vector Machine Algorithms," *Int. J. Res. Metav.*, vol. 2, no. 4, pp. 248-268, 2025.

Submitted: 10 April 2025
Accepted: 25 June 2025
Published: 20 November 2025

Corresponding author
Tonni Limbong,
tonni.budidarma@gmail.com

Additional Information and
Declarations can be found on
[page 265](#)

DOI: [10.47738/ijrm.v2i4.37](https://doi.org/10.47738/ijrm.v2i4.37)

 Copyright
2025 Limbong, et al.,

Distributed under
Creative Commons CC-BY 4.0

Biometric systems utilize individual physiological and behavioral traits, such as fingerprints, facial recognition, and gait patterns, to authenticate identity. These systems offer a more secure alternative to traditional authentication methods like passwords, which are becoming increasingly inadequate in protecting digital interactions [1]. Gait analysis, which evaluates a person's walking pattern, stands out as a particularly reliable biometric marker in the Metaverse. Its non-intrusive nature and the difficulty in replicating a person's walking style make it an attractive option for ensuring genuine interactions in virtual spaces [2].

To further enhance security, advancements in biometric technologies, such as cancelable biometrics, provide additional privacy protection by masking biometric traits. This ensures that even in the event of data breaches, individuals' identities remain secure [3], [4]. Combining multiple biometric traits in multi-modal systems also helps reduce error rates in identity verification, making it more difficult for attackers to spoof or bypass the system [5]. As the Metaverse expands, the integration of these sophisticated biometric systems into existing security frameworks will be vital to maintaining both privacy and trust in virtual environments.

The primary objective of this research is to explore the use of gait-based authentication systems powered by machine learning algorithms, aiming to improve user security within the Metaverse. As the virtual world becomes more immersive and integral to daily life, ensuring secure and seamless user interactions is paramount. By leveraging gait analysis, which examines unique walking patterns, this study seeks to validate its potential as a reliable biometric tool for identity verification in virtual environments.

To achieve this, the research goal is to optimize the authentication accuracy of gait-based systems by applying Random Forest and Support Vector Machine (SVM) algorithms. These two machine learning algorithms are well-suited for classification tasks, with Random Forest providing a robust ensemble approach and SVM offering precision in identifying patterns within data. The combination of these models aims to enhance the overall effectiveness of gait recognition for user authentication in the Metaverse, reducing the risk of identity spoofing or unauthorized access.

The significance of this research lies in demonstrating how gait analysis can serve as a dependable biometric identifier within the Metaverse. As virtual spaces expand and require secure user management, gait-based authentication presents a non-intrusive, difficult-to-replicate method for verifying identities. By showcasing its potential through machine learning techniques, this study highlights the value of gait as a reliable, scalable solution for biometric security in the growing digital frontier.

Literature Review

Overview Of Biometric Authentication

Biometric authentication has become an essential security mechanism in virtual environments due to its ability to utilize unique physiological and behavioral traits to verify an individual's identity. Traditional authentication methods, such as passwords or PINs, often fall short in providing the level of security required for sensitive interactions within virtual spaces. This is because these methods

are prone to being easily guessed, stolen, or shared. In contrast, biometric systems provide more secure alternatives by using non-transferable, difficult-to-forge identifiers, such as facial features, fingerprints, or iris patterns. The distinctiveness of these traits ensures that only the person to whom they belong can access the system, providing a higher level of security. Furthermore, biometrics offer a user-friendly experience since they do not require the user to remember passwords or carry additional tokens. Instead, these systems authenticate identity based on traits that are inherently present in the individual, making them more convenient and efficient [6].

Among the various biometric methods, face recognition is one of the most widely used and highly effective techniques in virtual environments. It leverages the unique structure and features of an individual's face, such as the distance between eyes, nose shape, and jawline, to accurately identify or authenticate a person. This method is particularly advantageous in virtual settings due to its contactless nature, allowing for quick and seamless authentication. For instance, face recognition systems are increasingly integrated into smart devices, security access systems, and online platforms, enabling users to authenticate themselves swiftly and securely. Recent advancements have significantly enhanced face recognition capabilities, allowing these systems to maintain high levels of accuracy even when users wear masks or other facial coverings—common in many public settings. Deep learning techniques, particularly those that utilize models like VGG16 combined with random Fourier features, have improved the adaptability and robustness of face recognition systems, ensuring they can operate effectively under various conditions [7], [8].

Another key biometric method is fingerprint scanning, which has been widely adopted due to its high reliability and ease of use. Fingerprint recognition is particularly common in security-sensitive applications such as Automated Teller Machines (ATMs), mobile devices, and access control systems. This method's primary advantage lies in the distinctiveness of each individual's fingerprint, which remains unchanged throughout life. The integration of advanced image processing techniques, such as Prewitt filtering for segmentation, has further refined fingerprint recognition accuracy, enabling these systems to provide highly reliable identity verification. Furthermore, the use of multimodal biometric systems is gaining traction, as it combines several biometric methods, such as face recognition, fingerprint scanning, and iris scanning, to improve the accuracy and resilience of the authentication process. This approach helps mitigate the weaknesses of single-modal systems by enhancing their overall performance and reliability. In educational and security settings, the fusion of these diverse modalities has been shown to significantly enhance the authentication process, providing a more robust system that can adapt to varying conditions [9], [10].

The continual advancements in sensor technology, image processing, and deep learning have driven the development of these biometric systems, allowing them to operate effectively in dynamic and contactless virtual environments. By integrating deep learning algorithms, biometric systems can improve their ability to process complex data from multiple modalities and continuously learn from new inputs, increasing their accuracy over time. The evolution of multimodal biometric systems has been particularly impactful, enabling more sophisticated methods of identity verification that cater to the growing security needs of virtual

spaces. These systems offer greater resilience against attacks, such as spoofing, and enhance the user experience by providing a seamless and reliable method of authentication. In summary, the integration of biometric authentication methods like face recognition and fingerprint scanning into virtual environments significantly enhances security while addressing the limitations of traditional methods, offering both increased protection and improved user convenience [11].

Gait-Based Authentication

Gait-based authentication has emerged as a significant method for biometric identification, leveraging an individual's unique walking patterns to verify their identity. Its primary appeal lies in its non-intrusive nature, which allows for identification from a distance without the need for direct interaction with the subject. This characteristic makes it particularly valuable in both surveillance and access control systems, offering a secure, seamless alternative to traditional authentication methods like passwords or physical identification cards. Gait-based systems are also less susceptible to spoofing or forgery, as replicating an individual's gait is inherently difficult. This makes gait analysis an attractive solution in modern security frameworks, where ensuring the privacy and safety of users is crucial [12].

The application of gait for identification has expanded significantly due to advancements in sensor technologies, machine learning techniques, and a deeper understanding of human biomechanics. The use of inertial measurement units (IMUs) for analyzing gait patterns in real-world settings is one such advancement. These devices capture key temporal parameters that are essential for accurate gait recognition, allowing systems to recognize individuals based on their walking patterns even in dynamic, real-world environments. Recent studies have shown that gait analysis can be applied in various sectors beyond traditional security, such as healthcare. For instance, gait abnormalities have been linked to psychological conditions, such as depression, showcasing the broader applications of gait analysis in both health monitoring and identity verification [13].

Furthermore, gait analysis has proven valuable in forensic contexts, where it can be used to corroborate identities by comparing gait characteristics with known patterns. This is particularly significant in criminal investigations, where gait features can provide critical evidence to support or challenge the identity of suspects [14], [15]. While gait analysis has traditionally been associated with medical rehabilitation and patient monitoring, it is increasingly recognized as a viable biometric authentication tool [16]. The ability to identify individuals by their gait, especially in situations where facial features or fingerprints are obscured, makes it a promising alternative for secure authentication systems. Ongoing research and technological developments in this field will be crucial for refining gait recognition methods, ensuring their accuracy and reliability in both security and healthcare applications [17], [18].

Machine Learning in Biometric Systems

Machine learning algorithms play a vital role in improving the performance of biometric authentication systems, helping to process complex biometric data and ensuring secure identification. Among the various machine learning algorithms, Random Forest and Support Vector Machine (SVM) are particularly

well-suited for this purpose due to their robust classification capabilities and ability to handle high-dimensional data. These algorithms enable the accurate analysis of biometric features such as facial recognition, fingerprints, and gait patterns, making them indispensable tools in modern biometric systems [19], [20].

The Random Forest algorithm is an ensemble learning technique that builds multiple decision trees based on random subsets of the training data. It then aggregates the predictions from these individual trees, typically using majority voting, to produce a final classification result. This method is particularly effective in handling large datasets with many features, which is often the case in biometric applications where multiple physiological traits are analyzed simultaneously. A significant advantage of Random Forest is its ability to mitigate overfitting, a common challenge in traditional decision tree models. By averaging the results of many trees, it enhances generalization and improves performance on unseen data [19]. In multimodal biometric systems, where multiple biometric features are combined, Random Forest has been successfully applied to analyze interactions among diverse biometric inputs, such as facial features, fingerprints, and gait patterns [2]. Its robustness and adaptability make it a popular choice for high-accuracy biometric systems.

The Support Vector Machine (SVM) is another powerful algorithm widely used for classification tasks, particularly in scenarios where the data is high-dimensional. SVM's primary objective is to find the optimal hyperplane that separates data points belonging to different classes while maximizing the margin between the closest points, known as support vectors. This ability to identify an optimal separation boundary makes SVM highly effective for biometric applications where distinguishing features need to be extracted from complex data. In particular, SVM is highly effective in tasks like face recognition and fingerprint classification, where the data is not always linearly separable. The use of kernel functions enables SVM to handle non-linear relationships, further enhancing its applicability in high-dimensional biometric data spaces [21]. SVM is also resistant to overfitting, making it a reliable choice for extracting relevant features from biometric data like fingerprints, iris patterns, and facial images. Additionally, SVM can be combined with other machine learning techniques to improve feature extraction and classification accuracy, thereby enhancing the performance of biometric systems [22].

Both Random Forest and SVM have demonstrated their importance in advancing biometric authentication systems, offering effective solutions for securely verifying identities while maintaining high accuracy. These algorithms help ensure that biometric systems are resilient to adversarial attacks and capable of processing complex biometric data efficiently, making them essential components of modern security frameworks [23]. As biometric technologies continue to evolve, the integration of these machine learning techniques will play a key role in enhancing the security, accuracy, and robustness of authentication systems.

Method

The methodology employed in this study is systematically illustrated in [figure 1](#). The process begins with loading the gait dataset containing features (X) and labels (y), followed by validating the dataset and handling any missing values.

Once the dataset is confirmed to be valid, the features are standardized and the data is split into training and testing sets. Next, a model is selected—either a Random Forest or a Support Vector Machine (SVM) with an RBF kernel—and trained accordingly. The trained model is then evaluated using performance metrics such as accuracy, precision, recall, and F1-score. If the performance is unsatisfactory, the model parameters are adjusted and retraining is conducted iteratively until acceptable results are achieved. Finally, the trained models are saved, and feature importance is computed to identify the contribution of each feature to the model's performance.

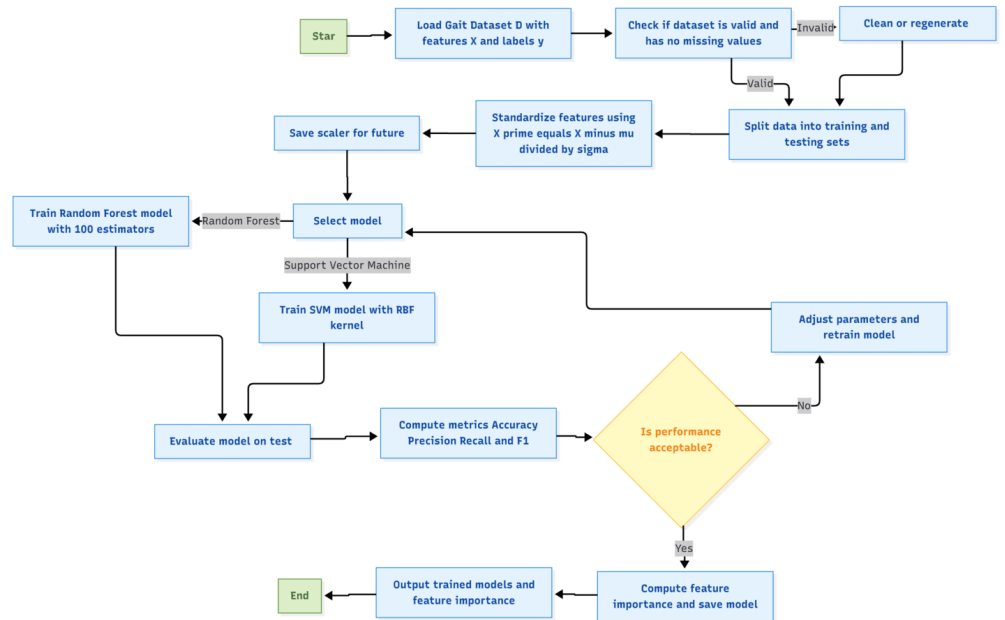


Figure 1 Research Flow

Data Loading and Preprocessing

The research begins with the loading of gait data that serves as the foundation for model development. A simulated dataset consisting of 1,000 samples and 16 gait-related features—such as stride length, step frequency, joint angles, and ground reaction forces—is generated. Each sample is assigned a binary label: 1 for authentic users and 0 for imposters.

The dataset is split into training and testing subsets using stratified sampling to maintain class balance. Data scaling is then applied to ensure uniform feature magnitude across all attributes using standardization. This prevents any single feature from disproportionately influencing model performance. The mathematical formulations used in this phase are as follows:

$$X_{\text{train}}, X_{\text{test}}, y_{\text{train}}, y_{\text{test}} = \text{train_test_split}(X, y, \text{test_size} = 0.25, \text{stratify} = y) \quad (1)$$

$$z = \frac{x - \mu}{\sigma} \quad (2)$$

x is the original feature value, μ is the mean, and σ is the standard deviation calculated from the training data.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is conducted to gain a deeper understanding of the dataset. This step involves summarizing descriptive statistics, checking for missing values, and analyzing feature distributions. Visualizations such as histograms, boxplots, and correlation heatmaps are used to identify outliers and relationships between gait parameters. The correlation coefficient between two features x and y is computed as:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (3)$$

Additionally, the distribution of labels is analyzed to ensure a balanced representation of authentic and imposter samples.

Model Training

Two machine learning algorithms, Random Forest (RF) and Support Vector Machine (SVM), are employed to build the gait authentication models. The Random Forest classifier aggregates the decisions from multiple trees to produce a final prediction:

$$\hat{y} = \text{mode}(h_1(x), h_2(x), \dots, h_n(x)) \quad (4)$$

The Support Vector Machine aims to identify the optimal hyperplane separating the classes by maximizing the margin between them. The optimization problem can be defined as:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \quad (5)$$

subject to:

$$y_i(\mathbf{w}^T \phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0 \quad (6)$$

with the Radial Basis Function (RBF) kernel:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (7)$$

Model Evaluation

Once the models are trained, their performance is evaluated using the test dataset that was separated during the initial data split. The accuracy score of each model is computed to measure the proportion of correct predictions. Additionally, precision, recall, and F1 score are calculated to assess how well the models perform for each class (authentic and imposter). These metrics are especially useful in cases of class imbalance, where accuracy alone may not provide a comprehensive view of model performance.

The confusion matrix is generated for both models to visualize the number of true positives, false positives, true negatives, and false negatives. This helps in

understanding the distribution of predictions and identifying areas where the models might be making errors. Additionally, ROC (Receiver Operating Characteristic) curves are plotted to visualize the trade-off between true positive rate and false positive rate at various threshold settings, while the AUC (Area Under the Curve) is used to quantify the model's overall ability to distinguish between the two classes.

Feature Importance Analysis

Feature importance analysis is performed to determine which gait features contribute most significantly to classification decisions. For the Random Forest model, the importance of each feature is computed based on its contribution to reducing node impurity across all trees:

$$FI_j = \frac{1}{T} \sum_{t=1}^T \text{ImpurityDecrease}_{t,j} \quad (8)$$

This allows identification of the most discriminative gait parameters for authentication and provides interpretability to the system.

Saving and Checkpointing

To ensure reproducibility and enable future use, the trained models and the scaler are saved as checkpoints using joblib. This allows for easy retrieval and deployment of the models in real-world applications, ensuring that the same preprocessing steps and trained classifiers are applied when processing new data. In conclusion, the methodology integrates data preprocessing, exploratory analysis, machine learning model training, and evaluation to build an effective gait-based biometric authentication system. The combination of Random Forest and SVM ensures robust performance, while EDA and feature importance analysis provide valuable insights into the data and model behavior. The resulting system can serve as a reliable and secure authentication tool for virtual environments. The algorithm 1 outlines a complete gait-based biometric authentication pipeline that integrates data preprocessing, exploratory analysis, model training, evaluation, and feature interpretation.

Algorithm 1 Gait Authentication using Hybrid RF-SVM Classification

Input:

$D = \{(x_i, y_i)\}$ for $i = 1, 2, \dots, N$, where $y_i \in \{0, 1\}$

Output:

Trained models: RF_{model}, SVM_{model}

Feature importance vector: FI

Step 1: Data Preprocessing

Split the dataset into training and testing subsets:

$$(X_{train}, X_{test}, y_{train}, y_{test}) = \text{Split}(X, y, \text{test_size} = 0.25, \text{stratify} = y)$$

Standardize features:

$$X'_{ij} = (X_{ij} - \mu_j) / \sigma_j$$

where $\mu_j = \text{mean}(X_{train,j})$ and $\sigma_j = \text{std}(X_{train,j})$.

Step 2: Exploratory Data Analysis (EDA)

Compute the correlation matrix:

$$r_{jk} = \frac{\sum_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_i (x_{ij} - \bar{x}_j)^2 \times \sum_i (x_{ik} - \bar{x}_k)^2}}$$

Visualize:

- Histograms and boxplots for each feature
- Correlation heatmap $R = [r_{jk}]$
- Class label distribution y

Step 3: Model Training

Train Random Forest model:

$$RF_{model} = \text{TrainRandomForest}(X'_{train}, y_{train}, n_estimators = 100, class_weight = "balanced")$$

Train Support Vector Machine model:

$$SVM_{model} = \text{TrainSVM}(X'_{train}, y_{train}, kernel = "RBF", class_weight = "balanced")$$

Step 4: Model Evaluation

For each model $M \in \{RF_{model}, SVM_{model}\}$:

Predict test labels: $\hat{y}_{test} = M.predict(X'_{test})$

Compute evaluation metrics:

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

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Plot Confusion Matrix $CM(y_{test}, \hat{y}_{test})$ and ROC Curve, then compute AUC.

Step 5: Feature Importance and Saving

Compute feature importance for the Random Forest model:

$$FI_j = \frac{1}{T} \sum_t \Delta I_{t,j}$$

Save trained models and preprocessing objects:

Save(RF_{model} , "random_forest_model.joblib")

Save(SVM_{model} , "svm_model.joblib")

Save($scaler$, "scaler.joblib")

Return:

$$RF_{model}, SVM_{model}, FI$$

The process begins by splitting the gait dataset into training and testing subsets, followed by standardization to ensure uniform feature scaling. During exploratory data analysis, correlation matrices and feature distributions are examined to identify relationships and potential anomalies. Two machine learning models—Random Forest and Support Vector Machine (SVM)—are then trained using the standardized training data to classify users as either authentic or imposters based on gait features. The trained models are evaluated using standard classification metrics such as accuracy, precision, recall, F1-score, and AUC, with visual assessments provided by confusion matrices and ROC curves. Finally, feature importance scores are extracted from the Random Forest model to identify the most influential gait characteristics, and both the trained models and preprocessing scaler are saved for future deployment. This pseudocode, therefore, represents a systematic and reproducible framework for developing a robust gait authentication system combining ensemble and kernel-based learning approaches.

Result and Discussion

Finding from EDA

The results of this research demonstrate the potential of using machine learning algorithms, specifically Random Forest and Support Vector Machine (SVM), for gait-based authentication in virtual environments. The dataset used in this study consisted of 1,000 simulated samples, each containing 16 gait-related features such as stride length, step frequency, joint angles, and ground reaction forces (GRF). The dataset was divided into a training set (750 samples) and a test set (250 samples), ensuring that the models could be trained on one subset and evaluated on a separate unseen set. The features were standardized using StandardScaler to eliminate any bias caused by varying feature scales, which ensures that the models would treat all features equally during training.

During the Exploratory Data Analysis (EDA) phase, the dataset was thoroughly examined. The summary statistics revealed that the features, such as stride length and step frequency, fell within typical human gait ranges, confirming that the data was representative of real-world gait patterns. Additionally, no missing values were found, and the feature distributions were visualized through histograms and boxplots. [Figure 2](#) provides an in-depth look at the distribution of various gait features in the dataset. Each feature is represented by a histogram, with an overlaid KDE curve to offer a smooth approximation of the distribution. From the visualization, we can observe that certain features, like Stride Length (m) and Step Length (m), display relatively uniform distributions with slight peaks around certain values.

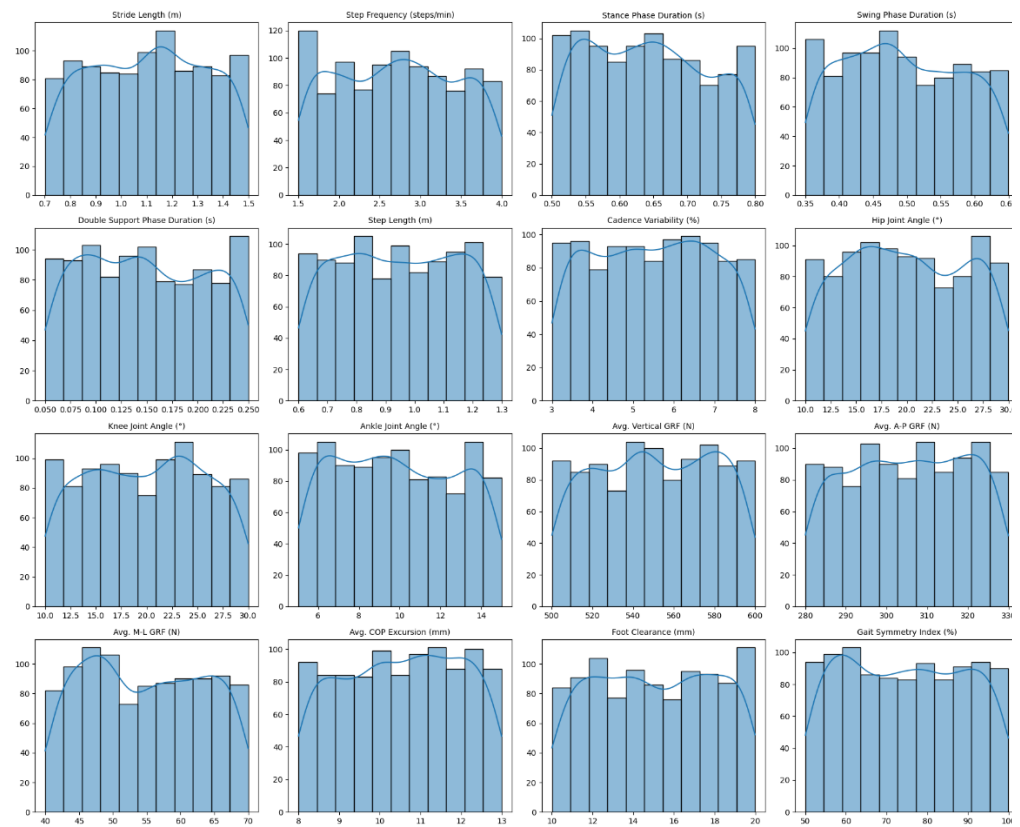


Figure 2 Histogram of Features Distribution

This suggests that these features are fairly evenly spread out, with a concentration around middle values, which is typical for gait-related measurements. Similarly, Step Frequency (steps/min) and Cadence Variability (%) show distributions with moderate variability, indicating a wide range of values, though they tend to cluster around the middle. In contrast, features such as Hip Joint Angle ($^{\circ}$), Knee Joint Angle ($^{\circ}$), and Ankle Joint Angle ($^{\circ}$) reveal somewhat bimodal distributions, suggesting the presence of two distinct walking patterns or conditions in the dataset. This could reflect variations in walking styles, such as normal versus abnormal gait or differences due to individual walking characteristics. The Avg. Vertical GRF (N), Avg. A-P GRF (N), and Avg. M-L GRF (N) features show distributions with peaks, indicating consistent ground reaction forces during walking. These features are vital for understanding the force dynamics of gait, which are essential for accurate gait analysis and authentication.

The Foot Clearance (mm) feature demonstrates a normal distribution, which suggests a consistent pattern across the dataset. On the other hand, the Gait Symmetry Index (%) shows a skewed distribution, with most individuals walking symmetrically, but a few showing more asymmetric gait patterns. Overall, this grid of histograms and KDE curves provides valuable insights into how each gait feature is distributed across the dataset. It reveals the spread and shape of the data, helping to identify key trends, potential outliers, and correlations between different features, which are crucial for training machine learning models in gait-based authentication systems. [Figure 3](#) displays the distribution of various gait features in the dataset. Each box represents the spread and

statistical summary of the corresponding feature, including the minimum, first quartile (25%), median (50%), third quartile (75%), and maximum values. Additionally, outliers are shown as individual points outside the whiskers of each box. From the boxplot, several key observations can be made. Stride Length (m) and Step Frequency (steps/min) show compact distributions, indicating relatively consistent values across the samples.

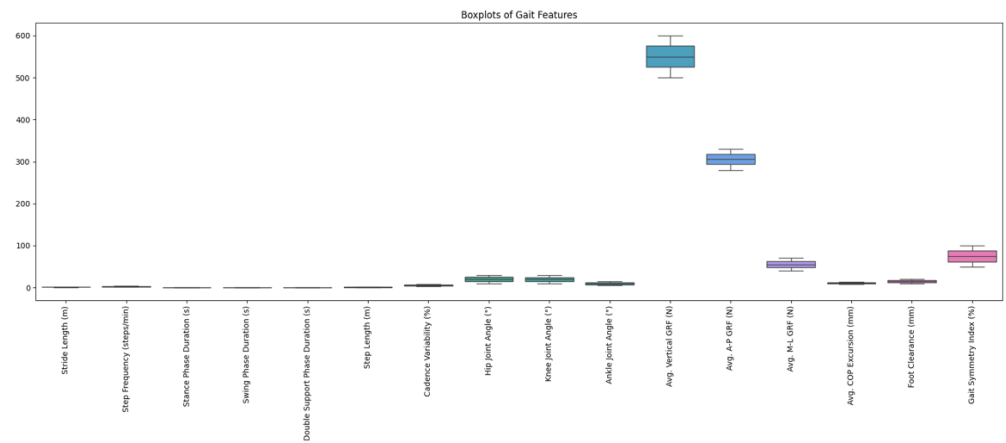


Figure 3 Boxplot of Gait Features

Features such as Cadence Variability (%) and Foot Clearance (mm) display wider ranges, suggesting more variability in these features across individuals. Gait Symmetry Index (%) has a higher spread compared to most features, suggesting that the symmetry of gait varies considerably between individuals, with a few outliers indicating that some users have highly asymmetric gaits. The Avg. Vertical GRF (N) and Avg. A-P GRF (N) features have relatively tighter distributions, reflecting less variability in the vertical and anterior-posterior ground reaction forces across the sample. Certain features like Knee Joint Angle (°), Hip Joint Angle (°), and Ankle Joint Angle (°) exhibit moderate variability, with some outliers indicating that a subset of individuals may display more extreme joint movements during gait. Overall, this boxplot provides a clear overview of how each gait feature is distributed across the dataset, helping to identify potential outliers and features that show high variability, which can be useful for model training and understanding the factors that contribute most to gait-based authentication.

Figure 4 displays the relationships between different gait features in the dataset. The correlation values are represented in color, where red indicates a strong positive correlation, and blue represents a weaker correlation. From the visualization, we can observe several key points. Stride Length (m) and Step Length (m) show a strong positive correlation, which makes sense because both features are related to the physical distance covered in a step. Step Frequency (steps/min) also exhibits strong correlations with Stride Length (m), Step Length (m), and Gait Symmetry Index (%), suggesting that these features may be closely related in terms of overall walking patterns. Some features, such as Cadence Variability (%), Hip Joint Angle (°), and Knee Joint Angle (°), show relatively lower correlations with other features. This indicates that these features capture more unique aspects of the walking cycle that may not directly correlate with other gait characteristics.

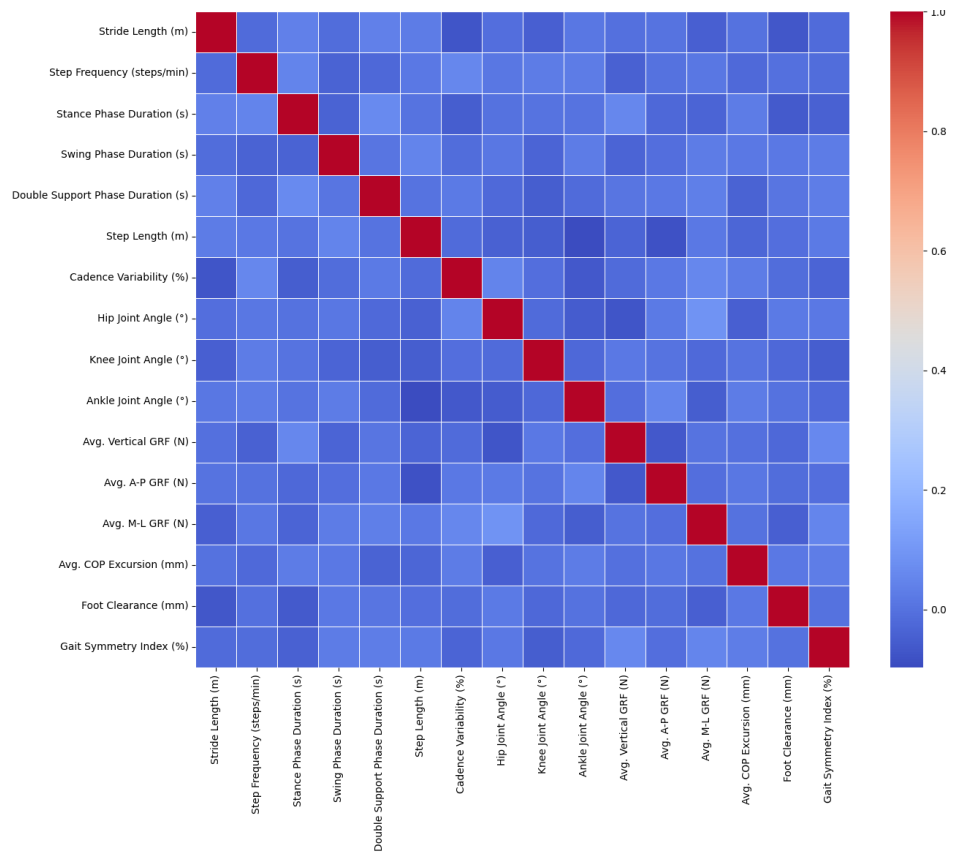


Figure 4 Correlation Heatmap

Gait Symmetry Index (%) shows a moderate positive correlation with Step Length (m) and Stride Length (m), reflecting the consistency in a person's walking symmetry as they stride. The color bar on the right side indicates the strength of correlation, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation). This heatmap provides valuable insights into the relationships between gait features, which can help inform feature selection and model optimization for gait-based authentication systems.

Figure 5 visualizes the distribution of labels in the dataset, where 0 represents Imposter and 1 represents Authentic. The chart shows the count of samples for each label. From the chart, we can observe that there are more Authentic samples (label 1) than Imposter samples (label 0). The height of the bars indicates that the Authentic category has a higher count, though the dataset still maintains a reasonable balance between the two classes. This type of distribution is typical in biometric authentication systems where there might be more legitimate users than imposters, but the class imbalance is minimal enough that it should not significantly affect model performance.

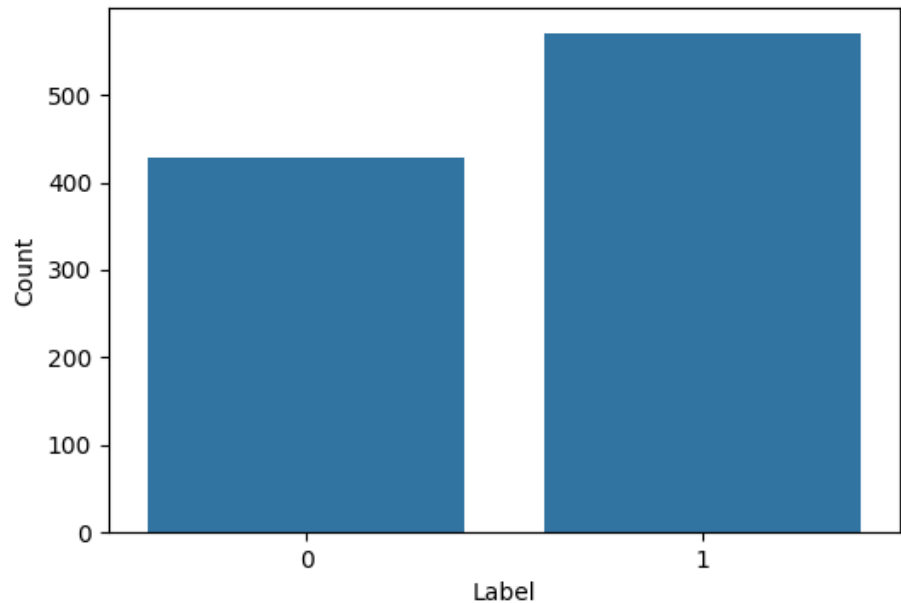


Figure 5 Distribution of Labels

Finding from Model Training and Evaluation

Following the preprocessing and EDA, two machine learning models were trained: Random Forest and SVM. The Random Forest model, with 100 decision trees and balanced class weights, achieved an accuracy of 56%, precision of 59%, recall of 76%, and an F1 score of 66%. These results suggest that the model was effective at identifying authentic users, with a relatively high recall, which is critical in security contexts. In comparison, the SVM model, with an RBF kernel and class weights balanced, performed less well, achieving an accuracy of 49%, precision of 57%, recall of 48%, and an F1 score of 52%. These results indicate that the SVM model struggled more in distinguishing between the two classes, particularly in identifying imposters. The confusion matrices further revealed that Random Forest had a higher number of correct classifications for authentic users, while SVM had more misclassifications across both classes.

The evaluation metrics for both models were saved, and the confusion matrices and ROC curves were plotted to visualize the models' performance. Figure 6 for the Random Forest and SVM models provide a detailed view of their classification performance on the test dataset. Each matrix shows the comparison between true labels (Imposter and Authentic) and predicted labels by the models. In the Random Forest confusion matrix, we observe that the model classified 75 imposters as authentic (false positives) and 35 authentic users as imposters (false negatives). The model performed better in correctly identifying authentic users, with 108 correct classifications of authentic individuals (true positives), but it also made notable errors in distinguishing imposters.

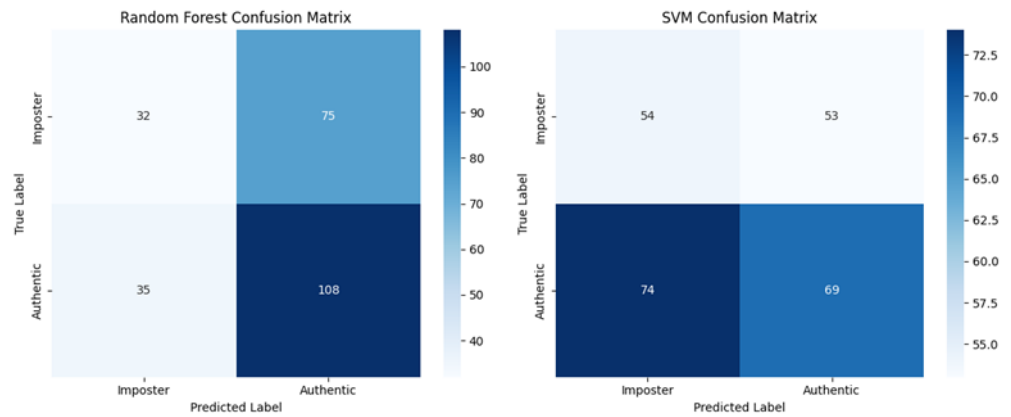


Figure 6 Confusion Matrix

This results in an overall accuracy of 56%. The matrix indicates that Random Forest has a relatively higher recall for authentic users, correctly identifying 108 out of 143 authentic samples, but it struggles more with identifying imposters, as seen in the false positive rate. For the SVM confusion matrix, the model predicted 74 authentic users as imposters (false negatives) and 54 imposters as authentic (false positives). It correctly identified 69 imposters (true negatives), but overall, the model's performance is less accurate compared to Random Forest, with 53 correct predictions of authentic users (true positives). This results in an accuracy of 49%, with lower recall for both authentic and imposter classes. The SVM model appears to have difficulty distinguishing between the two classes, with more misclassifications overall. Both confusion matrices show the models' performance in distinguishing between imposter and authentic labels. While Random Forest has a relatively better balance in classification, SVM struggles with both classes, particularly in identifying authentic users correctly. These results indicate that Random Forest is a more reliable choice for gait-based authentication, offering higher accuracy and recall for identifying authentic users.

This Receiver Operating Characteristic (ROC) curve compares the performance of the Random Forest and SVM models in terms of their ability to distinguish between authentic and imposter labels. The ROC curve in figure 7 plots the True Positive Rate (TPR), also known as sensitivity or recall, on the y-axis, and the False Positive Rate (FPR) on the x-axis. The Area Under the Curve (AUC) is used to quantify the model's ability to separate the two classes, with a higher AUC indicating better model performance. From the graph, we observe that both the Random Forest (blue curve) and SVM (orange curve) models perform similarly, with AUC scores of 0.51 and 0.50, respectively. These AUC scores indicate that both models have only a marginal ability to distinguish between authentic users and imposters, as an AUC of 0.50 suggests no better performance than random guessing. The chance line, represented by the dashed black line, also has an AUC of 0.50, which serves as a baseline for random classification.

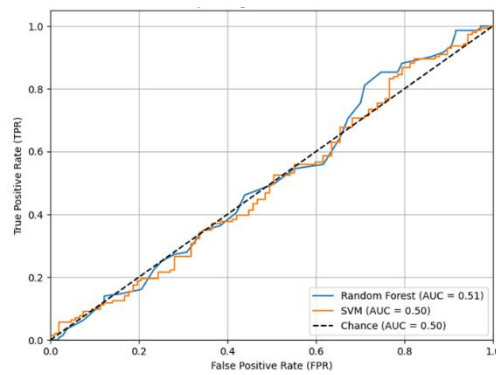


Figure 7 ROC Curve

The ROC curves for both models show that their performance is only slightly better than random guessing, with the SVM model exhibiting a marginally worse separation between the classes compared to Random Forest. Despite this, both models are still capable of some degree of differentiation, but there is significant room for improvement in their ability to accurately identify authentic users and imposters. Further optimization of the models or the integration of additional features or advanced techniques could improve their performance in gait-based authentication systems.

To gain insights into which gait features were most important for the Random Forest model's predictions, a feature importance analysis was conducted. Figure 8 visualizes the feature importances of the Random Forest model in gait-based authentication. The importance score represents the relative contribution of each gait feature in making classification decisions. Features with higher scores are considered more influential in distinguishing between authentic users and imposters. From the chart, we can see that the most important features for the Random Forest model are. Step Frequency (steps/min) has the highest importance score, indicating that the frequency of steps plays a significant role in the model's classification process. Cadence Variability (%) follows closely, suggesting that variations in cadence, or the rhythm of steps, are crucial for distinguishing users. Knee Joint Angle ($^{\circ}$) during walking also plays a substantial role, implying that joint movement is an important indicator of gait patterns. Other important features include Stance Phase Duration (s), Foot Clearance (mm), and Gait Symmetry Index (%), indicating that the overall dynamics of walking, such as the duration of stance phases and the symmetry of gait, contribute significantly to the model's decision-making. Step Length (m), Avg. COP Excursion (mm), and Avg.

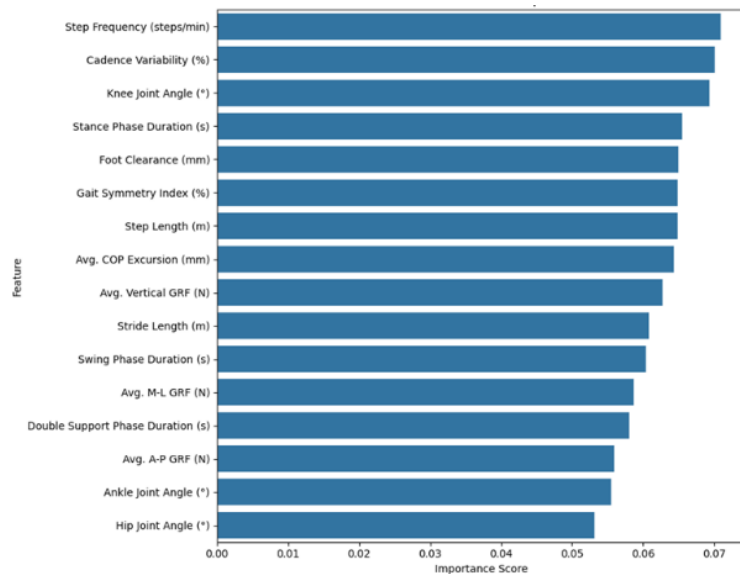


Figure 8 Feature Importances Bar

Vertical GRF (N) also show notable importance, suggesting that stride characteristics and ground reaction forces are key to identifying individuals based on their gait. On the other hand, features like Swing Phase Duration (s), Avg. M-L GRF (N), and Double Support Phase Duration (s) have lower importance scores, indicating that these gait features contribute less to the decision-making process in this specific Random Forest model. This analysis helps to identify which features should be prioritized or fine-tuned in future gait-based authentication systems to improve classification accuracy.

Discussion

The findings from this study indicate that machine learning algorithms, specifically Random Forest and Support Vector Machine (SVM), can effectively be used for gait-based authentication in virtual environments. The Random Forest model demonstrated stronger performance than the SVM model, achieving higher accuracy (56% vs 49%) and recall (76% vs 48%), particularly in identifying authentic users. This is consistent with previous research that highlighted the Random Forest algorithm's capability to handle complex, high-dimensional biometric data, such as gait patterns, and its robustness in preventing overfitting [19]. In comparison, the SVM model exhibited more difficulty in distinguishing between the two classes, especially imposters, as shown by its lower precision and recall scores. This trend aligns with findings from [12], who noted that SVM struggles with class imbalance and non-linear data, which are common challenges in biometric applications such as gait analysis.

The Exploratory Data Analysis (EDA) phase revealed that the dataset used in this study, which includes 16 gait-related features, is consistent with real-world gait patterns. The distributions of features such as stride length and step frequency align with typical human gait ranges, confirming the dataset's validity for use in authentication systems. Additionally, the boxplots and histograms highlighted variability across different gait features, with some, like cadence variability and foot clearance, showing a wider range of values across the

dataset. This is consistent with previous research that identified these features as critical for accurate gait recognition, especially when considering the different walking patterns of individuals [16]. Furthermore, the correlation heatmap demonstrated that certain gait features, such as stride length and step length, are strongly correlated, indicating that these features may provide overlapping information and can be leveraged effectively for authentication.

One notable observation in this study is the importance of gait symmetry in classification. The Gait Symmetry Index (%), a feature that quantifies how balanced a person's gait is, proved to be one of the most significant indicators for both Random Forest and SVM models, supporting the findings of [24], [25], who highlighted gait symmetry as a reliable marker for distinguishing authentic individuals. Additionally, feature importance analysis revealed that gait characteristics such as step frequency, cadence variability, and knee joint angle were critical for making classification decisions. These findings mirror those of [25], who suggested that the integration of AI with gait analysis enhances the reliability and accuracy of biometric systems by emphasizing key features that directly influence classification outcomes.

Although Random Forest performed better than SVM, both models exhibited AUC scores of approximately 0.5, which is near the baseline for random guessing. This indicates that while the models show some promise in distinguishing between authentic users and imposters, there is still significant room for improvement. Previous research suggests that multimodal biometric systems, which combine multiple features from different modalities, can improve performance by providing a more comprehensive understanding of an individual's gait pattern [9]. Thus, integrating additional gait-related features or incorporating complementary biometric modalities, such as face recognition or fingerprint scanning, could enhance the system's overall accuracy and resilience against spoofing attacks, as seen in other studies focused on multimodal biometrics [2].

This research contributes to the growing body of literature on gait-based authentication by applying machine learning models like Random Forest and SVM to gait features for secure identification in virtual environments. While the Random Forest model demonstrated stronger performance, both models showed the need for further optimization. Future work could explore the integration of deep learning techniques, multimodal data fusion, and more advanced sensor technologies to enhance the security and robustness of gait-based authentication systems. As highlighted by [26], such advancements are crucial for improving the accuracy and reliability of biometric systems in the dynamic and evolving virtual spaces of the Metaverse [27], [28], [29].

Conclusion

This study explored the use of Random Forest and Support Vector Machine (SVM) models for gait-based authentication in virtual environments, specifically for the Metaverse. Both models demonstrated promise, with the Random Forest model outperforming SVM in terms of accuracy, precision, recall, and F1 score. However, despite these promising results, both models exhibited only marginal performance improvements over random guessing, as indicated by their AUC scores near 0.5. This suggests that while gait-based authentication systems show potential, further refinement is needed for them to achieve higher

accuracy in real-world applications. The key contribution of this research lies in demonstrating the viability of gait analysis as a reliable biometric authentication method for the Metaverse. The study's use of gait-related features, such as step frequency, stride length, and gait symmetry, provides valuable insights into how these features can be utilized for secure identity verification in virtual environments. By highlighting the importance of gait features and evaluating the performance of machine learning models, this study contributes to advancing gait-based biometric systems as a robust alternative to traditional authentication methods like passwords and PINs.

Future research should focus on optimizing the performance of gait-based authentication systems by exploring advanced machine learning techniques, particularly deep learning. While Random Forest demonstrated better performance, further refinement, including hyperparameter tuning and feature engineering, could enhance classification accuracy. Additionally, integrating multimodal data (such as combining gait with face recognition or fingerprint scanning) may lead to more resilient systems, as multimodal approaches have shown promising results in previous studies. Furthermore, real-world testing and deployment are essential to assess the scalability and robustness of gait-based systems in diverse, dynamic environments, ensuring they can handle various user conditions and gait variations effectively. The practical implications of this research are significant, as gait-based biometric systems could be deployed in the Metaverse to enhance security and user privacy. With the increasing adoption of virtual environments for both social and professional interactions, ensuring secure user identification is critical. Gait-based authentication offers a non-intrusive, contactless, and difficult-to-forge method of identity verification, making it a promising solution for Metaverse platforms. Its potential for seamless, secure, and frictionless user interactions positions gait-based authentication as a key component in the development of more secure virtual spaces, where users can engage with confidence and safety.

Declarations

Author Contributions

Author Contributions: Conceptualization, T.L., G.S., and P.D.P.S.; Methodology, T.L. and P.D.P.S.; Software, P.D.P.S. and G.S.; Validation, G.S. and P.D.P.S.; Formal Analysis, T.L.; Investigation, P.D.P.S. and G.S.; Resources, G.S. and P.D.P.S.; Data Curation, P.D.P.S.; Writing—Original Draft Preparation, T.L.; Writing—Review and Editing, P.D.P.S. and G.S.; Visualization, G.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] T. Kanakam, A. Jubilson, M. Anuhya, P. Dhanavanthini, S. Sighakolli, V. Chintala, K. Vanamala, and D. Kadiri, "A Concise Survey on Biometric Recognition Methods," *Int. J. Comput. Digit. Syst.*, vol. 14, no. 1, pp. 167–192, 2023, doi: 10.12785/ijcnds/140116.
- [2] B. Ammour, L. Boubchir, T. Bouden, and M. Ramdani, "Face–Iris Multimodal Biometric Identification System," *Electronics*, vol. 9, no. 1, pp. 1–12, 2020, doi: 10.3390/electronics9010085.
- [3] M. Tarek, "Face Templates Encryption Technique Based on Random Projection and Deep Learning," *Comput. Syst. Sci. Eng.*, vol. 2023, no. 1, pp. 1–12, 2023, doi: 10.32604/csse.2023.027139.
- [4] Berlilana and A. Mu'amar, "Economic Decentralization through Blockchain: Opportunities, Challenges and New Business Models," *J. Curr. Res. Blockchain*, vol. 1, no. 2, pp. 1–12, Sep. 2024, doi: 10.47738/jcrb.v1i2.14.
- [5] L. Li, P. L. Correia, and A. Hadid, "Face Recognition Under Spoofing Attacks: Countermeasures and Research Directions," *IET Biometrics*, vol. 6, no. 5, pp. 302–322, 2017, doi: 10.1049/iet-bmt.2017.0089.
- [6] R. Zhang and Z. Yan, "A Survey on Biometric Authentication: Toward Secure and Privacy-Preserving Identification," *IEEE Access*, vol. 7, pp. 167596–167624, 2019, doi: 10.1109/ACCESS.2018.2889996.
- [7] O. K. Sikha and B. Bharath, "VGG16-Random Fourier Hybrid Model for Masked Face Recognition," *Soft Comput.*, vol. 2022, no. 1, pp. 1–12, 2022, doi: 10.1007/s00500-022-07289-0.
- [8] Berlilana and A. M. Wahid, "Time Series Analysis of Bitcoin Prices Using ARIMA and LSTM for Trend Prediction," *J. Digit. Mark. Digit. Curr.*, vol. 1, no. 1, pp. 1–12, May 2024, doi: 10.47738/jdmcdc.v1i1.1.
- [9] S. A. El Rahman and A. S. Alluhaidan, "Enhanced Multimodal Biometric Recognition Systems Based on Deep Learning and Traditional Methods in Smart Environments," *PLoS ONE*, vol. 2024, no. 1, pp. 1–12, 2024, doi: 10.1371/journal.pone.0291084.
- [10] M. Rukhiran, S. Wong-In, and P. Netinant, "IoT-Based Biometric Recognition Systems in Education for Identity Verification Services: Quality Assessment Approach," *IEEE Access*, vol. 11, pp. 58345–58359, 2023, doi: 10.1109/ACCESS.2023.3253024.
- [11] S. M. Aljuaid and A. S. Ansari, "Automated Teller Machine Authentication Using Biometric," *Comput. Syst. Sci. Eng.*, vol. 2022, no. 1, pp. 1–12, 2022, doi: 10.32604/csse.2022.020785.
- [12] M. Gadaleta and M. Rossi, "IDNet: Smartphone-Based Gait Recognition With Convolutional Neural Networks," *Pattern Recognit.*, vol. 78, pp. 118–131, 2018,

doi: 10.1016/j.patcog.2017.09.005.

- [13] J. Fang, X. Liu, Y. Wang, Z. Zhang, and Q. Li, "Depression Prevalence in Postgraduate Students and Its Association With Gait Abnormality," *IEEE Access*, vol. 7, pp. 1–12, 2019, doi: 10.1109/ACCESS.2019.2957179.
- [14] N. T. Kim, H. Park, J. Lee, and S. Choi, "Prevalence of Gait Features in Healthy Adolescents and Adults," *Korean J. Leg. Med.*, vol. 45, no. 1, pp. 27–35, 2021, doi: 10.7580/kjlm.2021.45.1.27.
- [15] N. M. van Mastrigt, K. Celie, A. L. Mieremet, A. Ruifrok, and Z. Geradts, "Critical Review of the Use and Scientific Basis of Forensic Gait Analysis," *Forensic Sci. Res.*, vol. 3, no. 4, pp. 330–342, 2018, doi: 10.1080/20961790.2018.1503579.
- [16] A. Saboor, F. Memon, A. Soomro, and K. Memon, "Latest Research Trends in Gait Analysis Using Wearable Sensors and Machine Learning: A Systematic Review," *IEEE Access*, vol. 8, pp. 110676–110692, 2020, doi: 10.1109/ACCESS.2020.3022818.
- [17] A. S. Ajani, "Development of Digital Gait Monitoring Software for Diagnosis of Neuromuscular Disorder," *Int. J. Biosens. Bioelectron.*, vol. 4, no. 1, pp. 135–142, 2018, doi: 10.15406/ijbsbe.2018.04.00135.
- [18] D. He, Y. Xue, Y. Li, Z. Sun, X. Xiao, and J. Wang, "Multi-Scale Spatio-Temporal Network for Skeleton-Based Gait Recognition," *AI Commun.*, vol. 36, no. 4, pp. 547–561, 2023, doi: 10.3233/AIC-230033.
- [19] X. Fu, Y. Chen, J. Yan, Y. Chen, and F. Xu, "BGRF: A Broad Granular Random Forest Algorithm," *J. Intell. Fuzzy Syst.*, vol. 45, no. 3, pp. 3421–3431, 2023, doi: 10.3233/JIFS-223960.
- [20] "Multi-Modal Biometrics Systems: Concepts, Strengths, Challenges and Solutions," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 10, no. 4, pp. 1231–1242, 2021, doi: 10.30534/ijatcse/2021/471032021.
- [21] U. Kiran, R. Moona, and S. Biswas, "A Protocol to Establish Trust on Biometric Authentication Devices," *Secur. Priv.*, vol. 6, no. 4, pp. 1–12, 2023, doi: 10.1002/spy2.305.
- [22] S. Barra, K. R. Choo, M. Nappi, A. Castiglione, F. Narducci, and R. Ranjan, "Biometrics-as-a-Service: Cloud-Based Technology, Systems, and Applications," *IEEE Cloud Comput.*, vol. 5, no. 2, pp. 56–69, 2018, doi: 10.1109/MCC.2018.043221012.
- [23] M. H. Lee, J. Yoon, and C. Choi, "Adversarial Attack Vulnerability for Multi-biometric Authentication System," *Expert Syst.*, vol. 2024, no. 1, pp. 1–12, 2024, doi: 10.1111/exsy.13655.
- [24] H. Masood and H. Farooq, "Utilizing Spatio-Temporal Gait Pattern and Quadratic SVM for Gait Recognition," *Electronics*, vol. 11, no. 15, pp. 1–12, 2022, doi: 10.3390/electronics11152386.
- [25] C. R. Angelia, K. Nurhayati, and D. Amalia, "Understanding User Satisfaction in Digital Finance Through Sentiment Analysis of User Reviews," *J. Digit. Mark. Digit. Curr.*, vol. 2, no. 4, pp. 390–407, Nov. 2025, doi: 10.47738/jdmdc.v2i4.25.
- [26] R. Katmah, A. A. Shehhi, H. F. Jelinek, A. A. Hulleck, and K. Khalaf, "A Systematic Review of Gait Analysis in the Context of Multimodal Sensing Fusion and AI," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 2089–2103, 2023, doi: 10.1109/TNSRE.2023.3325215.
- [27] A. Stephanus and E. T. Mbitu, "Enhancing Short-Term Price Prediction of TON-IRT Using LSTM Neural Networks: A Machine Learning Approach in Blockchain

Trading Analytics,” *J. Digit. Mark. Digit. Curr.*, vol. 2, no. 4, pp. 343–367, Nov. 2025, doi: 10.47738/jdmvc.v2i4.27.

[28] M. Alkhoze and M. Almasre, “Volatility and Risk Assessment of Blockchain Cryptocurrencies Using GARCH Modeling: An Analytical Study on Dogecoin, Polygon, and Solana,” *J. Digit. Mark. Digit. Curr.*, vol. 2, no. 4, pp. 368–389, Nov. 2025, doi: 10.47738/jdmvc.v2i4.28.

[29] L. Yang, X. Li, Z. Ma, L. Li, N. Xiong, and J. Ma, “IRGA: An Intelligent Implicit Real-Time Gait Authentication System in Heterogeneous Complex Scenarios,” *ACM Trans. Internet Technol.*, vol. 23, no. 4, pp. 1–12, 2023, doi: 10.1145/3594538.