Received June 5, 2025, July 30, 2025, date of publication September 1, 2025.

*Digital Object Identifier 10.21608/ijaici.2025.414064.1015*

# Real-Time Framework for Talent Swimmer Detection

**Hossam Fakher[1], Elsayed Badr[2,3], Ahmed Abdelfatah[4], and Ahmed Sara Sweidan[5,6].**

[1] Department of Artificial Intelligence, Faculty of Computers and Artificial Intelligence, Benha University, Egypt

[2] Department of Information Systems, College of Information Technology, Misr University for Science & Technology, Giza, Egypt

[3] Department of Scientific Computing, Faculty of Computer and Artificial Intelligence, Benha University, Benha, Egypt

[4] Department of Kinesiology, Specifications Biomechanics, Faculty of Sports Science, Damietta University, Damietta, Egypt

[5] Department of Artificial Intelligence, Faculty of Computer and Artificial Intelligence, Benha University, Egypt

[6] Faculty of Computer Science and Engineering, New Mansoura University, New Mansoura, Egypt

**ABSTRACT** This study presents a real-time framework for swimmer talent identification that integrates state-of-the-art pose estimation and machine learning classification techniques. To address the limitations of traditional pose estimation methods in aquatic environments, RTMPose is employed to extract reliable 2D joint keypoints. Temporal consistency across sequences is achieved using the RIFE interpolation model, selected for its efficiency in standardizing frame counts while avoiding the computational overhead of temporal deep learning models such as LSTMs or 3D CNNs. The dataset, consisting of underwater breaststroke footage, was augmented and balanced using SMOTE, with sensitivity analysis highlighting both its benefits for minority classes and the risk of overfitting. A comprehensive evaluation of twelve classifiers demonstrated that ensemble methods, particularly LightGBM, achieved superior results, yielding a cross-validation F1 score of 93.6% and a test F1 score of 96.8%. While the framework shows strong promise for practical use in sports analytics, its current evaluation is limited to breaststroke and underwater footage. Future work will expand to multiple swimming styles, above-water perspectives, and diverse pool environments to ensure broader generalization.

**INDEX TERMS** Object Detection, Pose estimation, RTMPose, Swimmer, Talent and Machine Learning.

## I. INTRODUCTION

Swimming is a globally recognized sport and a central discipline in international competitions. Talent identification and performance evaluation in swimming are critical for both athlete development and competitive success. Unlike land-based sports, swimming presents unique challenges for performance analysis, as the aquatic environment complicates visual observation. Coaches and analysts frequently struggle to perceive detailed biomechanical patterns in real time, while post-event evaluations rely heavily on replaying video recordings and manually tracking athletes' motions. Such methods are not only labor-intensive but also limited in precision, delaying feedback and hindering systematic talent development [1].

Over the past decade, advances in digital imaging and computer vision have transformed performance monitoring in aquatic sports. Automated algorithms can now interpret swimmer behavior with higher accuracy, minimizing dependence on manual observation. A key technology driving this progress is pose estimation, which provides a structured representation of human motion through skeletal keypoints [2, 3]. In swimming, pose estimation supports a wide range of applications, including athlete tracking, stroke recognition, biomechanical correction, and even drowning detection and rescue systems. Importantly, underwater imaging enhances the quality of pose estimation by reducing surface distortions such as splashing and reflections, offering clearer visualization of torso and limb movements [4, 5].

Despite these advances, current methods still face substantial challenges. Early attempts relied on graph-based models such as Deformable Part Models (DPMs) to capture swimmer poses. These approaches used handcrafted descriptors like Histogram of Oriented Gradients (HOGs), which were computationally expensive and rigid, often failing to adapt to large variations in human posture [6, 7]. With the rise of deep learning, regression-based models attempted direct coordinate prediction but often suffered from spatial generalization issues and overfitting in dynamic aquatic environments. In contrast, heatmap-based approaches, which encode joint probabilities as pixel intensities, have demonstrated superior accuracy and robustness. Yet, their systematic application in swimming,

especially for talent identification rather than stroke recognition, remains underexplored [8-10].

Another gap lies in temporal modeling of swimmer motion. Swimming strokes are cyclic and require consistent temporal representation for meaningful analysis. Traditional temporal models, such as Long Short-Term Memory (LSTM) networks and 3D Convolutional Neural Networks (3D CNNs), have been applied in other sports domains but are often unsuitable for aquatic analytics. They demand large, annotated datasets and impose high computational costs, limiting their feasibility for real-time deployment. This motivates the use of efficient interpolation methods such as Real-Time Intermediate Flow Estimation (RIFE) [11], which can normalize video length across samples while preserving temporal consistency at a fraction of the computational burden.

Furthermore, swimmer talent datasets often suffer from class imbalance. Highly skilled athletes (elite-level or "Talent A") are much rarer than average or non-talented swimmers, creating skewed distributions. This imbalance biases classifiers toward majority classes and risks overlooking minority talent groups. To address this, resampling strategies such as Synthetic Minority Oversampling Technique (SMOTE) have been introduced in machine learning. While SMOTE enhances representation of minority classes, it also raises the risk of overfitting, particularly when the original sample size is very small. Thus, careful sensitivity analysis is necessary to validate its impact in talent detection tasks.

Taken together, these limitations underscore the need for an integrated framework that combines robust pose estimation, efficient temporal normalization, and strategies for handling imbalanced data to support swimmer talent identification. Existing studies have primarily focused on biomechanical stroke analysis or performance monitoring; few have tackled the problem of systematic talent classification using pose-based machine learning.

This study addresses these challenges by introducing a real-time framework for swimmer talent detection. The framework integrates RTMPose for accurate underwater keypoint extraction, RIFE for temporal interpolation, and a comparative analysis of twelve machine learning classifiers. Special attention is given to ensemble methods such as LightGBM, which achieve superior accuracy and efficiency, as well as the role of SMOTE in mitigating class imbalance.

The main contributions of this study are as follows:

- Develop an end-to-end framework that integrates RTMPose-based pose estimation with machine learning classifiers for swimmer talent detection.
- Employ RIFE for temporal normalization, ensuring consistent video sequence length while avoiding the computational burden of LSTMs and 3D CNNs.
- Conduct a comparative evaluation of twelve classifiers, demonstrating the superior performance of ensemble-based models, particularly LightGBM.
- Apply SMOTE for class balancing and include a sensitivity analysis to examine its effects on minority classes, highlighting both benefits and potential risks.
- Construct and analyze a novel underwater breaststroke dataset, laying the groundwork for future research on multi-stroke and multi-environment swimmer analytics.

The remainder of this paper is structured as follows: Section 2 reviews related work on pose estimation and swimming analytics; Section 3 details the proposed framework, including preprocessing, temporal normalization, and classification; Section 4 presents the experimental results; and Section 5 concludes with limitations and directions for future research.

## II. Literature Review

Human pose estimation and action recognition in computer vision play a critical role in sports science, physical assessment, and talent identification. With the integration of deep learning techniques—such as Graph Convolutional Networks (GCNs) combined with spatiotemporal architecture, pose estimation technologies have become increasingly accurate, robust, and practical. As a result, biomechanical insights can be conveyed in real time, personalized training programs can be developed, and individual performance can be quantitatively assessed. Consequently, athletes, patients, and workers are increasingly encountering these technologies in sports coaching, rehabilitation, and safety evaluation.

This section reviews recent methodological advances, practical applications, and persisting challenges in pose estimation, with a particular focus on its role in athlete identification, swimming biomechanics, and sports performance enhancement.

### *A. Human pose estimation*

Consequently, the use of pose estimation technologies in athletic coaching, rehabilitation, and ergonomic hazard assessment has continued to grow. This section provides an overview of recent methodological developments, practical applications, and open challenges in the field, with particular emphasis on athlete identification, swimming biomechanics, and sports performance analysis [12, 13]. Deep learning has significantly advanced pose estimation by enabling models that rely on monocular image inputs, making the technology more accessible and scalable. For example, Convolutional Pose Machines (CPMs), introduced in [14], addressed the vanishing gradient problem in deep neural networks through the use of intermediate supervision layers. Similarly, the Stacked Hourglass Network proposed in [15] improved posture prediction accuracy by capturing joint properties at multiple spatial resolutions.

Pose estimation has also proven useful in evaluating athletic activity and improving performance outcomes. In [6], a hybrid approach was introduced that combined object detection and human posture estimation on sports-specific datasets. The method employed Gaussian Mixture Models (GMMs) together with the YOLO framework to identify human–object interactions (HOI). Skeleton-based models and GMM-driven elliptical fitting algorithms were used for body posture representation. The findings demonstrated the value of pose estimation in advancing sports analytics, particularly for athlete assessment and performance enhancement.

### *B. Recognition of actions*

The precise detection of human posture and motion is essential for accurate action recognition, particularly in sports scenarios. Deep learning architecture has often enabled this progress. For example, [16] introduced a Spatial-Temporal Long Short-Term Memory (ST-LSTM) model to capture both spatial relationships among skeletal joints and temporal dynamics across video frames. Similarly, [17] developed an enhanced LSTM architecture with a spatiotemporal attention mechanism, which allowed the network to focus on the most informative joints and frames during action recognition tasks.

In [18] , a two-stage human posture estimation system was proposed to improve action recognition. The method combined Convolutional Pose Machines (CPM) with an improved Single Shot Detector (SSD) to generate skeletal heatmaps. This approach proved effective in recognizing command actions in human–robot interaction scenarios, demonstrating the real-world utility of skeleton-based recognition. The integration of heatmap representations with convolutional neural networks improved accuracy and robustness, especially in demanding and dynamic environments.

Pose estimation has also become a fundamental method for sports performance analytics, supporting the quantitative assessment of athletic techniques. Beyond competitive sports, it has applications in personal health and fitness. For instance, in [19] , researchers compared four deep learning models—MediaPipe, PoseNet, OpenPose, and EpipolarPose—for yoga pose recognition. Using data from the S-VYASA dataset, which included five common yoga poses (e.g., warrior pose, and tree pose), the study showed that MediaPipe achieved the highest accuracy, up to 90.9%. This performance was attributed to its optimized two-stage detector–tracker architecture. These findings highlight the potential of pose estimation systems in supporting yoga practitioners by providing real-time feedback to ensure proper alignment and safety.

Pose estimation in swimming presents additional challenges caused by water refraction and frequent occlusions. In [20] , researchers addressed these difficulties by introducing a method adapted to different swimming styles, which improved joint localization accuracy by 16%. Their approach focused on continuous position estimation using swimming channel footage that included both above- and underwater views. They enhanced CPM models by incorporating contextual information such as swimming style (backstroke, freestyle, etc.) and temporal correlations across frames. The integration of activity-specific and time-dependent features yielded significant improvements, with the Percentage of Correct Keypoints (PCK) increasing by up to 16% compared to the baseline. This demonstrates the importance of domain-specific knowledge for overcoming visual complexities in aquatic environments.

Further contributions to the swimming domain were made by [21] who proposed a method for identifying key postures in cyclic swimming motion. Their framework used a pictorial structure model supplemented with “poselets” derived from Histogram of Oriented Gradients (HoG) features. This configuration enabled the detection of distinctive postures within repetitive stroke cycles. To estimate the most relevant key poses—critical for parameters such as stroke frequency—the authors applied a maximum likelihood framework that leveraged temporal consistency.

More recently, [22] employed the High-Resolution Network (HRNet) to predict swimmer poses from underwater video surveillance. Unlike conventional graph-based methods, HRNet preserved spatial detail by maintaining high-resolution feature maps throughout the pipeline. Using the HRNet-W48 variation, their model achieved an Average Precision (AP) of 95.6%. This study highlights the effectiveness of advanced deep learning frameworks in diverse aquatic contexts and emphasizes the advantages of underwater monitoring for full-body motion analysis.

Building on these findings, the current work integrates RTMPose with ensemble-based machine learning techniques to develop a real-time system for swimmer talent identification. Several key advancements influenced this framework: [20] ’s integration of swimming style and temporal dynamics, [21]’s cyclic motion modeling strategy for identifying critical postures, [22] ’s application of high-resolution deep learning models for underwater monitoring, and [19] ’s demonstration of pose estimation in fitness and wellness contexts. By addressing aquatic challenges such as occlusion, light refraction, and dynamic movement, our work advances human posture estimation in swimming. This study contributes to the growing use of pose estimation in athlete performance evaluation and talent discovery by developing accurate, real-time monitoring techniques.

## III. Methodology

Figure 1 shows the general structure of the suggested method, which includes a comprehensive machine learning pipeline intended for motion analysis-based swimming talent classification. In order to increase temporal granularity, the procedure starts with raw video data that has been temporally normalized using the RIFE method to interpolate frames and modify each sequence to 70 frames. The dataset is then supplemented in order to increase variability and enhance model generalization. Important joint coordinates in two dimensions (X and Y axes) that form the basis of the feature set are then extracted using the RTMPose-based architecture for pose estimation. Several preprocessing methods are applied to these features, such as class balancing, dataset partitioning, and normalization. Cross-validation is used to ensure model resilience and prevent overfitting. A series of machine learning classifiers are then trained using the revised feature data, classifying swimmers into four groups: Not Talent, Talent A, Talent B, and Talent C. This comprehensive solution combines state-of-the-art methods in machine learning and computer vision to provide accurate and scalable talent assessment in competitive swimming.

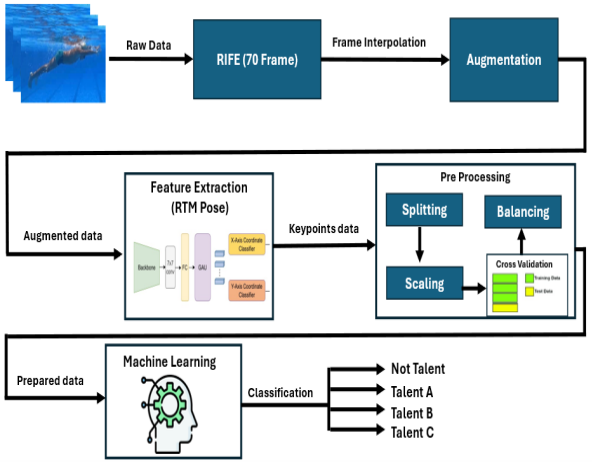


**Figure 1 Overall architecture of the proposed Talent Discovering Framework.**

### *A. Dataset Collection*

A dedicated dataset was constructed by capturing high-definition video recordings of swimmers using a Full HD camera set to a resolution of 1280×720 pixels. This resolution enabled the acquisition of detailed visual data, including fine-grained biomechanical movements and joint angles, serving as a reliable input source for model development. The dataset comprises approximately four hours of continuous footage, specifically targeting breaststroke techniques. To ensure comprehensive representation, swimmers across a broad range of proficiency levels—from novices to high-performance athletes—were included. Classification of swimmer expertise was carried out by experienced coaches, ensuring both accuracy and consistency in labeling based on their execution of the breaststroke stroke.

While this dataset provides a reliable foundation for model development, it is limited to breaststroke swimmers and underwater footage. This restriction narrows the generalizability of the findings, as the current framework has not yet been evaluated across other swimming strokes or environmental conditions. Future work will therefore expand the dataset to include multiple swimming styles (e.g., freestyle, butterfly, and backstroke), above-water perspectives, and recordings from different pool environments. Such extensions will allow for more comprehensive validation and ensure broader applicability of the proposed framework.

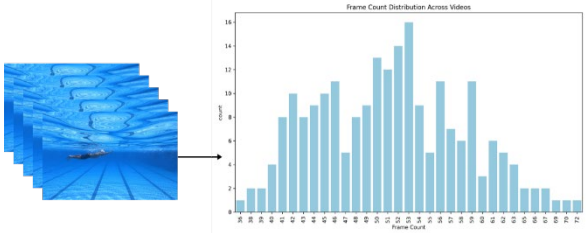


**Figure 2 Distribution of frame counts across the collected video samples.**

Achieving uniformity in the number of frames across video samples is crucial for consistent model training. In this study, an inconsistency in frame counts was identified among the collected video clips (see Figure 2), prompting the need for temporal standardization. The maximum observed frame count of 70 frames was used as the reference length to completely capture the swimming motion within each sample. To achieve this, we normalized all data using RIFE, an interpolation method based on neural networks [11].

RIFE estimates motion flows, including optical flow, between adjacent frames by warping and blending the inputs to produce new intermediate frames. The Intermediate Flow Network (IFNet) refines motion predictions using a hierarchical, coarse-to-fine strategy and generates smooth, high-quality interpolations with a fusion mask. This enables the upsampling of all video sequences to a consistent 70-frame format, ensuring temporal synchronization across the dataset and improving the effectiveness of subsequent machine learning applications [11].

Compared with alternative temporal models such as Long Short-Term Memory (LSTM) networks or 3D Convolutional Neural Networks (3D CNNs), RIFE offers several advantages. LSTMs and 3D CNNs require large amounts of labeled data to capture motion dynamics, are computationally intensive, and introduce significant latency, which makes them impractical for real-time swimmer analytics. By contrast, RIFE achieves temporal consistency through interpolation rather than recurrent training or spatiotemporal convolution, making it both data-efficient and computationally lightweight. This reduces the risk of overfitting on small, imbalanced

datasets while still producing temporally coherent sequences [11].

Additionally, the use of a privileged distillation framework in RIFE's training phase improves interpolation accuracy by allowing a teacher network to provide advanced supervisory signals to a student network. As a result, RIFE can synthesize intermediate motion frames with higher fidelity, accelerating convergence while maintaining efficiency during inference. Its lightweight design and real-time performance make it ideal not only for video-based sports analytics but also for use cases such as live video feeds, slow-motion rendering, and high frame rate display technologies. For these reasons, RIFE was selected as the temporal normalization method in this framework, ensuring both practical deploy ability and high-quality temporal alignment in swimmer talent detection systems [11].

### B. Feature extraction

Three fundamental techniques were investigated in order to create a reliable model for identifying swimming talent: pose estimation-based analysis, swimmer-specific tracking, and holistic video classification. The first strategy allows the model to assess the swimmer's movements in addition to contextual elements by analyzing the entire video without selecting particular segments. Although this technique records a great deal of visual information, it may be impacted by unrelated background activity, such as other swimmers or environmental objects. Additionally, analyzing entire video frames—especially those with high resolution—can require a significant amount of processing power. The small and unevenly distributed dataset exacerbates these issues and restricts the model's applicability.

The second strategy locates and tracks the swimmer using object detection algorithms; YOLOv8 is employed because it has a higher accuracy and inference speed than earlier versions like YOLOv3. This method improves the model's capacity to concentrate on the swimmer's technique by limiting the study to the region of interest (ROI). However, it significantly depends on the tracking algorithm's accuracy and adds more computing load during preprocessing. Errors in object tracking can lead to incorrect motion interpretation, which impairs classification performance. This problem is especially important when there is a shortage of training data.

This work uses RTMPose, a real-time, top-down deep learning architecture designed for effective multi-person keypoint recognition, for posture estimation. To start, a pre-trained object detector, like YOLOv3 or RTMDet, is used to identify the people in a frame. However, the extraction of relevant features for keypoint prediction is carried out by a specialized backbone network known as CSPNeXt. Once individuals are detected, their corresponding regions are cropped and passed through the CSPNeXt module, which captures detailed spatial features from the input image. These features are subsequently forwarded to SimCC—a compact coordinate classification component—which reformulates the 2D keypoint estimation task as a classification problem, diverging from traditional regression-based or heatmap-based approaches. The entire feature extraction and pose inference workflow is managed internally by RTMPose's integrated backbone and supporting modules. [23] (see Figure 3)

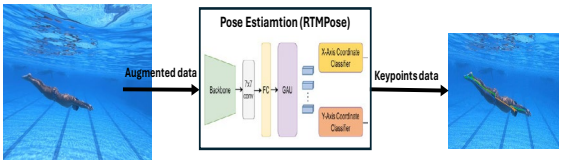


**Figure 3 Illustration of the keypoint extraction process using RTMPose.**

### C. Pre Processing

Each video was meticulously segmented to encompass the full execution of the breaststroke, capturing the stroke cycle from initiation to completion. This segmentation facilitated a granular examination of the swimmer's technique throughout the motion. Following this, the segmented clips were classified into four distinct categories based on performance level: (see Figure 4)

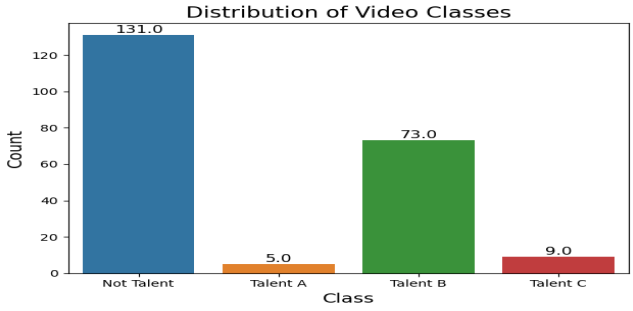


**Figure 4 Number of videos in each swimmer talent class (Talent A, Talent B, and Talent C) within the breaststroke dataset prior to augmentation.**

The classification into distinct categories was based on the swimmers' proficiency and effectiveness in performing the breaststroke technique. To ensure consistency and reliability, experienced coaches conducted the labeling process.

Given the limited size of the original dataset, video augmentation strategies were implemented to enhance variability and expand the data volume, effectively tripling the dataset. Each source video was augmented into eight distinct versions (see Figure 5) using a combination of transformations, including rotation, brightness modulation, and horizontal flipping. Rotational adjustments within a range of –15° to 15° introduced perspective variation, allowing the model to better accommodate different viewing angles. Alterations in brightness simulated a range of lighting scenarios, improving the model's ability to function under diverse illumination conditions. Flipping was applied to diversify the orientation of motion, promoting the learning of direction-invariant features and

thereby enhancing overall model generalization and performance. (see Figure 6)

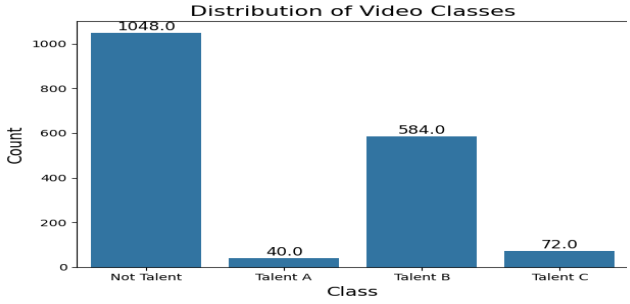


**Figure 5 Distribution of videos across swimmer talent classes (Talent A, B, and C) in the breaststroke dataset after augmentation. This figure highlights the initial class imbalance and the role of augmentation in improving class-level representation prior to training**

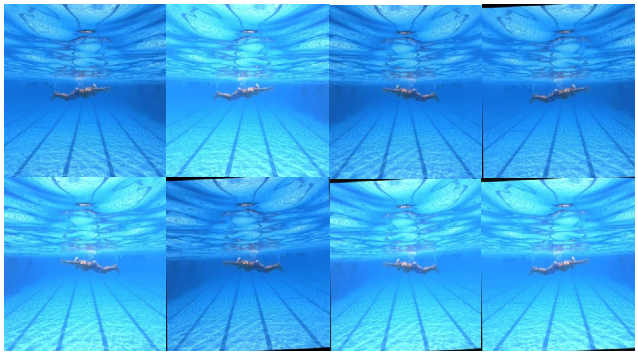

**Figure 6 Example frames from the augmented dataset, demonstrating variations in brightness, rotation, and scaling. These augmentations increase data diversity and improve the robustness of the classifiers against visual distortions.**

## Splitting

Following the augmentation process, the dataset was partitioned into training and testing subsets, with 80% allocated for model training and the remaining 20% reserved for evaluation. Consequently, 349 testing samples and 1,395 training samples were produced.

## Scaling

Standard scaling was applied to normalize the dataset. This technique transforms feature values by subtracting the mean and dividing by the standard deviation, thereby producing a distribution with zero mean and unit variance.

$$x_{scaled} = \frac{x - \mu}{\sigma} \quad (1)$$

## Cross Validation

By splitting the dataset into five separate and roughly equal groups, this method guarantees a consistent distribution of class labels across all folds; I use five folds.

## Balanced SMOTE

The dataset exhibited a clear imbalance across swimmer categories, with underrepresentation in high-talent groups such as Talent A. To mitigate this issue, we employed the Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic samples of minority classes by interpolating between existing samples. This method reduces bias toward majority classes, improves class-level representation, and enhances overall model accuracy compared to naïve oversampling.

performance measure

A range of performance criteria were used to assess the machine learning models' ability to detect swimming talent after the training phase. The evaluation framework's integration of accuracy, precision, recall, and F1-score allowed for a thorough analysis of overall performance.

## Evaluation Metrics

Accuracy: By figuring out the proportion of accurately classified events relative to all predictions. [24]

$$_{A}ccuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Precision: Precision quantifies the accuracy of the model's positive predictions by dividing the number of true positives by the total number of expected positives.[24]

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

Recall: Recall, also known as sensitivity, gauges how well the model can detect every real positive instance. [24]

$$recall = \frac{TP}{TP + FN} \quad (4)$$

F1-Score: The F1-score, which represents the harmonic means of precision and recall, offers a comprehensive evaluation of a model's performance, particularly when working with unbalanced datasets. [24]

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

To assess machine learning models for identifying swimming talent, a number of performance metrics, including accuracy, precision, recall, and F1-score.

## Classification models

To create a useful framework for identifying swimming talent, this study assessed a range of machine learning algorithms, including both traditional (SVM, KNN, Logistic Regression) and ensemble-based (CatBoost, LightGBM, XGBoost). Joint coordinate sequences acquired through pose estimation were used to create the structured data that the models were trained on. Because ensemble models can handle high-dimensional, noise, and incomplete data, they performed well, especially CatBoost and LightGBM. Interpretable baselines were provided by conventional classifiers. Notably, ensemble approaches also provided biomechanical insights into swimmer performance by highlighting important joint movements that affect classification.

## Model Hyperparameter Settings

Table 1 below lists all the machine learning models used in this study, along with the corresponding hyperparameters that were set up for each model during training and evaluation:

**Table 1 Parameters of Machine Learning Models**

| Models | Parameters | | | | |
|---|---|---|---|---|---|
| CatBoost | iterations=500 | | | | random_state=0 |
| Hist Gradient Boosting | loss='log_loss' | max_iter=100 | learning_rate=0.1 | | |
| Extra Trees | n_estimators=100 | criterion='gini' | max_features='sqrt' | | |
| LightGBM | boosting_type='gbdt' | num_leaves=31 | learning_rate=0.1 | n_estimators=100 | |
| Random Forest | n_estimators=100 | criterion='gini' | max_features='sqrt' | | |
| Bagging | n_estimators=10 | | | | |
| AdaBoosting | n_estimators=50 | | learning_rate=1.0 | | |
| XGBoost | use_label_encoder=True | | eval_metric='logloss' | | |
| Logistic Regression | penalty='l2' | C=1.0 | solver='lbfgs' | max_iter=100 | |
| K Nearest Neighbors | n_neighbors=5 | weights= 'uniform' | algorithm='auto' | | |
| Support Vector Machine | C=1.0 | kernel='rbf' | gamma='scale' | | |
| Decision Tree | criterion='gini' | | splitter='best' | | |

## IV. Results and discussions

The performance of machine learning models was evaluated on the dataset. The results were compared between training and testing datasets using metrics including Cros validation F1 score, Accuracy, Precision, Recall, and F1 Score. The results are detailed in Table 2

**Table 2 Evaluation of Machine Learning Models on Training and Testing Sets Using Benchmark Metrics**

| Models | Cross Validation F1 Score | Training | | | | | Testing | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | Precision | Recall | F1 Score | Time (S) | Accuracy | Precision | Recall | F1 Score | Prediction Time (S) |
| **LightGBM** | 93.6 | **100.0** | **100.0** | **100.0** | **100.0** | 15.071869 | **96.848** | **96.9** | **96.8** | **96.8** | 0.006209 |
| **Hist Gradient Boosting** | **94.6** | **100.0** | **100.0** | **100.0** | **100.0** | 29.568383 | 96.562 | 96.5 | 96.6 | 96.5 | 0.016641 |
| **Extra Trees** | 93.6 | **100.0** | **100.0** | **100.0** | **100.0** | 0.416143 | 96.275 | 96.4 | 96.3 | 96.2 | 0.031903 |
| **CatBoost** | 94.0 | **100.0** | **100.0** | **100.0** | **100.0** | 151.685011 | 95.702 | 95.8 | 95.7 | 95.7 | 0.008477 |
| **Random Forest** | 91.7 | **100.0** | **100.0** | **100.0** | **100.0** | 1.368065 | 95.129 | 95.2 | 95.1 | 95.1 | 0.032688 |
| **XGBoost** | 93.9 | **100.0** | **100.0** | **100.0** | **100.0** | 20.588125 | 95.129 | 95.4 | 95.1 | 95.1 | 0.005984 |
| **Bagging** | 87.5 | 99.3 | 99.3 | 99.3 | 99.3 | 4.155604 | 91.117 | 91.2 | 91.1 | 91.1 | 0.158459 |
| **Decision Tree** | 81.8 | **100.0** | **100.0** | **100.0** | **100.0** | 3.442314 | 87.966 | 88.4 | 88.0 | 88.0 | 0.001323 |
| **Logistic Regression** | 81.9 | 91.6 | 91.7 | 91.6 | 91.6 | 2.088439 | 82.521 | 82.6 | 82.5 | 82.5 | **0.001000** |
| **K Nearest Neighbors** | 74.9 | 90.5 | 91.6 | 90.5 | 90.7 | **0.098177** | 78.510 | 81.1 | 78.5 | 79.1 | 0.062718 |
| **Support Vector Machine** | 67.1 | 71.8 | 76.6 | 71.8 | 72.5 | 2.783184 | 71.920 | 77.4 | 71.9 | 72.5 | 0.470623 |
| **AdaBoosting** | 43.1 | 45.2 | 59.1 | 45.2 | 41.8 | 25.687481 | 43.553 | 55.8 | 43.6 | 39.8 | 0.041550 |

The comprehensive evaluation of twelve machine learning algorithms, with a particular emphasis on ensemble learning techniques, revealed clear differences in performance, generalization, and computational efficiency. As summarized in the results table, models such as LightGBM, Hist Gradient Boosting, Extra Trees, and CatBoost consistently outperformed others across nearly all key metrics, including cross-validation F1 score, precision, recall, and inference latency.

LightGBM emerged as a top performer. It achieved a cross-validation F1 score of 93.6%, perfect training accuracy, and a test F1 score of 96.8%. The model balanced accuracy and efficiency effectively, with a training time of just over 15 seconds and a prediction latency of 0.006 seconds. These characteristics make LightGBM particularly suitable for real-time systems that require both speed and reliability.

Hist Gradient Boosting, although marginally slower in training (~30 seconds) and prediction (0.017 seconds), achieved the highest cross-validation F1 score at 94.6%. This suggests superior generalization to unseen data. Its performance closely paralleled that of LightGBM, making it well-suited for applications where a small improvement in generalization justifies the additional computational cost.

Extra Trees offered a strong trade-off between accuracy and speed. Its test F1 score (96.2%) and cross-validation F1 score (93.6%) were comparable to the leading models. However, it excelled in efficiency, requiring only 0.42 seconds for training and 0.031 seconds for prediction. These attributes make Extra Trees particularly appealing for deployment in resource-constrained environments or for iterative model testing pipelines.

CatBoost also demonstrated strong predictive capability, with a cross-validation F1 score of 94.0% and a test F1 score of 95.7%. It combined excellent accuracy with very low prediction latency (0.008 seconds). Its primary limitation, however, was the lengthy training time (151.69 seconds), which may reduce its feasibility in scenarios requiring frequent retraining or real-time model updates.

By contrast, traditional models such as Logistic Regression and Decision Trees performed substantially worse. Their test F1 scores were 82.5% and 88.0%, respectively. Although interpretable and computationally lightweight, these models lacked the predictive strength needed for structured and complex datasets.

Overall, the evaluation highlights the superiority of ensemble-based methods, particularly LightGBM and Hist Gradient Boosting, for structured, keypoint-derived datasets such as those used in swimming talent detection. These models combine accuracy, scalability, and responsiveness, making them highly suitable for both academic exploration and practical deployment (see Figures 7, 8, and 9).

From a hardware perspective, the complete inference pipeline—including pose keypoint extraction with RTMPose—was executed on an Intel i7-9750H CPU. The keypoint estimation process for a 70-frame video sequence required approximately 14–16 seconds, consuming only 0.000271 kWh of energy. This corresponds to an estimated cost of 0.003 cents per sequence, underscoring the computational and economic efficiency of the proposed approach.

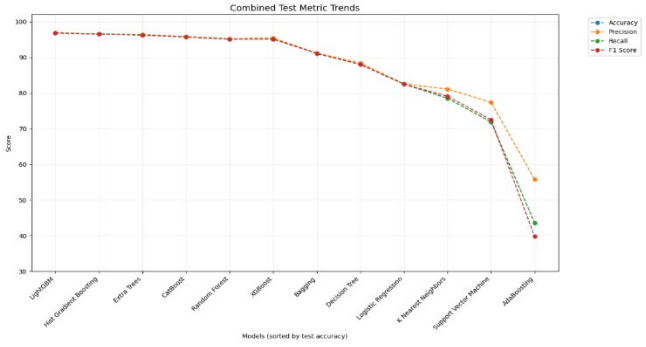


**Figure 7 Comparison of testing accuracy, precision, recall, and F1-score across twelve machine learning classifiers. The figure illustrates the relative performance of different models, highlighting the superior results achieved by ensemble-based methods such as LightGBM.**

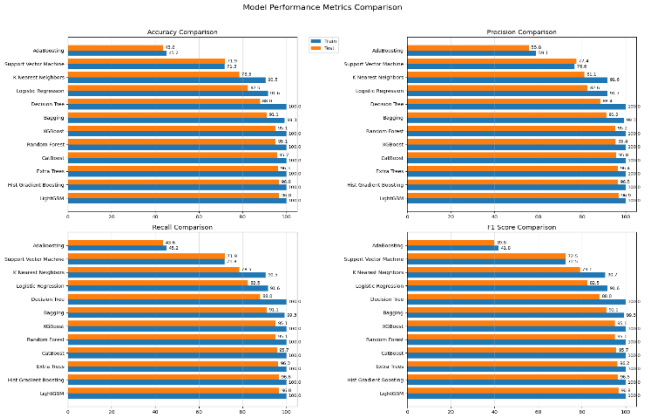


**Figure 8 Comparison of training and testing performance (accuracy, precision, recall, and F1-score) between the machine learning models. This comparison provides insights into model generalization and helps identify cases of potential overfitting or underfitting.**

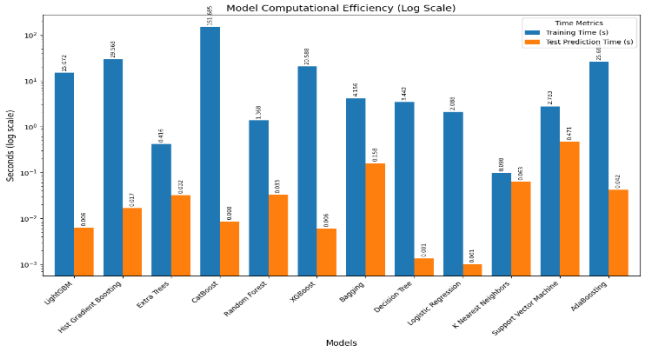


**Figure 9 training and testing time (in seconds) for each machine learning model. The figure emphasizes the trade-off between accuracy and computational efficiency, showing that LightGBM achieves both strong performance and reduced training time.**

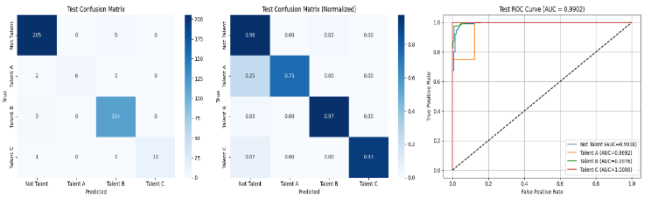


**Figure 10 Confusion matrix and ROC-AUC curve for the LightGBM model on the testing dataset. These visualizations confirm the model's ability to achieve high classification accuracy across all talent classes while maintaining robust discrimination between classes.**

Figure 10. Classification outcomes of the best-performing model (LightGBM). The confusion matrix (left) shows strong predictive performance across all swimmer talent categories, with minimal misclassifications. The normalized confusion matrix (center) highlights near-perfect recall for Not Talent (0.98), Talent B (0.97), and Talent C (0.93), while Talent A achieved lower recall (0.75) but perfect precision, reflecting its small sample size (n = 8). The ROC-AUC curves (right) further confirm discriminative power, with AUC values ranging from 0.994 to 1.000. Together, these results validate the model's robustness and suitability for pose estimation–based talent recognition.

**Table 3 Classification Performance Metrics by Talent Category**

| Class | Number | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| Not Talent | 210 | 98 % | 97 % | 98 % | 97 % |
| Talent A | 8 | 75 % | 100 % | 75 % | 86 % |
| Talent B | 117 | 97 % | 96 % | 97 % | 97 % |
| Talent C | 14 | 93 % | 100 % | 93 % | 96 % |

Table 3 presents the classification results for each swimmer category, including F1 score, recall, accuracy, and precision. The model achieved strong performance in most categories. For the Not Talent group, the F1 score was 97% with an accuracy of 98%, indicating highly reliable classification. Similarly, Talent B achieved 97% across all major evaluation metrics, confirming the model's stability in recognizing this category. The Talent C class also performed well, with an F1 score of 96%, perfect precision (100%), and recall of 93%. In contrast, Talent A achieved lower results, with an F1 score of 86% and recall and accuracy of 75%, despite maintaining perfect precision. These variations are likely due to class imbalance and the smaller sample size of underrepresented categories such as Talent A.

Overall, ensemble-based models—particularly LightGBM and Hist Gradient Boosting—demonstrated superior performance for swimmer talent classification. Both achieved high predictive accuracy (test F1 ≥ 96.5%) while maintaining inference speeds suitable for real-time deployment. Extra Trees provided a strong accuracy–efficiency trade-off, whereas CatBoost achieved excellent accuracy but required significantly longer training. In comparison, traditional models such as Logistic Regression and Decision Trees, though computationally lightweight, lagged behind in predictive power and were less suitable for structured, keypoint-derived datasets.

Class-wise analysis confirmed robust recognition of Not Talent, Talent B, and Talent C, with minor limitations for Talent A caused by imbalance. Furthermore, the complete pipeline, which integrated RTMPose for keypoint extraction, was computationally and economically efficient. These results reinforce the feasibility of deploying the proposed system for practical swimmer talent identification and broader applications in pose-based sports analytics.

### *A. Discussion*

**Comment 3:** While SMOTE is mentioned for class balancing, its impact on performance is not analyzed, nor are potential overfitting risks for small classes (e.g., "Talent A") addressed. Including a sensitivity analysis or an ablation study on the balancing method would enhance the robustness of the findings.

SMOTE. As shown in Table 4, the overall effect was minimal for most categories. Not Talent, Talent A, and Talent C maintained almost identical performance across precision, recall, and F1. The main difference appeared in Talent B, where recall improved from 0.96 to 0.97, while precision decreased slightly from 0.97 to 0.96, resulting in a marginal F1 increase (0.96 → 0.97). These results suggest that SMOTE broadened the decision boundary to capture more true positives for Talent B, with only a minor cost in precision. For Talent A, performance remained unchanged, reflecting the limitations of oversampling when the original sample size is extremely small (n=8).

**Table 4 Classification performance of LightGBM before and after applying SMOTE**

| | Before | | | After | | | |
|---|---|---|---|---|---|---|---|
| Class | Precision | Recall | F1 Score | Precision | Recall | F1 Score | Number |
| Not Talent | 96 % | 98 % | 97 % | 97 % | 98 % | 97 % | 210 |
| Talent A | 100 % | 75 % | 86 % | 100 % | 75 % | 86 % | 8 |
| Talent B | 97 % | 96 % | 96 % | 96 % | 97 % | 97 % | 117 |
| Talent C | 100 % | 93 % | 96 % | 100 % | 93 % | 96 % | 14 |

### *B. Limitations and Future Work*

Although the proposed framework demonstrates strong performance, its evaluation was limited to underwater breaststroke footage. This restriction constrains the external validity of the findings, as the model's ability to generalize to other swimming strokes, above-water perspectives, and diverse pool environments has not yet been tested. To address this limitation, future work will extend the dataset to include freestyle, butterfly, and backstroke techniques, along with above-water recordings and data collected from multiple facilities. Such expansions will enable a more comprehensive evaluation of the framework's robustness, scalability, and applicability, ultimately supporting its adoption in real-world swimming analytics.

## V. Conclusion

This study proposed an end-to-end framework for real-time swimmer talent detection that integrates RTMPose-based pose estimation with machine learning classifiers, supported by temporal normalization using RIFE and class balancing through SMOTE. The comparative evaluation of twelve classifiers demonstrated the advantages of ensemble-based methods, with LightGBM achieving the best trade-off between accuracy, computational efficiency, and interpretability. Sensitivity analysis further confirmed the benefits of SMOTE in improving recognition of minority classes, while also highlighting potential risks of overfitting in small groups such as Talent A.

The results confirm that the proposed framework is both accurate and computationally efficient, making it a promising solution for practical deployment in aquatic sports analytics. Nevertheless, the framework's external validity is currently constrained by its reliance on underwater breaststroke footage.

## REFERENCE

[1] Williams, A.M. and T. Reilly, *Talent identification and development in soccer.* Journal of sports sciences, 2000. **18**(9): p. 657-667.

[2] Abbott, A. and D. Collins, *Eliminating the dichotomy between theory and practice in talent identification and development: considering the role of psychology.* Journal of sports sciences, 2004. **22**(5): p. 395-408.

[3] Lidor, R., J. CôTé, and D. Hackfort, *ISSP position stand: To test or not to test? The use of physical skill tests in talent detection and in early phases of sport development.* International journal of sport and exercise psychology, 2009. **7**(2): p. 131-146.

[4] Tharatipyakul, A., T. Srikaewsiew, and S. Pongnumkul, *Deep learning-based human body pose estimation in providing feedback for physical movement: A review.* Heliyon, 2024.

[5] Knap, P., *Human modelling and pose estimation overview.* arXiv preprint arXiv:2406.19290, 2024.

[6] Girshick, R., et al. *Deformable part models are convolutional neural networks*. in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2015.

[7] López-Sastre, R.J., T. Tuytelaars, and S. Savarese. *Deformable part models revisited: A performance evaluation for object category pose estimation*. in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. 2011. IEEE.

[8] Chen, H., et al., *2D Human pose estimation: A survey.* Multimedia systems, 2023. **29**(5): p. 3115-3138.

[9] Wang, J., et al., *Deep 3D human pose estimation: A review.* Computer Vision and Image Understanding, 2021. **210**: p. 103225.

[10] Zheng, C., et al., *Deep learning-based human pose estimation: A survey.* ACM Computing Surveys, 2023. **56**(1): p. 1-37.

[11] Huang, Z., et al. *Real-time intermediate flow estimation for video frame interpolation*. in *European Conference on Computer Vision*. 2022. Springer.

[12] Wei, X., P. Zhang, and J. Chai, *Accurate realtime full-body motion capture using a single depth camera.* ACM Transactions on Graphics (TOG), 2012. **31**(6): p. 1-12.

[13] Shan, Y., Z. Zhang, and K. Huang. *Learning skeleton stream patterns with slow feature analysis for action recognition*. in *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part III 13*. 2015. Springer.

[14] Wei, S.-E., et al. *Convolutional pose machines*. in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2016.

[15] Newell, A., K. Yang, and J. Deng. *Stacked hourglass networks for human pose estimation*. in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. 2016. Springer.

[16] Liu, W., et al., *Recent advances of monocular 2D and 3D human pose estimation: A deep learning perspective.* ACM Computing Surveys, 2022. **55**(4): p. 1-41.

[17] Song, C., et al. *Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting*. in *Proceedings of the AAAI conference on artificial intelligence*. 2020.

[18] Sun, R., et al., *Human action recognition using a convolutional neural network based on skeleton heatmaps from two-stage pose estimation.* Biomimetic Intelligence and Robotics, 2022. **2**(3): p. 100062.

[19] Kishore, D.M., S. Bindu, and N.K. Manjunath, *Estimation of yoga postures using machine learning techniques.* International Journal of Yoga, 2022. **15**(2): p. 137-143.

[20] Einfalt, M., D. Zecha, and R. Lienhart. *Activity-conditioned continuous human pose estimation for performance analysis of athletes using the example of swimming*. in *2018 IEEE winter conference on applications of computer vision (WACV)*. 2018. IEEE.

[21] Zecha, D. and R. Lienhart. *Key-pose prediction in cyclic human motion*. in *2015 IEEE Winter Conference on Applications of Computer Vision*. 2015. IEEE.

[22] Cao, X. and W.Q. Yan, *Pose estimation for swimmers in video surveillance.* Multimedia Tools and Applications, 2024. **83**(9): p. 26565-26580.

[23] Jiang, T., et al., *Rtmpose: Real-time multi-person pose estimation based on mmpose.* arXiv preprint arXiv:2303.07399, 2023.

[24] Powers, D.M., *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.* arXiv preprint arXiv:2010.16061, 2020.