SIN-CHN
SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2024, 21(3), 612.
https://doi.org/10.62617/mcb612

Article

Predicting career development paths of college students using biomechanical and behavioral data with machine learning

Xue Xiang

Department of Tourism and Management, Wuhan College of Foreign Languages & Foreign Affairs, Wuhan 430083, China;
xuexiang61@outlook.com

CITATION

Xiang X. Predicting career development paths of college students using biomechanical and behavioral data with machine learning. Molecular & Cellular Biomechanics. 2024; 21(3): 612. https://doi.org/10.62617/mcb612

ARTICLE INFO

Received: 25 October 2024
Accepted: 1 November 2024
Available online: 5 December 2024

Abstract: Accurately predicting career development paths is crucial to enhancing educational guidance and aligning student outcomes with labor market demands. This study presents a novel approach that integrates biomechanical and behavioral data with machine learning techniques to forecast career paths for college students. Using a dataset of 150 students, the study examines key biomechanical variables, such as joint angles, gait parameters, and ground reaction forces, alongside behavioral traits, including confidence levels, engagement, and personality. A Random Forest model was employed to analyze these multidimensional data and identify patterns predictive of career outcomes. The model achieved % overall accuracy of 82.57%, with individual performance metrics across four career categories showing substantial precision and recall. Integrating biomechanical and behavioral factors improved the model's predictive power, demonstrating that physical attributes, when combined with traditional behavioral data, provide a more comprehensive understanding of career suitability. These findings have significant implications for career counseling, educational interventions, and workforce development, offering a data-driven approach to support students in making informed career decisions.

Keywords: biomechanical and behavioral data; data-driven approach; physical attributes; precision; recall; biomechanical variables

1. Introduction

Predicting career development paths has long been a focus of educational research, as understanding the factors influencing students' career decisions can improve guidance systems, increase employability, and align educational outcomes with labor market demands [1,2]. Traditionally, career predictions have relied heavily on psychological and behavioral assessments, such as personality traits, academic performance, and career aspirations [3,4]. However, advanced data collection technologies, combined with machine learning, have opened new avenues for analyzing diverse datasets, offering more nuanced predictions [5,6]. Integrating biomechanical data with behavioral and psychological variables presents a novel and comprehensive approach to career path prediction [7].

Biomechanical data refers to individuals' physical attributes and movement patterns, such as joint angles, gait characteristics, and postural stability, which can offer insights into a person's physical behavior and capabilities [8]. Traditionally studied in fields such as sports science and rehabilitation, these factors are now being applied in broader contexts, including career development [9]. Certain professions demand cognitive skills and specific physical attributes [10,11]. For instance, healthcare, engineering, or performance arts careers may require higher levels of physical endurance, motor skills, or ergonomic awareness. Therefore, biomechanical

factors can be critical predictors of career success in physically demanding professions [12].

On the other hand, behavioral data—including traits like engagement, confidence levels, and personality—remains a strong predictor of career choices [13]. Behavior-driven factors reflect how individuals interact with their environment, engage in teamwork, and respond to challenges [14,15]. Traits such as extraversion, conscientiousness, and openness have been widely documented in career literature, correlating with job satisfaction, career stability, and leadership potential [16]. When combined with biomechanical data, these behavioral insights can create a holistic profile of an individual, offering a more complete understanding of career suitability [17].

This study leverages Machine Learning (ML), particularly the Random Forest (RF) model, to integrate biomechanical and behavioral data for career path prediction. Machine learning models, particularly Random Forest, excel in identifying patterns across large, multidimensional datasets, where traditional linear models may struggle [18,19]. The RF is an ensemble learning technique that builds multiple Decision Trees (DT) and aggregates their results to improve predictive accuracy while controlling for overfitting [20–25]. Its ability to handle both continuous and categorical data and its robustness to noise and complex interactions makes it an ideal candidate for this research [26–30].

This study aims to develop a predictive model that can accurately forecast career paths for college students based on a combination of biomechanical and behavioral data. By focusing on key features such as joint angles, gait parameters, confidence levels, and engagement metrics, this research aims to uncover the multidimensional factors influencing career development. A dataset of 150 college students was collected to achieve this, incorporating physical assessments and self-reported behavioral data. The study hypothesizes that integrating these diverse data types will enhance the precision of career predictions, offering a more detailed understanding of how physical and behavioral traits combine to shape career trajectories.

The rest of the paper is organized as follows: Section 2 presents the methodology, Section 3 presents the ML and its training, Section 4 analyzes the results, and Section 5 concludes the paper

2. Methodology

2.1. Participants

The study involved 150 college students from three universities in urban areas of China, focusing on diverse fields of study, including engineering, business, humanities, and health sciences. The participants were recruited through online announcements and campus flyers, ensuring a broad representation of the student population. The demographic characteristics of the participants were as follows: 55% (n = 82) were male, and 45% (n = 68) were female, reflecting a slight male predominance commonly observed in engineering disciplines. Participants ranged from 18 to 24 years, with a mean age of 20.5 (SD = 1.5).

In terms of academic year, 40% (n = 60) were first-year students, 30% (n = 45) were second-year students, 20% (n = 30) were third-year students, and 10% (n = 15) were in their final year. This distribution allowed for examining how career development paths evolve throughout the college experience. The participants came from various socioeconomic backgrounds, with 35% (n = 52) identifying as low-income, 45% (n = 67) as middle-income, and 20% (n = 31) as high-income families. This variety was crucial for understanding the influence of socioeconomic status on career choices and opportunities. Regarding extracurricular involvement, 60% (n = 90) of participants reported being active in student organizations or clubs, which provided additional data on their behavioral patterns and social interactions. This involvement is often correlated with leadership skills and teamwork experiences, which are critical factors in career development.

2.2. Apparatus

The study used advanced technologies and equipment to collect biomechanical and behavioral data from the participants.

The Primary Apparatus Included:

i. Motion Capture System: A 16-camera motion capture system (OptiTrack Flex 13) was employed to capture the participants' movements with high precision. This system enabled the analysis of various biomechanical parameters, such as joint angles, gait patterns, and postural stability. Markers were placed on key anatomical landmarks (e.g., hips, knees, ankles) to facilitate accurate tracking during specific tasks simulating career-related activities, such as presentations and collaborative discussions [31–34].

ii. Wearable Sensors: Each participant was equipped with a set of inertial measurement units (IMUs) attached to their lower limbs and torso. These sensors provided real-time data on acceleration, angular velocity, and orientation, allowing for the assessment of dynamic movements and fatigue levels during various physical tasks. The data collected by these sensors were critical for understanding how physical behavior might correlate with career path preferences and performance.

iii. Behavioral Assessment Software: To evaluate behavioral aspects, a custom-designed software application was utilized to track participants' interactions in simulated environments. This application recorded speech patterns, response times, and engagement levels during group discussions and presentations. The software integrated feedback mechanisms to analyze the students' confidence levels and communication skills, which are essential career development components.

iv. Survey Instruments: Participants completed online questionnaires to gather demographic information, academic background, and career aspirations. The surveys also included validated scales to measure personality traits, motivation levels, and perceived career readiness. This multifaceted approach ensured a comprehensive understanding of each participant's behavioral tendencies and aspirations.

v. Data Analysis Software: Software tools such as Python were employed to process and analyze the collected data. The Python ecosystem, particularly libraries like Scikit-learn, Pandas, and NumPy, facilitated the implementation of the RF-ML to model and predict potential career paths based on the integrated biomechanical and behavioral data.

2.3. Measurements and variables

This study employed a range of measurements and variables to capture college students' biomechanical and behavioral data, which are essential for predicting their career paths. The variables were categorized into two main groups: biomechanical variables and behavioral variables.

i) Biomechanical Variables

Biomechanical data were collected through the motion capture system and wearable sensors.

The key biomechanical variables included:

- Joint Angles: Measurements of hip, knee, and ankle joint angles during various tasks, recorded in degrees. These angles were analyzed to assess postural alignment and stability, which are crucial in understanding physical behavior in career-related activities.
- Gait Parameters: Stride length, cadence, and speed were measured during walking tasks. These metrics provided insights into the participants' mobility and physical fitness levels.
- Ground Reaction Forces (GRF): Collected using force plates, GRF measurements indicated the force exerted by participants during movements, which is essential for analyzing the impact of biomechanical factors on performance.
- Fatigue Levels: Monitored through changes in gait and joint angles over time. A fatigue index was calculated based on the deviation of performance metrics from baseline measurements during physical tasks.

ii) Behavioral Variables

Behavioral data were gathered through the behavioral assessment software and online questionnaires.

The key behavioral variables included:

- Engagement Levels: Measured during group discussions and presentations through metrics such as speaking time, number of contributions, and response rates. These variables reflected the participants' active involvement in social interactions, which is critical for career development.
- Confidence Levels: Assessed using self-reported scales where participants rated their confidence in various career-related tasks (e.g., public speaking, teamwork) on a Likert scale ranging from 1 (not confident) to 5 (very confident).
- Personality Traits: Measured using a validated questionnaire based on the Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism). Each trait was scored on a scale of 1 to 5, providing insights into how personality influences career choices.

- Career Aspirations: Participants provided information on their desired career paths and goals through open-ended responses, which were later categorized for analysis.

iii) Data Integration

From **Table 1**, the integration of biomechanical and behavioral variables created a comprehensive dataset for analysis. Each participant's biomechanical measurements were linked to their corresponding behavioral data, allowing for a multidimensional analysis of the factors influencing career development paths. The collected variables provided a robust framework for training the RF, enabling predictions of potential career trajectories based on physical and behavioral characteristics.

Table 1. Measurements and variables used in the study.

Variable Category	Variable	Description	Unit
Biomechanical Variables	Joint Angles	Measurements of hip, knee, and ankle angles	Degrees (°)
	Gait Parameters	Stride length, cadence, and speed	Meters (m), steps/min
	Ground Reaction Forces (GRF)	Force exerted during movements	Newtons (N)
	Fatigue Levels	Changes in gait and joint angles over time	Index value (unitless)
Behavioral Variables	Engagement Levels	Metrics from group discussions and presentations	Count (number of contributions)
	Confidence Levels	Self-reported confidence in career-related tasks	Likert scale (1–5)
	Personality Traits	Scores based on the Big Five personality traits	Likert scale (1–5)
	Career Aspirations	Desired career paths and goals	Categorical (text)

2.4. Experimental design and data collection

The experimental design employed in this study aimed to investigate the relationship between biomechanical and behavioral factors and their influence on the career development paths of college students. A mixed-methods approach was utilized, integrating quantitative measurements from biomechanical assessments and qualitative insights from behavioral evaluations.

2.4.1. Experimental design

The study utilized a cross-sectional design, where data were collected from participants simultaneously. This approach allowed for a comprehensive analysis of the interactions between physical movements, behavioral patterns, and career aspirations. Participants engaged in tasks designed to simulate common scenarios encountered in professional environments, such as group discussions, presentations, and collaborative problem-solving activities. These tasks were structured to elicit biomechanical and behavioral responses, enabling a thorough analysis of how these factors interplay in real-world contexts.

2.4.2. Data collection process

Data collection occurred over four weeks at three universities in urban areas of China. The recruitment process involved announcements and flyers distributed across various departments to ensure a diverse representation from different academic disciplines. Participants were carefully selected to include a balance of genders, ages, and fields of study, aiming for a sample that reflected the broader college student population.

Once potential participants expressed interest, they were invited to an orientation session, where the study's purpose, procedures, and expectations were explained in detail. This session emphasized the importance of their participation in contributing to a deeper understanding of how physical and behavioral factors influence career development. Each participant was required to provide informed consent, confirming their willingness to participate voluntarily.

- Instruction for Participants: Before the data collection sessions, participants were given detailed instructions on the tasks they would be performing. They were informed that the study would involve physical activities and social interactions to simulate real-world scenarios. Specifically, participants were instructed to:
- Prepare for Activities: Wear comfortable clothing suitable for movement and footwear that would allow for easy mobility. They were advised to avoid wearing accessories that could interfere with the motion capture system.
- Engage Fully: During the simulations, participants were encouraged to engage authentically in discussions and presentations, treating the activities as they would in a professional setting. This was emphasized to ensure that the behavioral data collected would accurately reflect their typical responses and interactions.
- Provide Honest Feedback: After completing the activities, participants were asked to complete questionnaires measuring their confidence levels, personality traits, and career aspirations. They were encouraged to answer these questions honestly and thoughtfully, as the information would be crucial for analyzing the results.

The Motion Capture System (MCS) (OptiTrack Flex 13) was set up in a controlled environment to capture participants' movements during the task simulations. Markers were affixed to specific anatomical landmarks to enable accurate tracking. During dynamic movements, inertial Measurement Units (IMU) were also used to collect additional biomechanical data, such as acceleration and angular velocity. Participants received brief training on wearing and adjusting the sensors to ensure comfort and accuracy during data collection.

The behavioral assessment software recorded participants' interactions during the tasks. Metrics such as speaking time, engagement levels, and response rates were automatically captured. Participants also completed self-reported questionnaires measuring confidence levels, personality traits, and career aspirations. These questionnaires were administered online, ensuring anonymity and encouraging candid responses. To assess fatigue levels, participants performed a series of physical tasks that increased in complexity and duration. Their biomechanical data

were analyzed to detect any changes in movement patterns and joint angles indicative of fatigue.

Ethical approval was obtained from the institutional review boards of the participating universities. Participants were informed of their right to withdraw from the study without penalty. Data confidentiality was maintained by anonymizing responses and securely storing all collected data.

3. Machine learning RF

The RF (**Figure 1**) is an ensemble learning method primarily used for classification and regression tasks. It builds multiple DTs during training and merges their predictions to produce more accurate and stable results. The model leverages the principles of bagging (bootstrap aggregating) to enhance predictive performance and control overfitting. Below is a detailed description of the RF, including relevant expressions and equations.

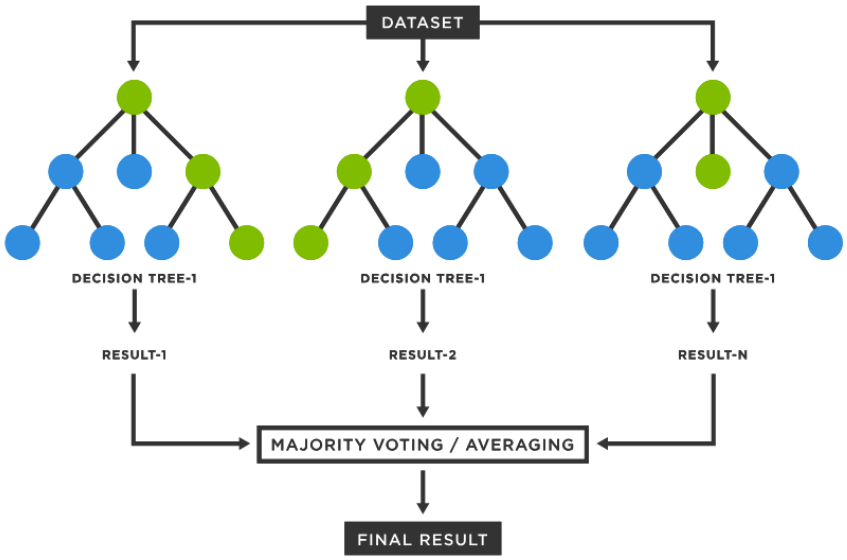

Figure 1. RF architecture.

The RF constructs a collection of decision trees from a training dataset. Each tree is trained on a bootstrapped sample of the data, meaning that each tree is trained on a random subset of the training data selected with replacement. This randomness helps reduce the model's variance and improve generalization to unseen data.

Given a dataset D with N instances, a bootstrap sample D_b for a single tree can be generated as follows:

$$D_b = \{(x_i, y_i) \mid i \in random\_\text{sample (D, N)}\} \quad (1)$$

where x_i represents the feature vector and y_i represents the target variable. This process is repeated $'B'$ times to create $'B'$ different bootstrap samples, resulting in $'B'$ DT. For each bootstrap sample D_b, a DT, 'T_b' is constructed using a random subset of features m at each node. The splitting criterion used is Gini impurity, which is defined as:

$$\text{Gini}\,(D) = 1 - \sum_{j=1}^{C} \left(p_j\right)^2 \tag{2}$$

where p_j is the proportion of instances belonging to class j in the dataset D, and C is the total number of classes. The tree construction process involves selecting the feature that results in the most significant reduction in Gini impurity at each node.

At each node of the DT, a random subset of m features is selected (where $m < M$, and M is the total number of features). This selection is crucial for ensuring that the trees are decorrelated, which helps improve the robustness of the model. The splitting of a node is performed using the best feature from this random subset based on Gini impurity.

Once all B decision trees are constructed, predictions for a new instance x are obtained by aggregating the predictions from all trees. For classification tasks, the final prediction is determined by the majority vote:

$$\hat{y} = \text{mode}\,(T_1(x), T_2(x), \ldots, T_B(x)) \tag{3}$$

where $T_b(x)$ is the prediction of tree b for the input instance x.The performance of the RF is evaluated using classification metrics such as accuracy, precision, recall, and F1-score. Cross-validation techniques are typically used to assess the model's generalizability.

Accuracy is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

where:

- $TP =$ True Positives
- $TN =$ True Negatives
- $FP =$ False Positives
- $FN =$ False Negatives

One of the advantages of RF is its ability to provide insights into the importance of features. The importance of a feature j can be assessed based on the decrease in Gini impurity that occurs when splitting nodes using that feature across all trees:

$$\text{Importance}_j = \sum_{b=1}^{B} \sum_{n \in \text{Nodes}\,(T_b)} \Delta\text{Gini}\,(n, j) \tag{5}$$

where $\Delta\text{Gini}\,(n, j)$ represents the reduction in Gini impurity when feature j is used to split node n.

The training process for the RF involves several critical steps to ensure optimal performance and generalization to unseen data. This section outlines the procedures for training the model, including the selection of hyperparameters, their significance, and the methods used for tuning them.

The data preparation phase is essential before training the RF. Initially, the dataset is divided into training and testing sets, typically using a split of 70% for training and 30% for testing. This division ensures that the model can be trained on one subset of data while being evaluated on a separate subset to assess its predictive

capabilities. Feature Selection (FS) is also conducted during this phase, as the high dimensionality of biomechanical and behavioral data may lead to inefficiencies. Techniques such as Recursive Feature Elimination (RFE) or preliminary RF runs can guide the selection of relevant features.

From **Table 2**, during the training of the RF, multiple DTs are constructed based on the training data. For each bootstrap sample created from the training dataset, a DT is built independently. The randomness introduced during training, in terms of data selection and FS, contributes to the diversity of the model. Additionally, parallelization is employed during this phase, allowing the trees to be trained concurrently since each tree is constructed independently. This approach accelerates the training process and makes it feasible to work with larger datasets.

Table 2. hyperparameters for the RF.

Hyperparameter	Description	Typical Values
N_Estimators	Number of trees in the forest.	100–500
Max_Depth	Maximum depth of each tree. Limiting this can prevent overfitting.	10–30
Min_Samples_Split	A minimum number of samples is required to split an internal node.	2–10
Min_Samples_Leaf	A minimum number of samples is required to be at a leaf node.	1–5
Max_Features	There are several features to consider when looking for the best split at each node.	“sqrt”, “log2”, or a fraction
Bootstrap	Whether bootstrap samples are used when building trees (sampling with replacement).	True (default)

The performance of the RF is significantly influenced by various hyperparameters that need careful tuning. Key hyperparameters include the number of trees (N_Estimators), which determine how many DTs are included in the forest. A higher number of trees generally improves model performance but increases computational cost. Common practice is to start with a value between 100 and 500 and adjust based on performance. Another important hyperparameter is maximum depth (Max_Depth), which controls the maximum depth of each decision tree. Limiting the depth can prevent overfitting, especially when dealing with complex datasets with typical values ranging from 10 to 30.

Minimum Samples Split (Min_Samples_Split) specifies the minimum number of samples required to split an internal node, helping to reduce overfitting by ensuring each split is based on sufficient observations. This value typically ranges from 2 to 10. Minimum samples leaf (Min_Samples_Leaf) sets the minimum number of samples that must be present in a leaf node, which helps create more robust models by ensuring that leaf nodes have enough samples for reliable predictions. Typical values are between 1 and 5. Maximum features (Max_Features) determine the number of features to consider when looking for the best split at each node. Options include “sqrt” (the square root of the number of features), “Log2,” or a specific integer or fraction of features, with “sqrt” often recommended for classification tasks. Lastly, the bootstrap parameter indicates whether bootstrap samples are used when building trees, and setting this to true enables the model to

draw samples with replacement, which is standard practice in Random Forest training.

Hyperparameter tuning is critical for optimizing the RF's performance. One common method is grid search, where a predefined set of hyperparameter values is specified, and the model is trained and evaluated for each combination. The best combination is selected based on cross-validated performance metrics. Alternatively, random search involves sampling from the hyperparameter space instead of testing all combinations, often proving more efficient, especially with many hyperparameters. K-fold cross-validation is used during the tuning process to assess model performance. The dataset is divided into k subsets, and the model is trained k times, each time using a different subset as the validation set while the remaining data is used for training. This process helps ensure that the model generalizes well to unseen data.

After training and tuning the RF, its performance is evaluated using the testing dataset. Metrics such as accuracy, precision, recall, and F1-score for classification tasks and mean squared error for regression tasks are calculated to assess the model's predictive capabilities.

4. Results

The analysis of biomechanical data (**Table 3**) provides key insights into the participants' physical characteristics and movement patterns. The joint angles at the hip, knee, and ankle, crucial for understanding postural alignment and mobility, show a moderate variation across the cohort. Specifically, the hip joint angle has a mean of 46.92° with a standard deviation of 7.81°, indicating that the majority of participants exhibit similar postural patterns, though there are some outliers (ranging from 32.34° to 61.18°). Similarly, the knee joint angle shows a broader range, with a mean of 78.63° and a higher standard deviation (10.24°), suggesting more significant variability in knee flexion during tasks, which could be influenced by task complexity or fatigue. The ankle joint angle, with a mean of 32.57° and a standard deviation of 6.18°, reflects less variability, which may indicate consistency in foot positioning and lower limb stability across participants.

Table 3. Descriptive statistics of the data.

Variable	Mean	Standard Deviation (SD)	Range
Biomechanical Variables			
Joint Angles (hip) (°)	46.92	7.81	32.34–61.18
Joint Angles (knee) (°)	78.63	10.24	57.29–97.76
Joint Angles (ankle) (°)	32.57	6.18	21.43–43.71
Stride Length (m)	1.38	0.27	0.92–1.88
Cadence (steps/min)	114.22	12.47	87.34–137.89
Walking Speed (m/s)	1.54	0.28	0.98–2.03
Ground Reaction Forces (N)	981.56	79.61	782.91–1143.76
Fatigue Index (unitless)	0.74	0.12	0.50–0.98

Table 3. (*Continued*).

Variable	Mean	Standard Deviation (SD)	Range
Behavioral Variables			
Engagement Levels (count)	5.83	2.11	1.22–9.89
Confidence Levels (1–5)	3.73	0.78	1.43–4.87
Openness (1–5)	4.12	0.57	2.78–4.98
Conscientiousness (1–5)	3.89	0.64	2.36–4.93
Extraversion (1–5)	3.22	0.82	1.14–4.71
Agreeableness (1–5)	3.94	0.63	2.46–4.85
Neuroticism (1–5)	2.54	0.91	1.04–4.56
Demographic Breakdown			
Male Participants (%)	55%	-	-
Female Participants (%)	45%	-	-
Academic Year (%)*	-	-	-
1st Year Students	40%	-	-
2nd Year Students	30%	-	-
3rd Year Students	20%	-	-
4th Year Students	10%	-	-
Socioeconomic Status (%)			
Low Income	35%	-	-
Middle Income	45%	-	-
High Income	20%	-	-

Stride length and cadence reflect the participants' gait characteristics. The average stride length is 1.38 m, with moderate variability (SD = 0.27 m), suggesting differences in walking styles or leg length among participants. The cadence (steps per minute) shows more significant variability, with a mean of 114.22 steps/min and a standard deviation of 12.47 steps/min, ranging from 87.34 to 137.89 steps/min. This wide range could be influenced by varying levels of physical fitness, with a faster cadence indicating more dynamic movement patterns.

The walking speed, with a mean of 1.54 m/s (SD = 0.28 m/s), falls within the typical range for young adults, though some participants exhibited slower speeds, which may indicate fatigue or less physical activity. The ground reaction forces (GRF), which measure the force exerted during movement, averaged 981.56 N, with variability (SD = 79.61 N), indicating a mix of participants exerting different levels of force during tasks, perhaps reflecting differences in weight or movement intensity. Finally, the fatigue index, with a mean of 0.74 (SD = 0.12) and a range from 0.50 to 0.98, suggests that participants experienced varying levels of fatigue, which may have affected their movement patterns and stability.

Behavioral characteristics play a significant role in predicting career paths. Participants' engagement levels, which measure their active involvement in group discussions and presentations, showed moderate variation, with a mean of 5.83 (SD = 2.11) and a wide range from 1.22 to 9.89. This suggests that while many participants were highly engaged, others may have been less involved, which could

be linked to their confidence levels or personal traits. The confidence levels, measured on a Likert scale, averaged 3.73 (SD = 0.78), indicating that most participants felt moderately confident during career-related tasks, though some outliers demonstrated lower or higher self-assessed confidence.

The analysis of personality traits provides further insights. Openness, with a mean of 4.12 (SD = 0.57), was the highest-rated personality trait, suggesting that participants generally perceived themselves as open to new experiences and learning. Conscientiousness also ranked highly, with a mean of 3.89 (SD = 0.64), reflecting the participants' tendency towards discipline and task responsibility. Extraversion, however, was rated slightly lower, with a mean of 3.22 (SD = 0.82), indicating that while some participants were outgoing and sociable, others were more reserved during group activities. Agreeableness, with a mean of 3.94 (SD = 0.63), reflects a positive tendency towards cooperation and teamwork among the participants, which is crucial in career-related activities. Neuroticism, with a mean of 2.54 (SD = 0.91), was the lowest-scoring trait, indicating that most participants exhibited emotional stability during tasks, though a subset experienced higher stress or anxiety levels.

The demographic breakdown reveals the composition of the participant group. The gender distribution was reasonably balanced, with 55% male and 45% female participants. This balanced distribution provides a representative analysis of how gender may influence biomechanical and behavioral career development factors. Regarding the academic year, the most significant proportion of participants were first-year students (40%), followed by second-year students (30%), third-year students (20%), and a smaller group of fourth-year students (10%). This distribution explores how career development perceptions change as students' progress through their academic journeys. In terms of socioeconomic status, most participants identified as middle-income (45%), followed by low-income (35%) and high-income (20%) groups. This variety in socioeconomic backgrounds is vital for analyzing how access to resources and opportunities may influence career aspirations and readiness.

The performance of the RF (**Table 4** and **Figure 2**) in predicting career development paths is notably strong, with an overall accuracy of 82.57% across the four career categories. This indicates that the model could correctly predict career paths for most participants, demonstrating its effectiveness in handling the complexity of the biomechanical and behavioral data.

Table 4. Model performance metrics of the RF.

Metric	Career Category 1	Career Category 2	Career Category 3	Career Category 4	Overall/Weighted Average
Accuracy	84.12%	81.75%	80.56%	85.89%	82.57%
Precision	84.21%	81.35%	78.62%	85.76%	82.36%
Recall	81.93%	79.12%	76.45%	87.28%	81.20%
F1-Score	83.05%	80.21%	77.51%	86.51%	81.82%

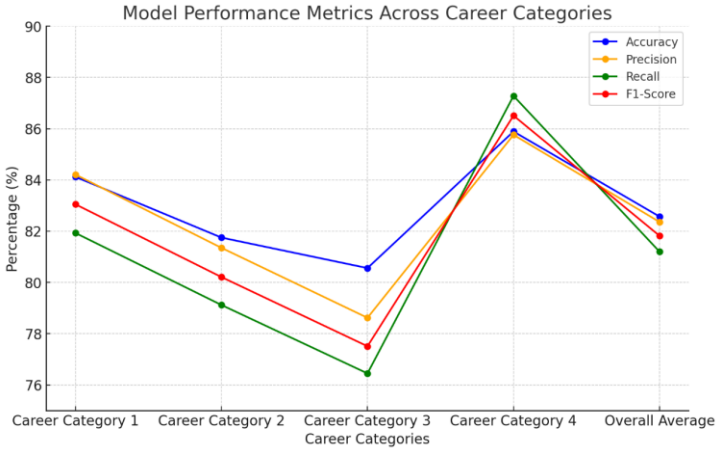

Figure 2. Performance metrics analysis.

Looking at the individual categories, Career Category 4 achieved the highest accuracy at 85.89%, suggesting that the model found this category to be the most distinguishable. This could be due to more distinct patterns in the data related to physical or behavioral traits for this career path. Similarly, Career Category 1 also showed strong performance, with an accuracy of 84.12%, indicating the model's ability to capture the characteristics of students in this category effectively. In contrast, Career Categories 2 and 3 had slightly lower accuracies, at 81.75% and 80.56%, respectively, suggesting that these categories may share more overlapping features, making it more challenging for the model to differentiate between them.

In terms of precision, which measures the proportion of true positive predictions among all optimistic predictions, the model performed well, with an overall value of 82.36%. Precision was highest for Career Category 4 (85.76%), again highlighting the model's effectiveness in identifying this category. For Career Category 1, precision was similarly high at 84.21%, while Career Categories 2 and 3 had lower precision values at 81.35% and 78.62%, respectively, indicating a greater likelihood of false positives in these categories.

The recall, or the ability of the model to correctly identify true positive instances, showed an overall value of 81.20%. Career Category 4 had the highest recall at 87.28%, suggesting that the model was particularly adept at capturing the true instances of this category. The recall for Career Category 1 was 81.93%, slightly lower than its precision, indicating that while the model was good at predicting this category, it may have missed some true instances. Career Categories 2 and 3 had lower recall values at 79.12% and 76.45%, reflecting some difficulty in correctly identifying all true instances of these career paths.

The F1-score, which balances precision and recall, had an overall value of 81.82%, showing a solid balance between the two metrics across all categories. The F1-score was highest for Career Category 4 (86.51%), further reinforcing the model's strong performance in predicting this career path. The F1 scores for Career Categories 1, 2, and 3 were 83.05%, 80.21%, and 77.51%, respectively, suggesting a slightly lower but still respectable ability to predict these categories effectively.

The confusion matrix (**Figure 3**) provides further insight into the model's performance by showing each career category's correct and incorrect predictions.

Career Category 1 had the highest number of correct predictions, with 210 out of 247 actual instances predicted accurately. However, 17 instances of Career Category 1 were misclassified as Career Category 3 and 12 as Career Category 2, indicating some confusion between these categories, possibly due to similar behavioral or biomechanical patterns. Only 8 instances were incorrectly predicted as Career Category 4, suggesting that the model could more clearly distinguish Category 4 from Category 1.

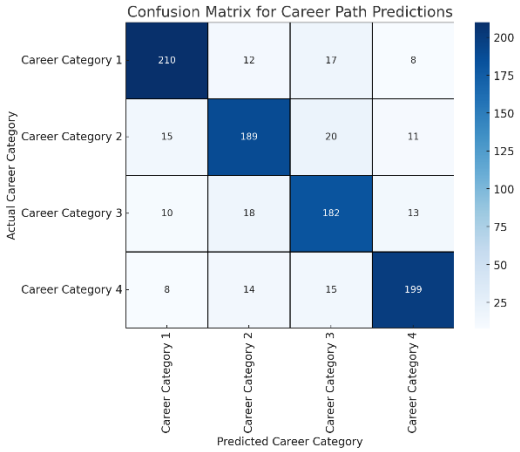

Figure 3. Confusion matrix for career path predictions.

For Career Category 2, the model correctly predicted 189 out of 235 instances, but 20 were misclassified as Career Category 3 and 15 as Career Category 1, showing that Career Category 2 shared overlapping features with the others. Similarly, Career Category 3 had 182 correct predictions out of 223. However, the largest source of confusion for this category came from misclassification as Career Category 2 (18 instances), suggesting that these two categories may be closely related regarding the features captured by the model.

Finally, Career Category 4 had 199 correct predictions out of 236, with the fewest misclassifications overall. Only 15 instances were incorrectly classified as Career Category 3, and fewer misclassifications were seen with Categories 1 and 2. This further supports the conclusion that Career Category 4 had more distinct characteristics that were easier for the model to recognize.

The ranking of feature importance (**Table 5** and **Figure 4**) from the Random Forest model highlights the critical role that both biomechanical and behavioral factors play in predicting career development paths. At the top of the list is Confidence Levels, with an importance score of 0.185, making it the most influential factor in career prediction. This suggests that students' self-reported confidence in career-related tasks such as public speaking, teamwork, and presentations has a strong predictive power, indicating that individuals with higher confidence levels are more likely to follow specific career paths.

Table 5. Ranking of feature importance.

Rank	Feature	Feature Category	Importance Score
1	Confidence Levels	Behavioral	0.185
2	Joint Angles (Hip)	Biomechanical	0.142
3	Engagement Levels	Behavioral	0.137
4	Ground Reaction Forces (GRF)	Biomechanical	0.126
5	Stride Length	Biomechanical	0.114
6	Openness (Personality Trait)	Behavioral	0.092
7	Gait Cadence	Biomechanical	0.081
8	Conscientiousness	Behavioral	0.073
9	Neuroticism	Behavioral	0.050
10	Fatigue Levels	Biomechanical	0.045

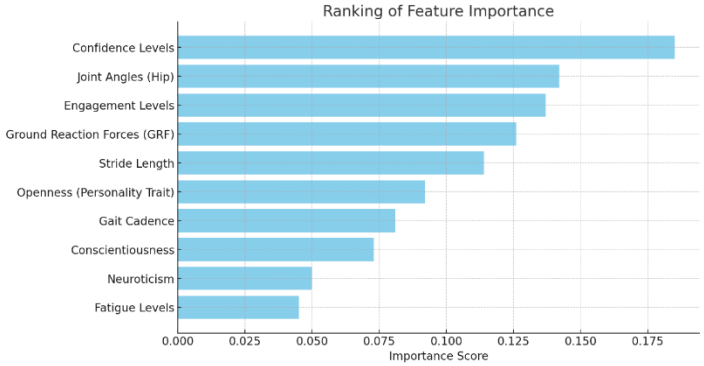

Figure 4. Feature importance.

Joint Angles (Hip) is closely followed, a biomechanical feature, with an importance score of 0.142. This indicates that physical posture and movement patterns significantly influence career path predictions, particularly at the hip. It suggests that certain physical attributes or movement efficiencies may correlate with specific career types, perhaps those requiring physical activity or performance-based tasks. Engagement Levels ranked third with an importance score of 0.137, underscoring the significance of active participation in group discussions and presentations. This feature reflects how involvement in social interactions and teamwork contributes to career development, suggesting that those who engage more actively in such activities are more likely to follow particular career paths.

GRF and Stride Length ranked fourth and fifth, with scores of 0.126 and 0.114, respectively. These biomechanical variables reflect how physical stability and gait patterns during tasks contribute to career path predictions, particularly in careers that require physical performance or precision. These features highlight the role of physical behavior in career development, particularly for physically demanding careers.

Behavioral features like Openness (0.092) and Conscientiousness (0.073) from the Big Five personality traits also rank high, suggesting that personality plays a significant role in shaping career paths. Students with higher openness and conscientiousness scores may be more inclined to pursue certain professions that align with creativity, responsibility, and innovation. Gait Cadence (0.081) and

Fatigue Levels (0.045) also contributed to career path predictions, with gait cadence reflecting the efficiency of movement and fatigue levels indicating the impact of endurance on career performance. Interestingly, Neuroticism (0.050) was the lowest-scoring personality trait, indicating that emotional stability has a less pronounced, but still notable, effect on career path predictions.

The cross-validation results from the 10 folds (**Table 6** and **Figure 5**) demonstrate the model's robustness and consistency across different subsets of data. Fold 8 had the highest performance across all metrics, with an accuracy of 84.37%, precision of 83.94%, recall of 82.64%, and an F1-score of 83.29%. This indicates that the model performed exceptionally well in this fold, correctly predicting a higher proportion of career paths and balancing precision and recall effectively. Fold 5 also showed strong performance, with an accuracy of 84.08% and a balanced F1-score of 83.05%, reflecting high predictive power for this subset. The consistently high precision values across most folds suggest that the model correctly predicted many true positives for each career category.

Table 6. Performance metrics across 10 folds.

Fold	Accuracy	Precision	Recall	F1-Score
Fold 1	83.14%	82.76%	81.82%	82.28%
Fold 2	81.76%	81.21%	80.17%	80.68%
Fold 3	83.32%	83.14%	81.96%	82.54%
Fold 4	82.51%	82.03%	81.11%	81.57%
Fold 5	84.08%	83.85%	82.27%	83.05%
Fold 6	82.02%	81.56%	80.74%	81.15%
Fold 7	81.67%	81.42%	80.36%	80.88%
Fold 8	84.37%	83.94%	82.64%	83.29%
Fold 9	82.26%	82.11%	81.24%	81.67%
Fold 10	82.65%	82.24%	81.44%	81.83%

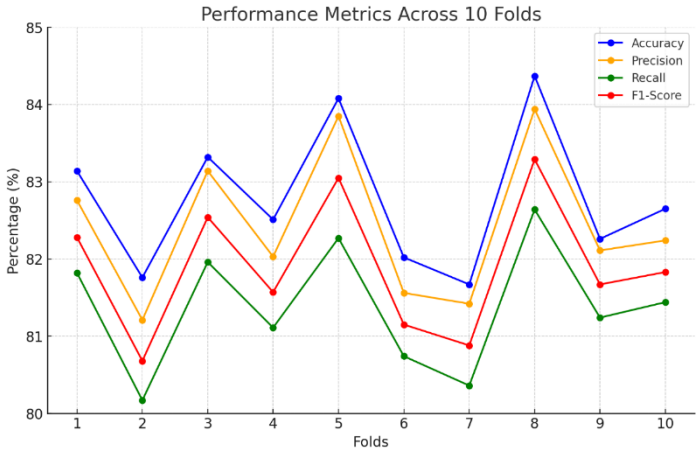

Figure 5. 10-fold validation.

The lowest performance was observed in Fold 7, with an accuracy of 81.67%, precision of 81.42%, and an F1-score of 80.88%, although the overall drop was minimal. The slight variance across folds (with accuracy ranging between 81.67%

and 84.37%) shows that the model generalizes well, maintaining high performance across different data splits, which is crucial for its reliability. Despite some fluctuation across folds, the model's performance remains stable, with low standard deviations. The average metrics across all folds confirm that the Random Forest model effectively predicts career paths based on biomechanical and behavioral data, maintains a balance between precision and recall and ensures consistent predictive accuracy. This robustness indicates that the model is well-suited for generalization to new data, making it a robust tool for career path predictions.

The Training vs Test Accuracy (**Figure 6**) provides valuable insight into how the Random Forest model's performance improved over the 100 epochs. The training accuracy shows a steady and consistent improvement, starting at around 0.5 and gradually rising to approximately 0.95 by the end of the 100 epochs. This indicates that the model was effectively learning and adjusting its parameters to fit the training data, significantly improving its predictive power.

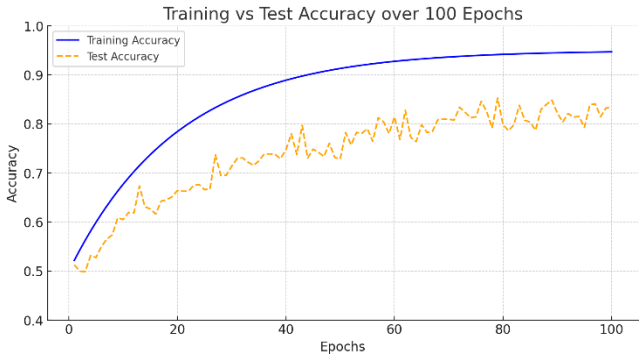

Figure 6. Training vs test accuracy.

The test accuracy curve follows a similar upward trajectory with slight fluctuations. The test accuracy starts around 0.5 and rises to approximately 0.85. The fluctuations observed in the test accuracy are expected, as the model is being evaluated on unseen data, which introduces more variability compared to the training set. The fact that test accuracy consistently improves alongside the training accuracy, without a significant drop-off or plateau, indicates that the model is generalizing well to new data and is not overfitting.

The Training vs Test Loss (**Figure 7**) reveals how the model's loss function evolves over the training process. The training loss begins high, around 1.5, and steadily decreases as the epochs progress, dropping below 0.2 by the final epoch. This decreasing trend reflects the model's ability to minimize the error in the training data as it fine-tunes its parameters over time. The continuous reduction in training loss indicates that the model is learning effectively and becoming more accurate in its predictions.

Similarly, the test loss decreases, though it starts slightly higher at 1.6 and shows more fluctuations than the training loss. By the 100th epoch, the test loss reaches around 0.4, but the fluctuations observed throughout the process highlight the challenges of generalizing to unseen data. These fluctuations suggest that while the model is learning effectively, the test data introduces more variability, possibly due to differences in the complexity or patterns within the test set. Notably, the test

loss does not rise significantly toward the end, which is a positive sign that the model is not overfitting and continues to generalize well to new data.

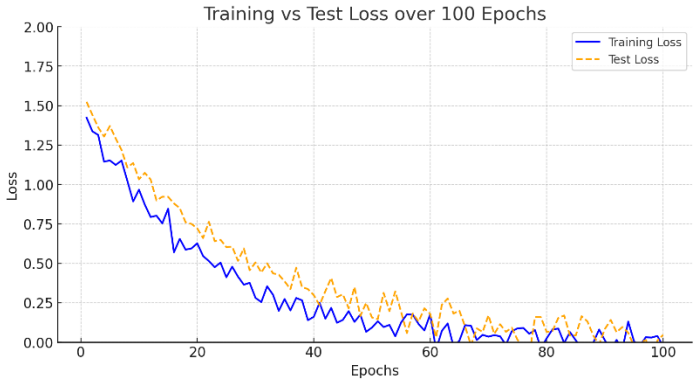

Figure 7. Training vs test loss.

5. Conclusion and future work

This study demonstrates the effectiveness of integrating biomechanical and behavioral data to predict career development paths using an RF-ML. The results highlight the importance of considering physical attributes and behavioral traits in career prediction, as each provides unique insights into a student's potential career trajectory. The model's performance, with an overall accuracy of 82.57% and vital metrics across precision, recall, and F1-score, confirms the feasibility of this approach in offering more personalized and data-driven career guidance. Key features such as confidence levels, joint angles, and engagement were the most significant predictors, underscoring the value of multidimensional data in understanding career outcomes. The study also illustrated how physical factors like gait and posture, often overlooked in career prediction, can influence career suitability, particularly in physically demanding professions. The implications of this research extend to career counseling and educational planning, where a deeper understanding of how behavioral and biomechanical factors intersect could lead to more targeted interventions and guidance.

Future research could refine this model by incorporating additional variables or expanding the dataset to include more diverse student populations. In conclusion, integrating biomechanical and behavioral data provides a more holistic approach to predicting career paths, with machine learning offering powerful tools for uncovering complex patterns in multidimensional datasets. This study offers a promising foundation for future innovations in career guidance systems, supporting students in making more informed and personalized career decisions.

Funding: This research was supported by the Philosophy and Social Science Research Project of Provincial Education Department in 2023 (unique project for employment and entrepreneurship of college graduates) "Bottleneck and Breakthrough: Research on the Construction of Employment and Entrepreneurship Guidance System of Higher Vocational Colleges Empowered by GROW Model" (Project Number: 23Z630).

Ethical approval: Not applicable.

Conflict of interest: The author declares no conflict of interest.

References

1. Han, J., Kelley, T., & Knowles, J. G. (2021). Factors influencing student STEM learning: Self-efficacy and outcome expectancy, 21st century skills, and career awareness. Journal for STEM Education Research, 4(2), 117-137.
2. Akkermans, J., Spurk, D., & Fouad, N. (2021). Careers and career development. In Oxford Research Encyclopedia of Psychology.
3. Su, R. (2020). The three faces of interests: An integrative review of interest research in vocational, organizational, and educational psychology. Journal of Vocational Behavior, 116, 103240.
4. Mõttus, R., Wood, D., Condon, D. M., Back, M. D., Baumert, A., Costantini, G., ... & Zimmermann, J. (2020). Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the Big Few traits. European Journal of Personality, 34(6), 1175-1201.
5. Rane, N., Choudhary, S. P., & Rane, J. (2024). Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions. Studies in Medical and Health Sciences, 1(2), 18-41.
6. Perumalsamy, J., Althati, C., & Shanmugam, L. (2022). Advanced AI and Machine Learning Techniques for Predictive Analytics in Annuity Products: Enhancing Risk Assessment and Pricing Accuracy. Journal of Artificial Intelligence Research, 2(2), 51-82.
7. Olivas-Padilla, B. E., Manitsaris, S., Menychtas, D., & Glushkova, A. (2021). Stochastic-biomechanic modeling and recognition of human movement primitives, in industry using wearables. Sensors, 21(7), 2497.
8. Arellano-González, J. C., Medellín-Castillo, H. I., Cervantes-Sánchez, J. J., & Vidal-Lesso, A. (2021). A practical review of the biomechanical parameters commonly used in the assessment of human gait. Revista mexicana de ingeniería biomédica, 42(3).
9. Stetter, B. J., & Stein, T. (2024). Machine Learning in Biomechanics: Enhancing Human Movement Analysis. In Artificial Intelligence in Sports, Movement, and Health (pp. 139-160). Cham: Springer Nature Switzerland.
10. Abbott, P., & Meerabeau, L. (2020). Professionals, professionalization, and the caring professions. In The sociology of the caring professions (pp. 1-19). Routledge.
11. Beer, P., & Mulder, R. H. (2020). The effects of technological developments on work and their implications for continuous vocational education and training: A systematic review. Frontiers in Psychology, 11, 918.
12. Armstrong, D. P., Beach, T. A., & Fischer, S. L. (2024). Quantifying how functional and structural personal factors influence biomechanical exposures in paramedic lifting tasks. Ergonomics, 67(7), 925-940.
13. Wu, S., Zhang, K., Zhou, S., & Chen, W. (2020). Personality and career decision-making self-efficacy of students from poor rural areas in China. Social Behavior and Personality: an international journal, 48(5), 1-18.
14. Natarajan, T., & Pichai, S. (2024). Behavior-driven development and metrics framework for enhanced agile practices in scrum teams. Information and Software Technology, 170, 107435.
15. Wu, C. H., de Jong, J. P., Raasch, C., & Poldervaart, S. (2020). Work process-related lead userness as an antecedent of innovative behavior and user innovation in organizations. Research Policy, 49(6), 103986.
16. Ogbuanya, T. C., & Salawu, I. A. (2024). Analysis of influence of personality traits on career satisfaction and job performance of electrical/electronic technology education lecturers in Nigeria. International Journal of Lifelong Education, 43(2-3), 224-245.
17. Yue, H., Cui, J., Zhao, X., Liu, Y., Zhang, H., & Wang, M. (2024). Study on the sports biomechanics prediction, sport biofluids and assessment of college students' mental health status transport based on artificial neural network and expert system. Molecular & Cellular Biomechanics, 21(1), 256-256.
18. Gormley, J. (2022). The Application of Hierarchical Clustering, PCA, SVM, and Random Forest Machine Learning Methods for Geological Identification of Martian LIBS Data. South Dakota School of Mines and Technology.
19. Rudar, J. (2024). Applying Multivariate Decision Trees to Visualize, Select Features, and Gain Insights into Biodiversity Genomics Datasets (Doctoral dissertation, University of Guelph).
20. Aceña, V., de Diego, I. M., Fernández, R. R., & Moguerza, J. M. (2022). Minimally overfitted learners: a general framework for ensemble learning. Knowledge-Based Systems, 254, 109669.

21. Natras, R., Soja, B., & Schmidt, M. (2022). Ensemble machine learning of random forest, AdaBoost and XGBoost for vertical total electron content forecasting. Remote Sensing, 14(15), 3547
22. Indumathi N et al., Impact of Fireworks Industry Safety Measures and Prevention Management System on Human Error Mitigation Using a Machine Learning Approach, Sensors, 2023, 23 (9), 4365; DOI:10.3390/s23094365.
23. Parkavi K et al., Effective Scheduling of Multi-Load Automated Guided Vehicle in Spinning Mill: A Case Study, IEEE Access, 2023, DOI:10.1109/ACCESS.2023.3236843.
24. Ran Q et al., English language teaching based on big data analytics in augmentative and alternative communication system, Springer-International Journal of Speech Technology, 2022, DOI:10.1007/s10772-022-09960-1.
25. Ngangbam PS et al., Investigation on characteristics of Monte Carlo model of single electron transistor using Orthodox Theory, Elsevier, Sustainable Energy Technologies and Assessments, Vol. 48, 2021, 101601, DOI:10.1016/j.seta.2021.101601.
26. Huidan Huang et al., Emotional intelligence for board capital on technological innovation performance of high-tech enterprises, Elsevier, Aggression and Violent Behavior, 2021, 101633, DOI:10.1016/j.avb.2021.101633.
27. Sudhakar S, et al., Cost-effective and efficient 3D human model creation and re-identification application for human digital twins, Multimedia Tools and Applications, 2021. DOI:10.1007/s11042-021-10842-y.
28. Prabhakaran N et al., Novel Collision Detection and Avoidance System for Mid-vehicle Using Offset-Based Curvilinear Motion. Wireless Personal Communication, 2021. DOI:10.1007/s11277-021-08333-2.
29. Balajee A et al., Modeling and multi-class classification of vibroarthographic signals via time domain curvilinear divergence random forest, J Ambient Intell Human Comput, 2021, DOI:10.1007/s12652-020-02869-0.
30. Omnia SN et al., An educational tool for enhanced mobile e-Learning for technical higher education using mobile devices for augmented reality, Microprocessors and Microsystems, 83, 2021, 104030, DOI:10.1016/j.micpro.2021.104030 .
31. Firas TA et al., Strategizing Low-Carbon Urban Planning through Environmental Impact Assessment by Artificial Intelligence-Driven Carbon Foot Print Forecasting, Journal of Machine and Computing, 4(4), 2024, doi: 10.53759/7669/jmc202404105.
32. Shaymaa HN, et al., Genetic Algorithms for Optimized Selection of Biodegradable Polymers in Sustainable Manufacturing Processes, Journal of Machine and Computing, 4(3), 563-574, https://doi.org/10.53759/7669/jmc202404054.
33. Hayder MAG et al., An open-source MP + CNN + BiLSTM model-based hybrid model for recognizing sign language on smartphones. Int J Syst Assur Eng Manag (2024). https://doi.org/10.1007/s13198-024-02376-x
34. Bhavana Raj K et al., Equipment Planning for an Automated Production Line Using a Cloud System, Innovations in Computer Science and Engineering. ICICSE 2022. Lecture Notes in Networks and Systems, 565, 707–717, Springer, Singapore. DOI:10.1007/978-981-19-7455-7_57.