Intelligent Assessment of Athlete Physical Fitness: Addressing Data Imbalance

Janyarat Phrueksanant[1*], Chayanont Awikunprasert[2], Jirachai Karawa[3], and Sutthirak Wisetsang[4]
[1]Department of Information Technology and Computer Innovation,
Faculty of Management Sciences and Information Technology, Nakhon Phanom University, Nakhon Phanom, Thailand
[2,3,4]Department of Sports Science, Faculty of Management Sciences and Information Technology,
Nakhon Phanom University, Nakhon Phanom, Thailand
E-mail: janyarat@npu.ac.th*, chayanona@yahoo.com, jckwmosza@gmail.com, watini8939@gmail.com

Received: February 28, 2025 / Revised: May 28, 2025 / Accepted: May 29, 2025

***Abstract*— This study aims to mitigate the impact of imbalanced data through the use of the oversampling technique and to develop supervised learning models for assessing the physical fitness of youth athletes. The dataset comprises the physical fitness test results from 75 athletes aged 11 to 16 years. The dataset presents two major challenges: A limited sample size and a significant class imbalance, with certain fitness levels being underrepresented. This class imbalance can substantially degrade the performance of classification models, as it often leads to biased predictions favoring the majority class while failing to learn the characteristics of minority classes, those that may be most critical in practice.**

To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) was employed to synthetically balance the class distribution. Five supervised learning algorithms were evaluated: Light Gradient Boosting Machine, Decision Tree, Random Forest, Neural Network, and Multinomial Logistic Regression. The Light Gradient Boosting Machine model yielded the highest accuracy at 87.76%, followed by Decision Tree, Random Forest, and Neural Network models, each with an accuracy of 79.59%. The Multinomial Logistic Regression model achieved the lowest accuracy at 75.51%. On average, the classification accuracy across all models improved to 81.41%, representing a 12.23% increase compared to using the original imbalanced dataset. The results demonstrate that applying oversampling techniques such as SMOTE can effectively alleviate the effects of class imbalance and enhance the predictive performance of machine learning models in the context of physical fitness assessment.

***Index Terms*—Athletes, Data Imbalance, Physical Fitness Performance, SMOTE, Supervised Learning**

I. Introduction

In recent years, the integration of data analytics into sports has advanced significantly, enabling more informed decision-making for both individual athletes and teams. Two case studies highlight this trend. The first involves a trainer for a women's college soccer team who utilizes wearable devices to collect internal (e.g., heart rate, body temperature, respiration) and external load data (e.g., running distance, speed, acceleration). These insights are used to monitor training intensity, prevent injuries, and ensure players are physically prepared for competition. Additional tools, such as single-leg squat tests, concussion assessments, sleep tracking, and periodic brain scans, contribute to a comprehensive understanding of each athlete's condition. The second case centers on a college football team led by a head coach and supported by a team operations expert. This team leverages annotated game footage, decision tree models, heatmaps, and time-series analytics to study opponents' tactics and optimize in-game strategy. Data-driven analysis is also applied to recruitment, incorporating advanced performance metrics such as reaction time, spatial awareness, and route-running precision, moving beyond traditional physical measurements. These examples demonstrate how data-driven approaches are transforming sports by enhancing athletic performance, minimizing risk, and informing strategic planning at both individual and organizational levels [1].

Physical fitness is a vital component of health and well-being, serving as the foundation for daily activities [2], [3]. Enhancing physical fitness not only improves quality of life but also serves as a key indicator of physical development over time. In athletic contexts, fitness assessments are commonly used to evaluate health status, identify individual strengths and weaknesses, and guide personalized training strategies aimed at optimizing performance.

This study aims to develop supervised learning models for assessing physical fitness levels among athletes. The dataset used was collected from physical fitness tests conducted at Nakhon Phanom Sports School, Thailand. As is common in real-world datasets, the collected data in this study are imbalanced, with one or more classes significantly underrepresented. Class imbalance is a well-documented challenge in data analytics, particularly in classification tasks, as it can negatively impact the accuracy and generalizability of predictive models. A relevant example is a previous study that utilized real-world data consisting of a small and imbalanced dataset to explore academic performance among IT students. The study employed Principal Component Analysis (PCA) and clustering techniques, and despite having only 115 samples, it successfully extracted meaningful insights. This highlights the effectiveness of dimensionality reduction and unsupervised learning methods in constrained data environments [4]. Given that most supervised learning algorithms assume balanced class distributions, several techniques have been proposed to address this issue [5]-[7]. In this study, the Synthetic Minority Oversampling Technique (SMOTE) is applied to augment the minority class, aiming to improve model robustness and classification performance.

The remainder of this paper is structured as follows. Section II provides a review of supervised learning techniques, the Synthetic Minority Oversampling Technique (SMOTE), and related literature. Section III describes the dataset, outlines the preprocessing procedures, and explains the implementation of SMOTE, as well as the models and evaluation metrics employed in the study. Section IV presents and analyzes the experimental results. Finally, Section V concludes the paper by summarizing the key findings and suggesting directions for future research.

II. Literature Review

A. *State-of-the-Art Supervised Learning Techniques for Classification*

Classification is one of the most extensively explored tasks in data analytics, aiming to categorize input data into predefined classes or labels. Supervised learning techniques form the foundation of classification models by learning patterns from labeled training data. Traditional methods such as Logistic Regression, Decision Trees, and Support Vector Machines (SVM) have been widely adopted due to their interpretability and effectiveness on structured datasets. More recent advancements include Ensemble Methods like Random Forests and Gradient Boosting Machines (e.g., XGBoost, LightGBM), which combine multiple weak learners to achieve higher accuracy and robustness. In addition, Artificial Neural Networks (ANNs) and deep learning architectures, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have demonstrated exceptional performance in complex, high-dimensional data such as images and sequences. These state-of-the-art models offer improved predictive power but often require larger datasets and computational resources. The selection of an appropriate classifier depends on factors such as data characteristics, model interpretability, scalability, and computational constraints. This section provides an overview of commonly used techniques, highlighting their strengths and state-of-the-art approaches.

1) Traditional Machine Learning Techniques

Logistic Regression (LR): Logistic Regression is a simple yet effective method for binary classification. It models the probability of a data point belonging to a class using a logistic function. Despite its simplicity, it performs well with linearly separable data and is widely used in various fields such as healthcare and finance.

Support Vector Machines (SVM): SVM is a robust classifier that aims to find the optimal hyperplane that separates data points of different classes with the maximum margin. Kernel methods allow SVM to handle non-linear data by transforming it into higher-dimensional spaces.

Decision Trees (DT): Decision Trees are hierarchical models that split data into subsets based on feature thresholds. They are interpretable and can handle both categorical and numerical data.

Ensemble Methods: Ensemble methods like Random Forest and Gradient Boosting (e.g., XGBoost, LightGBM, CatBoost) combine multiple weak learners to improve predictive performance.

2) Neural Networks and Deep Learning

Artificial Neural Networks (ANNs): ANNs are the foundation of modern deep learning. They consist of layers of interconnected nodes (neurons) that learn hierarchical representations of input data. ANNs are versatile and can model complex, non-linear relationships.

Convolutional Neural Networks (CNNs): CNNs are specialized for image and spatial data classification. They use convolutional layers to automatically extract features from raw data, making them state-of-the-art for tasks like object detection and image recognition.

Recurrent Neural Networks (RNNs) and Variants: RNNs, including Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), are used for sequential data like text, speech, and time series. They capture temporal dependencies in data.

3) Emerging Techniques and Trends

Explainable AI (XAI) in Classification: Modern classification tasks increasingly focus on interpretability. Techniques like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) help demystify complex models like deep learning.

Hybrid Models: Combining traditional machine learning with deep learning is becoming a trend, such as integrating feature engineering from domain knowledge into neural network architectures.

Supervised learning techniques for classification have evolved significantly, from traditional algorithms like logistic regression and support vector machines to advanced methods involving neural networks and transformers. Each technique has its strengths and limitations, making them suitable for specific types of data and tasks. Emerging trends, such as explainable AI, few-shot learning, and hybrid models, are paving the way for more robust and interpretable classification systems. As data continues to grow in complexity and volume, leveraging the right combination of techniques will be critical to advancing the field and solving real-world problems [8]-[12].

B. Synthetic Minority Oversampling Technique

Synthetic Minority Oversampling Technique (SMOTE) is a popular method for addressing class imbalance in datasets. It was introduced by Chawla et al. in 2002 [13]. The technique works by creating synthetic examples in the feature space to increase the representation of the minority class, rather than duplicating existing instances. For each sample in the minority class, it identifies its k-nearest neighbors within the minority class. Typically, k=5 is used. Then a random neighbor is chosen from the k-nearest neighbors. A new synthetic sample is created by interpolating between the original sample and the chosen neighbor. The formula is:

$$\mathcal{X}_{synthetic} = \mathcal{X}_i + (\mathcal{X}_j - \mathcal{X}_i).\lambda \quad (1)$$

Where:

$\mathcal{X}_i$ is the original minority class sample.

$\mathcal{X}_j$ is the chosen neighbor.

λ is a random number between 0 and 1.

This approach helps improve the performance of machine learning models, particularly those sensitive to class imbalance, like decision trees and neural networks.

C. Related Work

Studies on imbalanced data classification span diverse domains, employing advanced techniques to address challenges associated with class imbalance. One study focused on predicting osteoarthritis conditions in elderly patients, utilizing a dataset of 370 records divided into four classes: No symptoms, early symptoms, moderate symptoms, and severe symptoms. The study applied ADASYN and SMOTE to balance the data, using a 10-fold cross-validation method along with one-vs-one and one-vs-all multi-class classification approaches. The combination of ADASYN and the one-vs-one method achieved an impressive accuracy of 97.31% [14]. Another study examined career path predictions for 2,005 university graduates using employment status data. By leveraging oversampling, undersampling, and SMOTE, alongside decision tree and random forest classifiers, the study found that random forests, applied to oversampled data, yielded the best performance, achieving 67.17% accuracy, with corresponding precision, recall, and F-measure values of 0.66, 0.67, and 0.66, respectively [15]. Another interesting study that addresses the challenge of imbalanced datasets in predicting first-year engineering students' performance employs oversampling methods—SMOTE, Borderline-SMOTE, SVMSMOTE, and ADASYN—to balance the data, followed by classification using models such as Multi-Layer Perceptron (MLP), Gradient Boosting, AdaBoost, and Random Forest. The findings indicate that combining Borderline-SMOTE with various classifiers enhances the prediction accuracy for minority classes, thereby aiding in the early identification of students at risk of underperformance. This research underscores the effectiveness of integrating oversampling techniques with robust classifiers to improve predictive accuracy in educational settings, particularly for imbalanced [16]. Another study on educational data used the High School Longitudinal Study of 2009 (HSLS: 09) dataset to classify students into those likely to pursue higher education and those expected to defer or drop out. Sampling methods like ROS, RUS, SMOTE-NC, and hybrid techniques were tested, with hybrid resampling performing best for highly imbalanced datasets when paired with random forest classifiers [17]. Lastly, a study on physical fitness data classification analyzed relationships in students' physical fitness using a dataset of 812 records. The data, which included six fitness-related attributes such as BMI and exercise performance metrics, was classified using decision trees, random forests, and association rule mining techniques. The decision tree technique achieved a maximum accuracy of 100%, while random forests reached 99.8%. Additionally, association rule mining via Apriority and FP-Growth produced consistent patterns in the data. Collectively, these studies underscore the importance of advanced data balancing methods like SMOTE and hybrid resampling, as well as robust machine learning models such as random forests, in effectively addressing class imbalances across various datasets and applications. Tools like WEKA and Python were pivotal in facilitating these analyses, enabling efficient preprocessing, data balancing, and model evaluation [18]. Collectively, these studies highlight the effectiveness of data balancing techniques, particularly SMOTE and hybrid methods, and machine learning models like random forests in improving classification performance. These findings emphasize the critical role of advanced sampling methods and robust algorithms in addressing imbalanced data challenges across various domains, enabling more accurate predictions and better insights into complex datasets.

III. Methodology

A. Dataset

This section outlines the procedures used for data collection, which were carefully designed to ensure accuracy, consistency, and participant safety. The key steps involved are as follows:

Participants:

The study involved 75 male football student-athletes from Nakhon Phanom Sports School, Nakhon Phanom Province, Thailand, aged between 11 and 16 years. The participants were selected based on the following criteria:

• Being enrolled in the youth football training program under the school's athletic development system

• Being in good physical health with no injuries at the time of testing

• Having received informed consent from a parent, guardian, or coach

Data Collection Procedure:

1) *Preparation Phase*

Participants were informed in advance about the testing procedures. All athletes underwent a physical condition check and were instructed to perform warm-up exercises before each test. Testing environments were arranged according to the requirements of each physical test to ensure safety and measurement accuracy.

2) *Testing Phase*

All physical fitness tests were conducted individually. Each measurement was administered by trained evaluators or members of the research team. Data from each test was recorded immediately using standardized forms and later entered into a computer system for further analysis. The tests are divided into four aspects, which include:

Flexibility: Measured using the Sit and Reach test, with measurements recorded in centimeters to evaluate forward trunk flexibility.

Leg Muscle Strength: Measured using a Leg Strength Dynamometer, with results expressed in kilograms per body weight (kg/BW) to account for individual weight differences and provide a normalized measure of lower limb strength.

Muscular Endurance and Strength: Assessed using the 1-Minute Sit-Up test, with the number of correctly completed sit-ups recorded within one minute. The results were measured in repetitions.

Body Fat: Measured using a Body Composition Analyzer (The ACCUNIQ BC-360). The results were expressed as a percentage (%) of total body mass.

3) Data Cleansing

Collected data were verified for accuracy and consistency. Outliers were identified by comparing values against standard physical fitness benchmarks set by the Sports Authority of Thailand, Region 3 [19]. In cases where anomalies or measurement errors were detected, retesting was conducted for the concerned athlete, if necessary.

The physical fitness test results (Fitness Level) could be categorized into five levels: Very low (poor), low (fair), average, good, and very good (excellent). An example of the athletes' physical fitness test data is presented in Table I, while Table II summarizes the basic statistical values across the four main components of physical fitness.

TABLE I
EXAMPLES OF PHYSICAL FITNESS TEST DATA

No.	Sit and Reach (cm)	Leg Strength (kg/BW)	Sit Up 1 min (number)	Body Fat (%)
1	10.00	3.34	45	14.20
2	12.00	2.42	43	18.50
3	17.00	2.62	47	10.90
4	15.00	2.69	49	13.50
5	10.00	2.95	35	19.00

TABLE II
DESCRIPTIVE STATISTICS FOR PHYSICAL FITNESS TEST RESULTS

Finess Test	Min	Max	Average	Deviation
Sit and Reach (cm)	3.00	28.00	15.60	5.51
Leg Strength (kg/BW)	1.38	4.54	2.48	0.61
Sit Up 1 min. (number)	18.00	62.00	43.69	7.17
Body Fat (%)	3.00	28.60	14.49	5.77

From the analysis and evaluation of the physical fitness of 75 participants, it was found that 5 athletes were in the excellent category, 17 athletes were in the Good category, 23 athletes were in the Average category, 22 athletes were in the Fair category, and 8 athletes were in the Poor category. These figures represent 6.70%, 22.70%, 30.70%, 29.20%, and 10.70% of the total population, respectively. A bar chart showing the number of participants categorized by physical fitness test results is presented in Fig. 1.

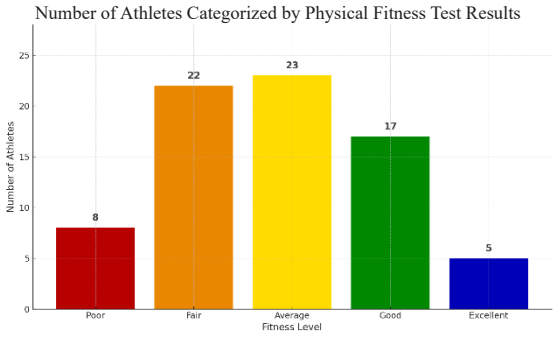

Fig. 1. Distribution of athletes by fitness level

It can be observed that the sample size is relatively small, and the proportion of data in each class varies significantly. The approach to addressing the issue of class imbalance will be discussed in the next section.

B. The Application of SMOTE to Handling Imbalanced Data

Several sampling methodologies have been developed to address class imbalance in supervised learning, including both under-sampling and oversampling techniques. Under sampling methods, such as Random Under sampling and Tomek Links, reduce the majority class size to balance the dataset, but risk losing valuable information. Oversampling approaches, on the other hand, expand the minority class by duplicating existing data or generating synthetic samples. Among these, SMOTE (Synthetic Minority Oversampling Technique) and ADASYN (Adaptive Synthetic Sampling) are widely used. While both generate synthetic samples, SMOTE creates them uniformly between minority class neighbors, whereas ADASYN focuses more on difficult-to-learn areas near the decision boundary. In this study, SMOTE was preferred over ADASYN due to its stability and balanced sample generation, which are especially important when working with small datasets. ADASYN, while effective in complex scenarios, can amplify noise and increase the risk of overfitting in limited or noisy datasets. Therefore, SMOTE was selected to ensure more controlled and interpretable oversampling outcomes [13], [20]. A total of 86 new samples were synthesized, resulting in a total of 161 records. These were classified as follows: 30 records for athletes with excellent performance, 32 records with good performance, 34 records with average performance, 33 records with fair performance, and 32 records with poor performance. This corresponds to proportions of 18.60%, 19.90%, 21.10%, 20.50%, and 19.90%, respectively. Table III shows the actual data and synthetic data for each class. Fig. 2 presents a bar chart showing actual data together with synthetic data classified by class. Section 4 presents and compares the experimental results.

TABLE III
THE NUMBER OF THE ACTUAL DATA AND SYNTHETIC DATA IN EACH CLASS

Class	Actual data	Synthetic data	Total
Poor	8	24	32
Fair	22	11	33
Average	23	11	34
Good	17	15	32
Excellent	5	25	30
TOTAL	**75**	**86**	**161**

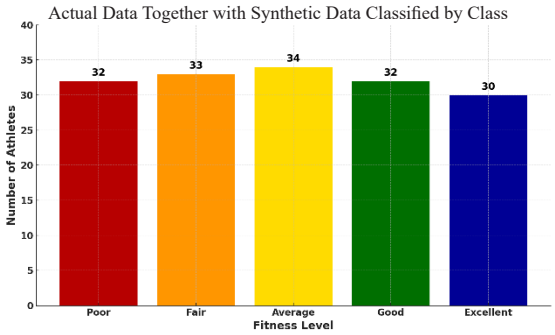

Fig. 2. Actual and synthetic data classified by class

It can be observed that this study synthesized a greater number of new data records compared to the original dataset. From the review of related studies, it was found that several studies have successfully applied the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples, often exceeding the number of original minority class samples, to address class imbalance and enhance model performance. Notable examples include: [13], [21]-[23]. These studies collectively demonstrate the successful application of SMOTE and its variants in generating synthetic samples that exceed the original number of minority class samples, leading to improved classifier performance in imbalanced datasets.

C. Tools used for Data Analysis

This study utilized RapidMiner Studio 9.10 [24] in conjunction with the Weka Extension [25] and the Python Scripting Extension [26] for both data preparation and developing models to evaluate the physical fitness of athletes. RapidMiner Studio is an integrated software designed to provide convenience and efficiency for data science tasks such as data mining and machine learning. It offers a wide range of operators, categorized by problem type, allowing users to select the most appropriate solution for their tasks. It could also be seamlessly integrated with other tools such as Weka Extension, Python Script, and Jupyter Notebook.

D. The Supervised Learning Model for Evaluating the Physical Fitness of Athletes

This study employed five supervised learning models, including Decision Tree, Random Forest, Neural Network, Multinomial Logistic Regression, and LightGBM (Python Learner). These models were implemented to evaluate datasets effectively, offering unique advantages and trade-offs. The brief details and working principles of these models are as follows [24]-[27]:

1) Decision Tree

The Decision Tree model is a simple yet powerful algorithm for classification and regression tasks. It works by splitting the dataset into subsets based on feature values, forming a tree-like structure. Each node represents a decision rule, and leaf nodes represent outcomes. Decision Trees are easy to interpret and visualize, suitable for datasets with mixed data types, but are prone to overfitting, especially with noisy data, unless pruned effectively.

2) Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and combines their outputs to improve accuracy and robustness. It addresses the overfitting issue of single decision trees by averaging predictions. Random Forests offer high accuracy and robustness to noise, are effective for handling large datasets and high-dimensional data, but are computationally intensive compared to single decision trees.

3) Neural Network

Neural Networks mimic the structure and functionality of the human brain, comprising layers of interconnected neurons. This model excels at identifying complex patterns and relationships in data. Neural Networks are versatile in handling structured and unstructured data, deliver high performance for large datasets with sufficient computational power, but require careful tuning of hyper parameters and may be prone to overfitting without sufficient training data.

4) Multinomial Logistic Regression

Multinomial Logistic Regression is a statistical model used for multi-class classification problems. It extends logistic regression by modeling the probability of each class as a function of input features. This model is simple and easy to implement, performs well with linearly separable data, but has limited ability to capture non-linear relationships in complex datasets.

5) LightGBM (Python Learner)

Python Learner in RapidMiner allows integration with Python scripts, offering flexibility for implementing custom algorithms. LightGBM (Light Gradient Boosting Machine), a widely used model implemented using Python Learner, is a high-performance gradient boosting framework based on decision tree algorithms. This model is fast, scalable, and efficient for large datasets, supports both continuous and categorical features, and is particularly effective for structured data. Unlike simpler models, LightGBM captures complex feature interactions and does not assume feature independence, making it well-suited for a wide range of real-world classification and regression tasks.

E. Model Performance Evaluation

This study utilizes the Confusion Matrix, a widely used tool for evaluating the outcomes of predictions or estimations made by developed models [27]. An example of a 2x2 Confusion Matrix is illustrated in Fig. 3.

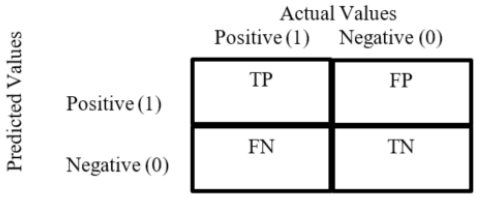

Fig. 3. Confusion matrix structure

Where:

TP: True Positive

FN: False Negative

FP: False Positive

TN: True Negative

The accuracy can be calculated by taking the percentage of correct predictions out of the total number of samples. Correct predictions are defined as instances where the predicted attribute value matches the actual target attribute value. The accuracy can be calculated as follows:

$$Accuracy = \left(\frac{Correct\ Predictions}{Total\ Samples}\right) * 100 \qquad (2)$$

IV. Experimental Results

For training and testing the model, the physical fitness test data was divided into two subsets: 70% was used to train the models, while the remaining 30% was reserved for evaluating model performance. Although the data was initially shuffled, stratified sampling was applied to ensure that the class distribution in both the training and testing sets reflected that of the overall dataset. It is important to note that the performance evaluation in this study was conducted solely on the test dataset.

This section presents the assessment results of athletes' physical fitness using both the original dataset and the newly synthesized dataset.

A. The Assessment Results of the Original Dataset

The assessment results of physical fitness assessment using Decision Tree, Random Forest, Neural Network, Multinomial Logistic Regression, and LightGBM techniques to classify the original dataset consisting of 75 records, as shown in Tables IV to VIII, respectively.

TABLE IV
THE ASSESSMENT RESULTS USING THE DECISION TREE: THE ORIGINAL DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	0	0	0	0	0	0.00%
pred. Good	1	3	0	0	0	75.00%
pred. Average	0	1	5	0	0	83.33%
pred. Fair	0	1	2	6	0	66.67%
pred. Poor	0	0	0	1	2	66.67%
class recall	0.00%	0.60%	71.43%	85.71%	100.00%	

Accuracy: 72.73%

TABLE V
THE ASSESSMENT RESULTS USING THE RANDOM FOREST: THE ORIGINAL DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	0	0	0	0	0	0.00%
pred. Good	1	2	0	0	0	66.67%
pred. Average	0	3	5	1	0	55.56%
pred. Fair	0	0	2	5	1	62.50%
pred. Poor	0	0	0	1	1	50.00%
class recall	0.00%	40.00%	71.43%	71.43%	50.00%	

Accuracy: 59.09%

TABLE VI
THE ASSESSMENT RESULTS USING THE NEURAL NETWORK: THE ORIGINAL DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	1	1	0	0	0	50.00%
pred. Good	0	3	0	0	0	100.00%
pred. Average	0	1	5	0	0	83.33%
pred. Fair	0	0	2	6	0	75.00%
pred. Poor	0	0	0	1	2	66.67%
class recall	100.00%	60.00%	71.43%	85.71%	100.00%	

Accuracy: 77.27%

TABLE VII
THE ASSESSMENT RESULTS USING THE MULTINOMIAL LOGISTIC REGRESSION: THE ORIGINAL DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	0	1	1	0	0	0.00%
pred. Good	1	3	0	0	0	75.00%
pred. Average	0	1	5	0	0	83.33%
pred. Fair	0	0	1	6	0	85.71%
pred. Poor	0	0	0	1	2	66.67%
class recall	0.00%	60.00%	71.43%	85.71%	100.00%	

Accuracy: 72.73%

TABLE VIII
THE ASSESSMENT RESULTS USING THE LightGBM: THE ORIGINAL DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	0	0	1	0	0	0.00%
pred. Good	0	3	1	0	0	75.00%
pred. Average	1	2	4	0	0	57.14%
pred. Fair	0	0	1	6	2	66.67%
pred. Poor	0	0	0	1	0	0.00%
class recall	0.00%	60.00%	57.14%	85.71%	0.00%	

Accuracy: 59.09%

B. The Assessment Results of the New Dataset

The Confusion Matrix tables present the assessment results of physical fitness assessment using Decision Tree, Random Forest, Neural Network, Multinomial Logistic Regression, and LightGBM techniques to classify the original dataset combined with synthetic data, 161 records in total shown in Tables IX-XIII, respectively.

TABLE IX
THE ASSESSMENT RESULTS USING THE DECISION TREE: THE NEW DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	6	2	1	0	0	66.67%
pred. Good	3	8	0	0	0	72.73%
pred. Average	0	0	6	1	0	85.71%
pred. Fair	0	0	3	9	0	75.00%
pred. Poor	0	0	0	0	10	100.00%
class recall	66.67%	80.00%	60.00%	90.00%	100.00%	

Accuracy: 79.59%

TABLE X
THE ASSESSMENT RESULTS USING THE RANDOM FOREST: THE NEW DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	6	2	0	0	0	75.00%
pred. Good	1	8	0	0	0	88.89%
pred. Average	0	0	7	2	0	77.78%
pred. Fair	2	0	3	8	0	61.54%
pred. Poor	0	0	0	0	10	100.00%
class recall	66.67%	80.00%	70.00%	80.00%	100.00%	

Accuracy: 79.59%

TABLE XI
THE ASSESSMENT RESULTS USING THE NEURAL NETWORK: THE NEW DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	8	2	0	0	0	80.00%
pred. Good	1	7	2	0	0	70.00%
pred. Average	0	1	7	1	0	77.78%
pred. Fair	0	0	1	9	2	75.00%
pred. Poor	0	0	0	0	8	100.00%
class recall	88.89%	70.00%	70.00%	90.00%	80.00%	

Accuracy: 79.59%

TABLE XII
THE ASSESSMENT RESULTS USING THE MULTINOMIAL LOGISTIC REGRESSION: THE NEW DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	7	3	0	0	0	70.00%
pred. Good	2	6	0	0	0	75.00%
pred. Average	0	1	9	0	0	90.00%
pred. Fair	0	0	1	8	3	66.67%
pred. Poor	0	0	0	2	7	77.78%
class recall	77.78%	60.00%	90.00%	80.00%	70.00%	

Accuracy: 75.51%

TABLE XIII
THE ASSESSMENT RESULTS USING THE LightGBM: THE NEW DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	9	1	0	0	0	90.00%
pred. Good	0	8	0	0	0	100.00%
pred. Average	0	1	9	0	0	81.82%
pred. Fair	0	0	1	2	2	75.00%
pred. Poor	0	0	0	8	8	100.00%
class recall	100.00%	80.00%	90.00%	90.00%	80.00%	

Accuracy: 87.76%

C. *Comparison of Results*

The assessment accuracy of all five models on the two physical fitness datasets is presented in Table XIV.

TABLE XIV
A SUMMARY OF THE ASSESSMENT ACCURACY OF ALL FIVE MODELS ON FIVE MODELS ON THE TWO PHYSICAL FITNESS DATASETS

List	The Original Dataset (%)	The New Dataset (%)
Decision Tree	72.73	79.59
Random Forest	59.09	79.59
Neural Network	77.27	79.59
Multinomial Logistic Regression	72.73	75.51
LightGBM	59.09	87.76
Average	68.18	80.40

The table demonstrates that in the evaluation of physical fitness using the original dataset, the Neural Network model performed the best, achieving the highest accuracy of 77.27%. It was followed by the Decision Tree and Multinomial Logistic Regression models, both with an accuracy of 72.73%. The Decision Tree and LightGBM models showed the lowest accuracy at 59.09%. For the dataset consisting of 161 records, original data combined with 86 synthetic records generated using the SMOTE method, the LightGBM (Python Learner) model achieved the highest accuracy of 87.76%. The Decision Tree, Random Forest, and Neural Network models followed, with the highest accuracy being 79.59%. The Multinomial Logistic Regression model showed the lowest accuracy in evaluating this dataset, with an accuracy of 75.51%. As observed, the evaluation of physical fitness using the original dataset combined with the newly synthesized data, totaling 161 records, showed improved performance. The highest accuracy reached 87.76%, achieved by the LightGBM (Python Learner) model, which had an average accuracy of 81.41%. This represents an increase of 12.23% compared to the average performance of the evaluation using only the original dataset. The analysis revealed that LightGBM outperformed other supervised learning models due to its highly efficient gradient boosting framework. Unlike traditional tree construction methods, LightGBM builds trees in a leaf-wise manner, resulting in greater loss reduction and improved accuracy. Compared to other models such as logistic regression, decision trees, or Support Vector Machines (SVMs), LightGBM handles large-scale and high-dimensional datasets with

superior speed and precision, thanks to its histogram-based computation and advanced sampling strategies. Moreover, it natively supports missing values and categorical features, reducing the need for extensive pre-processing. When implemented through the Python Learner in RapidMiner Studio, LightGBM offers flexible hyper parameter tuning and integrates seamlessly into the analytical workflow, making it a powerful and scalable solution for predictive modelling tasks. While synthetic data generation techniques such as SMOTE can effectively address class imbalance and improve model performance, they may also introduce certain biases. For instance, synthetic samples are created by interpolating between existing minority class instances, which can lead to the over-representation of specific regions in the feature space while neglecting others. This may result in models that generalize poorly or are overly confident in areas where no real data exists. Additionally, if the original dataset contains noise or mislabelled instances, synthetic sampling may inadvertently amplify these issues. Therefore, careful validation and data quality checks are essential when using synthetic data to ensure that it enhances rather than distorts the learning process.

V. Conclusions

This research collected physical fitness test data from 75 student-athletes aged 11-16 at the Nakhon Phanom Sports School. The tests were divided into four aspects: flexibility, leg muscle strength, endurance and muscle strength, and body fat. The collected data underwent cleaning (Data Cleansing) to ensure accuracy and was initially evaluated using the standard youth athlete fitness test criteria of the Sports Authority of Thailand, Region 3. The physical fitness test results (Fitness Level) were classified into five levels: Poor, Fair, Average, Good, and Excellent. Analysis revealed that 5 students were in the Excellent category, 17 in Good, 23 in Average, 22 in Fair, and 8 in Poor, accounting for 6.70%, 22.70%, 30.70%, 29.20%, and 10.70% of the total population, respectively. This dataset is imbalanced, which could affect the performance of models used for data processing. To address this imbalance and improve supervised learning model performance, this study utilized the Synthetic Minority Oversampling Technique (SMOTE), a special oversampling method that generates new synthetic data points instead of replicating existing ones. The machine learning models used for evaluating the student-athlete dataset included Decision Tree, Random Forest, Neural Network, Multinomial Logistic Regression, and LightGBM (Python Learner). The findings revealed that applying SMOTE to increase the dataset size to 161 records with SMOTE improved performance, achieving a maximum accuracy of 87.76%. The LightGBM (Python Learner) model demonstrated an average accuracy of 81.41%, representing a 12.23% improvement compared to the average performance with the original dataset. In summary, the results indicate that using oversampling techniques like SMOTE can mitigate data imbalance issues and enhance data classification or evaluation performance when the additional data is generated in an appropriate quantity. However, other factors may also influence model performance and data classification efficiency. In summary, these models, such as LightGBM and other supervised learning techniques, could be implemented in real-world athlete training programs to monitor performance, predict injury risks, and personalize training plans based on physiological data. By leveraging predictive analytics, coaches and sports scientists can make data-driven decisions that optimize performance and reduce overtraining. Future research could explore the integration of real-time data from wearable devices and expand predictive modelling to include psychological, nutritional, and environmental factors for a more holistic view of athlete development.

Acknowledgment

We would like to express our gratitude to the Nakhon Phanom Sports School for their support in facilitating the physical fitness testing of student-athletes. Additionally, we extend our sincere thanks to the Faculty of Management Sciences and Information Technology, Nakhon Phanom University, for providing funding support for this study.

References

[1] R. Sharda, D. Delen, and E. Turban, *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*, 4th ed. Harlow, UK: Pearson, 2018, pp. 33-36.

[2] R. C. Prati, G. E. Balista, and M. C. Monard, "Data mining with imbalanced class distributions: concepts and methods," in *Proc. Indian Int. Conf. Artif. Intell., IICAI 2009*, 2009, pp. 359-376.

[3] G. Lemaître, F. Nogueira, and C. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1-15, Sep. 2016.

[4] U. Ninrutsirikun, H. Imai, B. Watanapa, and C. Arpnikanondt, "Principal component clustered factors for determining study performance in computer programming class," *Wireless Pers. Commun.*, vol. 115, no. 4, pp. 2897-2916, Dec. 2020.

[5] R. Liu, "A novel synthetics minority oversampling technique based on relative and absolute densities for imbalanced classification," *Appl. Intell.*, vol. 53, pp. 786-803, Apr. 2003.

[6] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263-1284, Sep. 2009.

[7] Q. Yang and X. Wu, "10 challenging problems in data mining research," *Int. J. Inf. Technol. Decis. Mak.*, vol. 5, no. 4, pp. 597-604, Dec. 2006, https://doi.org/10.1142/S0219622006002258

[8] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825-2830, Oct. 2011.

[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 85-794.
[10] A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5998-6008, Aug. 2017, https://doi.org/10.48550/arXiv.1706.03762
[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 4765-4774, Nov. 2017, https://doi.org/10.48550/arXiv.1705.07874
[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770-778.
[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321-357, Jun. 2002, https://doi.org/10.1613/jair.953
[14] P. Thanathammathee and Y. Sirisathitkul, "Improved classification techniques for imbalanced dataset of elderly's knee osteoarthritis," *J. Sci. Technol.*, vol. 27, no. 6, pp. 1164-1178, Nov.-Dec. 2019.
[15] S. Wannont and R. Muangsarn, "Improving prediction models of student business career using sampling techniques for learning in multi-classes imbalance dataset," *Chaiyaphum Parithat J.*, vol. 4, no. 1, pp. 39-49, Jan.-Apr. 2021.
[16] N. Rachburee and W. Punlumjeak, "Oversampling technique in student performance classification from engineering course," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 4, pp. 3567-3574, Aug. 2021.
[17] T. Wongvorachan, S. He, and O. Bulut, "A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining," *Information*, vol. 14, no. 1, p. 54, Jan. 2023.
[18] C. Masachai, N. Srisahno, W. Masachai, and W. Buathong, "Relationship physical fitness assessment results of students with data mining at Rajaprachanugroh 1 school," *Ind. Technol. Lampang Rajabhat Univ.*, vol. 14, no. 2, pp. 1-11, Jul.-Dec. 2021.
[19] S. Kusum, C. Chiewsakul, J. Naksri, N. Mudchanthuek, and W. Deeniwong, "Physical fitness test and standard guidelines for youth athletes," *Regional Sports Sci. Work, Sports Authority of Thailand, Region 3, Ministry of Tourism and Sports*, 2019. [Online]. Availabe: https://set3.org [Accesied: Apr. 25, 2024]
[20] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, 2008, pp. 1322-1328.
[21] D. Elreedy and A. F. Atiya, "A theoretical distribution analysis of the synthetic minority oversampling technique," *Mach. Learn.*, vol. 111, no. 1, pp. 157-180, Jan. 2024.
[22] R. Liu, "A novel synthetic minority oversampling technique based on relative and absolute densities for imbalanced classification," *Appl Intell*, vol. 53, p. 786803, Jan. 2023, https://doi.org/10.1007/s10489-022-03512-5
[23] G. Douzas and F. Bacao, "Geometric SMOTE: Effective oversampling for imbalanced learning through a geometric extension of SMOTE," *Inf. Sci.*, vol. 465, pp. 1-20, Sep. 2017.
[24] RapidMiner, Inc., "RapidMiner Studio, version 9.10." *Docs. Rapidminer. Com*. 2024. [Online]. Available: https://www.rapidminer.com [Accessed May 25, 2024].
[25] M. Hall, E. Frank, and G. Holmes, "The WEKA data mining software: An update," *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10-18, Nov. 2009, https://doi.org/10.1145/1656274.1656278
[26] G. van Rossum and F. L. Drake, "Python 3 reference manual," *(S. l.) ACM DL*. 2009. [Online]. Available: https://dl.acm.org/doi/abs/10.5555/1593511 [Accessed Mar. 20, 2009].
[27] R. Kohavi and F. Provost, "Glossary of terms," *Mach. Learn.*, vol. 30, no. 2-3, pp. 271-274, Jan. 1998.

Janyarat Phrueksanant received her M.Sc. in Information Technology from the School of Information Technology, King Mongkut's University of Technology Thonburi, and Ph. D. in Systems Engineering from the School of Engineering, Cardiff University. She currently works as a lecturer at the Department of Information Technology and Computer Innovation, Faculty of Management Sciences and Information Technology, Nakhon Phanom University.

Chayanont Awikunprasert received his M.Sc. in Human Development from the Faculty of Graduate Studies, Mahidol University, and Ph.D. in Exercise and Sport Science from the Faculty of Sport Science, Burapha University. He currently works as a lecturer at the Department of Sports Science, Faculty of Management Sciences and Information Technology, Nakhon Phanom University.

Jirachai Karawa received his B.Sc. in Sport Science from the Faculty of Education, Mahasarakham University, and M.Sc. in Exercise and Sport Sciences from the Graduate School, Khon Kaen University. He currently works as a lecturer at the Department of Sports Science, Faculty of Management Sciences and Information Technology, Nakhon Phanom University.

Sutthirak Wisetsang received his B.Ed. and M.Ed. in Physical Education from the Faculty of Education, Ramkhamhaeng University. He currently works as a lecturer at the Department of Sports Science, Faculty of Management Sciences and Information Technology, Nakhon Phanom University.