Optimizing Hyperparameters in Machine Learning Models for Accurate Fitness Activity Classification in School-Aged Children

Britsel Calluchi Arocutipa[1], Magaly Villegas Cahuana[2], Vanessa Huanca Hilachoque[3], Marco Cossio Bolaños[4]
Ingeniería de Sistemas, Universidad Nacional De San Agustín De Arequipa, Arequipa, Perú[1, 2, 3]
Universidad Católica Del Maule, Talca, Chile[4]

***Abstract*—Classification using machine learning algorithms in physical fitness tests carried out by students in educational centers can help prevent obesity and other related diseases. This research aims to evaluate physical fitness using percentiles of the tests and machine learning algorithms with hyperparameter optimization. The process followed was knowledge discovery in databases (KDD). Data were collected from 1525 students (784 women, 741 men) aged 6 to 17, selected non-probabilistically from five public schools. For the evaluation, anthropometric parameters such as age, weight, height, sitting height, abdominal circumference, relaxed arm circumference, oxygen saturation, resting heart rate, and maximum expiratory flow were considered. Physical Fitness tests included sitting flexibility, kangaroo horizontal jump, and 20-meter fly speed. Within the percentiles observed, we took three cut-off points as a basis for the present research: > P75 (above average), p25 to p75 (average), and < P25 (below average). The following machine learning algorithms were used for classification: Random Forest, Support Vector Machine, Decision tree, Logistic Regression, Naive Bayes, K-nearest neighbor, XGBboost, Neural network, Cat Boost, LGBM, and Gradient Boosting. The algorithms were hyperparameter optimized using GridSearchCV to find the best configurations. In conclusion, the importance of hyperparameter optimization in improving the accuracy of machine learning models is highlighted. Random Forest performs well in classifying the "High" and "Low" categories in most tests but struggles to correctly classify the "Normal" category for both male and female students.**

***Keywords*—Machine learning; classification; physical fitness; schoolchildren; hyperparameters**

I. Introduction

Machine learning (ML) is a subset of AI that involves building computer models capable of learning and making independent predictions or decisions based on the provided data [1]. In its operation, ML allows you to train a model to categorize data based on selected characteristics. It is classified into two broad categories: supervised and unsupervised. Unsupervised Machine Learning is used to conclude from data sets that contain input data without labeled responses. On the other hand, supervised machine learning attempts to discover the relationship between input attributes (independent variables) and a target attribute (dependent variable) [2]. This approach has a wide range of applications, including sectors like healthcare, education, and technological advancements. The applications are varied and can be integrated with the use of wearable technologies to track physical activity and monitor health conditions [3].

Physical activity is any body movement that results in an increase in energy expenditure above the resting level. Regular physical activity has been shown to help prevent and control non-communicable diseases, such as heart disease, stroke, diabetes, and several types of cancer. According to the WHO, more than 80% of adolescents worldwide have an insufficient level of physical activity, and it recommends that children between 5 and 17 years old dedicate at least an average of 60 minutes a day to moderate to intense physical activities, mainly aerobic, throughout the week [4].

Lack of physical activity, poor eating habits, and sedentary behaviors, such as excessive use of technology to watch television, play video games, or use cell phones, and even lack of sleep, have led to an increase in the prevalence of overweight in recent years [5].

According to a UNICEF report from 2023, in Latin America and the Caribbean, there are nearly 49 million children and adolescents between 5 and 19 years old who are overweight, which represents 30.6% of the population, above the global prevalence of 18.2 percent. South America has the highest number of people affected, with 30 million overweight, followed by Central America with 16 million and the Caribbean with three million. Argentina, Bahamas, Chile, and Mexico have the highest prevalence, with more than 35 percent. Peru also has a high prevalence, at 25 percent. Furthermore, the report shows differences by sex: the prevalence is 27 percent in men and 27.9 percent in women [6].

The analysis by Andermo et al. [7] underlines the effectiveness of school initiatives that promote physical activity among children and young people. According to these results, these actions reduce anxiety, strengthen resilience, improve well-being, and promote positive mental health.

The motivations of interest concern the Health of Students, the high prevalence of overweight, and the significant lack of physical activity among students. This public health problem requires innovative solutions that can be implemented on a large scale. The potential of machine learning with the application of specifically supervised ML algorithms can provide new insights and tools to classify and evaluate the physical fitness of schoolchildren. This will allow for a more targeted and personalized intervention. Innovation in Physical

Education: Integrating advanced technologies such as ML into physical fitness assessment can revolutionize how physical activity is understood and promoted in educational centers. Optimizing hyperparameters in ML models ensures that predictions and classifications are as accurate as possible, which is crucial for designing effective interventions.

Therefore, the paper aims to explore the level of physical fitness and the application of machine learning algorithms optimized by hyperparameters to optimally classify the physical fitness of schoolchildren from educational centers so that artificial intelligence techniques can contribute to health and academic contexts.

Creating a supervised machine learning model optimized for classifying the physical fitness level of schoolchildren is a significant contribution. This model can accurately evaluate different physical parameters and provide a detailed classification that facilitates personalized intervention. Implementing and analyzing hyperparameter optimization techniques will improve the accuracy and effectiveness of predictive models. This methodological approach can be applied in other areas of study that use machine learning, providing a framework to improve the quality of predictions. The research will provide a detailed analysis of how machine learning can evaluate and improve school initiatives that promote physical activity.

The article is organized as follows: Section I with the introduction, Section II has the literature review, Section III develops the methodology used, and Section IV presents the results obtained. Finally, the discussion and conclusion are given in Section V and Section VI respectively.

II. Literature Review

There is more than one way to measure physical activity levels (whether manually, with questionnaires, wearable technology, or smart devices), which will help classify them into levels. In this sense, we present the main works carried out with the topic under study.

Trejo et al. [8] indicate that obesity is a problem worldwide. Even more so, with the advancement of technology, many schoolchildren lead sedentary lifestyles due to being immersed in social networks and virtual video games. According to the results obtained, children with obesity spend an average of three hours watching television programs or playing video games.

According to the study by [9], Physical Activity, Diet Quality, and Physical Condition should be assessed early, considering it a physiological need to contribute to a healthy lifestyle and improve the child's future quality of life. The school framework is taken at an early average age of 8 - 12 since it is considered the ideal environment to promote good, healthy behaviors.

Zhou et al. [10] show that predicting adherence or commitment to physical activity is of utmost importance since it prevents a relapse in exercise, using automatic prediction messages with Logistic Regression and Support Vector Machine (SVM) models. This research conducted tests on sedentary people, testing their resistance to physical activity for a specific time. The Logistic Regression model demonstrated a slightly better performance than the Support Vector Machine (SVM). It should be noted that the precision in both models is high.

Alsareii et al. [11] mention physical activity is essential in controlling obesity and maintaining a healthy life. Tracking physical activities using state-of-the-art automatic techniques can promote healthy living and control obesity. This work introduces novel techniques to identify and record physical activities using machine learning techniques and wearable sensors.

Ahmadi, Pavey, and Trost [12] mention as the objective of the study the evaluation of the accuracy of Random Forest (RF) activity classification models for preschool children trained with data from free-living accelerometers in children in the range of four to nine years concluding that RF activity classification models trained with free-living accelerometer data provide accurate recognition of young children's movement behaviors under real-world conditions.

In study [13], students from public schools in Arequipa (Peru) were evaluated, and they performed the classification of motor competence based on the evaluation with the most popular machine learning algorithms optimized by their hyperparameters. The tests assessed using wearable technology were related to motor competence in schoolchildren aged 6 to 17. Different data was captured using a pedometer, accelerometer, and heart rate sensors. As a result, percentiles of schoolchildren were created. Regarding classification, the gradient boosting algorithm with its optimization of hyperparameters with the RandomizedSearchCV technique was the one that obtained the best precision of 0.95 and in the ROC-AUC curves with 0.98. Additionally, they developed software that implemented the model that was built.

Current research has identified several areas of intervention to optimize the classification of lack of physical activity in schoolchildren. Still, it has also highlighted significant gaps that must be addressed to improve the effectiveness of prevention and treatment strategies. Among them is the improvement of predictive models, which can be achieved by incorporating the configuration of hyperparameters. The comparison of tracking techniques, making an exhaustive comparison of different automatic fitness activity tracking techniques in terms of cost-effectiveness and accessibility, proposing practical solutions for implementation in school programs. The classification models can be integrated by incorporating Fitness Activity classification models into school programs and providing guidelines for their use by educators and parents. Algorithm optimization, performing a comparative analysis of the machine learning algorithms used to classify Fitness Activity, determining the most effective and accessible implementation in various contexts.

III. Methodology

It is necessary to perform data analysis as a fundamental process for classifying schoolchildren and obtaining valuable information on physical activity results. There are different methods and techniques to classify and perform data analysis. The KDD (Knowledge Discovery in Databases) process was

used for this research to extract knowledge from large volumes of data. It consists of a series of defined stages applied before using data mining techniques to search for hidden patterns in the data and analyze the patterns found. KDD uses a structured set of stages to address data mining projects, from understanding data to obtaining knowledge, as shown in Fig. 1; this represents how we use the methodology following the stages in the data process of physical activity, generating knowledge. The methodology used was KDD, which consists of six stages: Starting from Selection, followed by Preprocessing, Transformation, Mining, and Interpretation to reach the knowledge we want. Fig. 1 shows the KDD process.

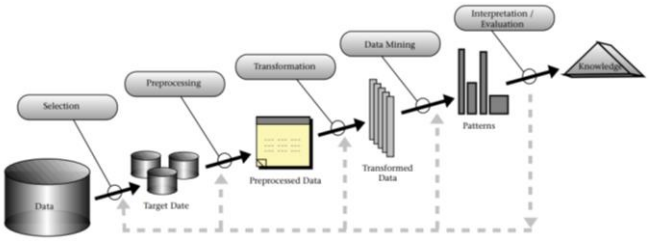

Fig. 1. KDD process [14].

A. Data Selection

The selection of data sources consisted of searching the data for appropriate input attributes of physical activity, obtaining 1525 students (784 women, 741 men) from 6 to 17 years old, including children and adolescents; the sample selection was non-probabilistic. This means knowing what you want to obtain and what data will facilitate this process to achieve the results.

The evaluations of the motor competence tests were carried out in public schools in the city of Arequipa (Peru). The tests were carried out during physical education sessions. The schoolchildren were previously informed about the evaluation to be carried out.

For this research, anthropometric measurements were carried out in the facilities of each school, working with children and adolescents with the authorization of their parents or guardians. Data collection and evaluation were carried out by two physical education teachers with experience in similar work. It was assessed according to the standardized protocol of Ross and Marfell-Jones to capture the standing height and weight measurements. Body weight (in kilograms) was measured using a BC-730 electronic scale, with a range of 0 to 150 kg and a precision of 100 grams. Standing height was measured according to the Frankfurt plane using a portable stadiometer with an accuracy of 0.1 mm. To divide abdominal fat (AF) into categories by sex and age, the suggestions described by Fernández in [15] were followed.

B. Pre-Processing

Data was collected during the different physical education sessions to pre-process this research. Once collected, the following steps were carried out, shown in Table I, necessary for use in the classification algorithms:

1) Data cleaning: In this step, the data was cleaned, including incomplete data (where there are missing attributes or attribute values), noise (incorrect or unexpected values), and inconsistent data (containing values and characteristics with different names). Conflicting data were eliminated because they would allow inadequate analysis and incorrect results.

The data cleaning tasks to be executed by the Jupyter panel were written in Python 3. The Pandas library, used for data manipulation and analysis, uses the Python Scikit-learn library [16], as shown in greater detail in Table II.

Fig. 2 shows the distribution of the students' classes: High, Normal, and Low physical activity for both sexes.

2) Data transformation: The data was normalized to be on the same scale since some machine learning algorithms are sensitive to scales. The physical fitness tests considered according to the specialists were:

a) Flexibility (cm): The dorsal-lumbar flexibility, sitting posture, and modified reach were measured.

b) Horizontal jump (cm): The horizontal jump was measured by the number of attempts in the "kangaroo" test.

c) Speed 20m (seconds): It was evaluated ten times in the 20-meter race and assessed in seconds with a stopwatch.

The percentile tables were used to consider the research [17]. These percentiles were divided into male and female ranges, and the parameters for each test taken for this investigation were shown.

TABLE I. DATA PRE-PROCESSING

Phases	Description
Data Cleaning	Errors in the data, such as empty spaces and punctuation marks not allowed, were detected and corrected.
Data integration	The data collected from the sessions was combined to provide a unified view by integrating them into a single format.
Data transformation	In this stage, categorical variables were normalized and standardized, and coding was carried out, transforming the data into a uniform format.
Data reduction	The data to be processed were identified for each record having ten input attributes.

TABLE II. APPLICATION OF TECHNOLOGIES

Technology	Description
Python 3.6	Python is the programming language we will use to analyze and process algorithms.
Colab	It is the environment that can run and program in Python. It provides a flexible environment for Python programming and other scientific computing tasks.
Scikit-learn	Python library includes various supervised and unsupervised machine learning algorithms, which are widely used for their ease of use, and a wide range of data analysis and modeling tools.

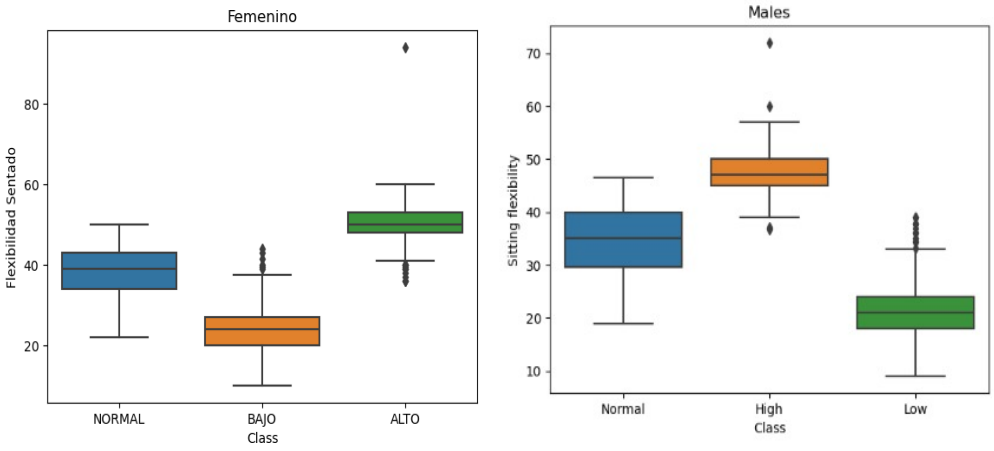

Fig. 2. Boxplot of the classes of female and male school children.

TABLE III. PERCENTILES OF THE FLEXIBILITY, HORIZONTAL JUMP, AND SPEED 20M TESTS FOR MEN AGES 6 TO 18

	Age												
Percentile/Test	6	7	8	9	10	11	12	13	14	15	16	17	18
Flexibility													
P25	39	18.5	24	21.8	24	26.5	42	35	24.8	25	26	28	29
P50	41.5	22	36	32.5	36	38	46	43.7	31	34	32	32	31
P75	43.1	36.1	44	43	43	44.5	47	47.6	42.5	45	43	38	36
Horizontal Jump													
P25	82.8	76.5	87	99.3	98	105	115	110	110	119	121	132	144
P50	91	93	110	110	115	120	120	123	130	134	143	150	155
P75	97	104	117	119	125	130	120	142	150	160	165	178	180
Speed 20m													
P25	5.02	4.32	4.22	4.15	4.2	4	3.32	4.08	3.6	3.4	3	2.98	2.9
P50	5.59	4.68	4.92	4.86	4.65	4.2	3.75	4.68	3.82	3.9	3.2	3.2	3
P75	6.14	5.1	5.38	5.43	4.89	4.59	3.8	4.97	4.28	4.31	3.5	3.6	3.2

TABLE IV. PERCENTILES OF THE FLEXIBILITY, HORIZONTAL JUMP, AND SPEED 20M TESTS FOR FEMALES AGES 6 TO 18

	Age												
Percentile/Test	6	7	8	9	10	11	12	13	14	15	16	17	18
Flexibility													
P25	36.5	21	22.3	25	29	27	44	40	27	33	29.8	27	30.5
P50	39.5	29.8	38	40	41	39.5	47	45	34.5	44	39.5	34	34
P75	43.8	39.4	46.8	47	47	46.5	53	50	48	51	49.3	39	36
Horizontal Jump													
P25	70	69.8	80	90	90	90	105	107	95.3	104	90	102	102
P50	81	82	95	102	105	105	111	113	110	112	109	117	110
P75	88	88.5	105	111	112	114	114	119	118	122	120	129	126
Speed 20m													
P25	5.4	4.58	4.64	4.3	4.44	3.9	4.1	4.52	3.8	4	3.78	3.6	3.65
P50	5.99	4.94	5.2	4.91	4.66	4.4	4.12	4.87	4.25	4.5	4.36	3.7	3.8
P75	6.2	5.4	5.67	5.54	5.29	4.94	4.63	5.35	4.81	5.19	5.12	4	4.05

Tables III and IV present the results of the evaluated tests and their corresponding percentiles: P25, P50, and P75. These percentiles were used as cut-off points in this research. Values close to the P75th percentile are considered excellent, while values at the P25th percentile are rated poor. The grade depends on the type of physical test and the objective. According to the studies, the following cut-off points were proposed for the diagnosis of physical fitness: 75 (High), 50 (Normal), and 25 (Low). The previously studied percentiles were fundamental for this research, providing a solid basis for establishing the cut-off points.

C. Transformation

For balance, the records were first sorted randomly, and then 80% were selected for training and 20% for testing. In summary, this process consisted of three phases: defining and determining the types of errors, finding and identifying instances containing errors, and correcting the discovered errors.

The registration of schoolchildren had the problem of imbalance in data sets, which is a significant problem in classification operations. When one class is significantly more frequent than others, machine learning algorithms can become biased toward the majority class, resulting in poor performance in predicting the minority class. Data balancing helps mitigate this problem; for this reason, data balancing was carried out to correctly predict the minority classes for a data set in the data analysis.

D. Classification of Data Mining

This stage consists of searching for patterns of interest that can be expressed as a model based on Machine learning algorithms applied to physical fitness tests in schoolchildren. Data analysis determined that the classification results from physical activity tests are labeled- Low, Normal, and High. Likewise, classification was the most appropriate type of prediction for this research. For the classification model, a comparison of supervised machine-learning techniques is made. The most used optimizers and algorithms, according to the literature, are:

TABLE V. DESCRIPTION OF CLASSIFICATION ALGORITHMS USED IN THE STUDY

Algorithm	Description	Advantages	Limitations
Decisión Tree	Separates the data of an ensemble into smaller subsets with an increase in the depth of the tree. The objective increases the prediction using decision nodes [18].	Visual interpretation of all possible results requires little data cleaning, is unaffected by different values, and uses numerical and categorical variables.	The calculation is complex, which implies more time is needed to train the model; a slight change in the data can cause a significant change in the tree's structure.
Random Forest	Models are made up of many decision trees; when training each tree, it learns from a random sample of the data points and a subset of features [19].	The final predictions of the Random Forest are made by averaging the predictions of each tree, reducing the problem of overfitting and variance.	Regression algorithms have a higher computational cost and longer training time than decision trees. They do not predict beyond the range in the training data.
Naive Bayes	Probabilistic learning algorithm that uses the rule of Bayes' theorem together with prior knowledge, whose characteristics depend on the independence provided by the class [20].	The advantage of Naive Bayes is when the cost of incorrectly classifying a result as positive is high, it is crucial to have high precision to minimize false results.	Naive Bayes is limited by underperforming when faced with splitting data sets with high dimensionality. As the number of features increases.
Support Vector Machine	SVMs find a line or hyperplane between different data and calculate a maximum margin that leads to a homogeneous division of all data points [21].	Efficient in memory usage when processing.	Choosing the correct kernel and parameters can be computationally expensive.
Logistic Regression	Used for binary classification problems, the basis of logistic regression is the logistic function (sigmoid) that takes any number with an accurate value and assigns it a value between 0 and 1 [22].	It is a simple but effective algorithm closely related to neural networks. The training time is less than that of other algorithms.	Predicting complex data is problematic because it has a linear decision surface; in high-dimensional data sets, this can generate overfitting.
Neural Network	It uses interconnected nodes with a layered structure that resembles the human brain. It creates an adaptive system that computers use to learn from their mistakes and continuously generate improvements [23].	Neural networks help computers make intelligent decisions with limited assistance. They can learn and model complex, non-linear input-output data relationships.	Its limitation is that it has a forward propagation network, which uses a feedback process to improve predictions over time.
k-Nearest Neighbors KN	It uses K-nearest neighbor to make classifications or predictions about the clustering of a data point. The main idea is that all data points are close to each other to belong to the same class [24].	Speeds training time by storing the training set and learning from it only when making predictions.	They are computationally expensive. They must be preprocessed and scaled, and the observations will be used only during prediction, so this step is costly.
XG Boost	It implements Gradient Boosting to maximize training speed and model performance [25].	Training optimizes memory resources and distributed computing, allowing for handling large data sets.	High flexibility produces many hyperparameters that strongly interact with the model's behavior.
Gradient Boos	With an ensemble algorithm with numerical optimization, the objective is to minimize the loss of the model by sequentially adding decision trees [26].	Good predictive accuracy, flexibility to adjust to different kinds of data, and predictions are made by a majority vote of weak learners.	Gradient boosting will continue to improve to minimize all errors that cause excessive overfitting.

Hyperparameters are defined as extra parameters or parameters that the learning algorithm does not memorize directly. Hyperparameters are external configuration variables that, when performing data analysis, help us manage the training of Machine Learning models. We also call model hyperparameters, which are manually configured before training a model. Optimization was used by configuring hyperparameters; among the main fields, the following were chosen:

Random Forest:

- n_estimators: Number of trees in the forest.
- max_depth: Maximum depth of each tree.
- min_samples_split: Minimum number of samples required to divide a node.
- min_samples_leaf: Minimum number of samples required in a leaf node.
- max_features: Number of features to consider for the best split.
- bootstrap: Whether to use bootstrap sampling when building trees.

Gradient Boosting:

- n_estimators: Number of impulse stages to perform.
- learning_rate: Learning rate.
- max_depth: Maximum depth of the trees.
- min_samples_split: Minimum number of samples required to divide a node.
- min_samples_leaf: Minimum number of samples required in a leaf node.
- max_features: Number of features to consider for the best split.
- subsample: The fraction of samples used to train each base tree.

IV. RESULTS

We worked with the Anaconda Navigator platform with Jupyter Notebook because it already had established, well-supported libraries. Table V shows the different machine-learning techniques applied in this research.

For the work of the data obtained already cleaned, it was used for shaping, 80% allocated for training, and 20% designated for testing, together with the Jupyter Notebook, a commonly used tool for machine learning, using the scikit-learning library.

The configuration used for the machine learning algorithms was made regarding the hyperparameters. The configuration achieved by the best results obtained in Random Forest and Gradient Boosting, respectively, is shown:

'n_estimators': 200,
'max_depth': 10,
'min_samples_split': 5,
'min_samples_leaf': 2,
'max_features': 'sqrt',
'bootstrap': False

'n_estimators': 100, 20,
'learning_rate': 0.1,
'max_depth': 3,
'min_samples_split': 5,
'min_samples_leaf': 1,
'max_features': 'sqrt',
'subsample': 1.0

Tables VI to VIII show the results obtained once processed. In male schoolchildren, they compare traditional and enhanced machine learning techniques with configured hyperparameters for flexibility, speed, and horizontal jump data. Accuracy metrics, such as F1 score, recall, and precision, are also shown for schoolchildren.

Tables IX to Table XI show the results of applying machine learning algorithms with hyperparameter optimization compared to traditional ones in female schoolchildren.

TABLE VI. COMPARISON OF RESULTS FOR THE MALE FLEXIBILITY TEST

Algorithm	Decision Tree	SVM	Random Forest	Naive Bayes	Logistic Regression	KNN	MLP	Gradient Boost	XGB	LGBM	CatBoost
Accuracy	0.92	0.79	0.94	0.85	0.86	0.73	0.81	0.96	0.96	0.98	0.95
Accuracy optimized Hyperparameter	0.95	0.89	0.97	0.85	0.91	0.82	0.83	0.95	0.89	0.98	0.95
F1-score	0.95	0.86	0.97	0.93	0.94	0.83	0.85	0.97	0.97	0.98	0.97
Recall	0.97	0.92	0.95	0.96	0.95	0.77	0.81	0.95	1.00	1.00	0.96
Precision	0.92	0.80	0.99	0.90	0.94	0.90	0.89	1.00	0.95	0.96	0.99

TABLE VII. COMPARISON OF RESULTS FOR THE MALE HORIZONTAL JUMP TEST

Algorithm	Decision Tree	SVM	Random Forest	Naive Bayes	Logistic Regression	KNN	MLP	Gradient Boost	XGB	LGBM	CatBoost
Accuracy	0.91	0.81	0.92	0.65	0.91	0.82	0.81	0.93	0.94	0.93	0.93
Accuracy optimized Hyperparameter	0.91	0.93	0.98	0.67	0.94	0.84	0.90	0.94	0.90	0.93	0.94
F1-score	0.94	0.82	0.94	0.71	0.93	0.85	0.83	0.96	0.96	0.96	0.95
Recall	0.94	0.78	0.93	0.70	0.96	0.83	0.80	0.97	0.95	0.95	0.98
Precision	0.94	0.87	0.95	0.73	0.91	0.88	0.86	0.95	0.98	0.98	0.93

TABLE VIII. COMPARISON OF RESULTS FOR THE MEN'S 20M SPEED TEST

Algorithm	Decision Tree	SVM	Random Forest	Naive Bayes	Logistic Regression	KNN	MLP	Gradient Boost	XGB	LGBM	CatBoost
Accuracy	0.85	0.54	0.81	0.61	0.77	0.48	0.45	0.91	0.90	0.93	0.90
Accuracy Optimized Hyperparameter	0.91	0.78	0.81	0.62	0.85	0.74	0.71	1.00	0.96	0.93	0.93
F1-score	0.88	0.62	0.81	0.70	0.85	0.60	0.61	0.90	0.88	0.88	0.91
Recall	0.92	0.56	0.76	0.64	0.88	0.56	0.68	0.88	0.93	0.93	0.88
Precision	0.83	0.70	0.87	0.76	0.83	0.64	0.56	0.93	0.85	0.85	0.95

TABLE IX. COMPARISON OF RESULTS FOR THE FEMALE FLEXIBILITY TEST

Algorithm	Decision Tree	SVM	Random Forest	Naive Bayes	Logistic Regression	KNN	MLP	Gradient Boost	XGB	LGBM	CatBoost
Accuracy	0.92	0.79	0.85	0.72	0.79	0.75	0.71	0.85	0.86	0.88	0.91
Accuracy Optimized Hyperparameter	0.91	0.88	0.91	0.75	0.85	0.78	0.72	0.82	0.85	0.89	0.85
F1-score	0.93	0.88	0.90	0.74	0.85	0.78	0.72	0.83	0.86	0.89	0.90
Recall	0.93	0.89	0.91	0.75	0.79	0.82	0.71	0.84	0.88	0.90	0.84
Precision	0.94	0.87	0.88	0.76	0.81	0.80	0.70	0.83	0.86	0.88	0.85

TABLE X. COMPARISON OF RESULTS FOR THE FEMALE HORIZONTAL JUMP TEST

Algorithm	Decision Tree	SVM	Random Forest	Naive Bayes	Logistic Regression	KNN	MLP	Gradient Boost	XGB	LGBM	CatBoost
Accuracy	0.90	0.83	0.87	0.67	0.81	0.74	0.71	0.84	0.87	0.88	0.90
Accuracy Optimized Hyperparameter	0.93	0.88	0.98	0.75	0.87	0.78	0.74	0.88	0.81	0.88	0.89
F1-score	0.94	0.89	0.91	0.75	0.89	0.78	0.82	0.87	0.92	0.90	0.93
Recall	0.93	0.90	0.94	0.85	0.82	0.73	0.86	0.86	0.89	0.91	0.94
Precision	0.94	0.88	0.88	0.67	0.70	0.85	0.78	0.89	0.94	0.89	0.92

TABLE XI. COMPARISON OF RESULTS FOR THE WOMEN'S 20M SPEED JUMP TEST

Algorithm	Decision Tree	SVM	Random Forest	Naive Bayes	Logistic Regression	KNN	MLP	Gradient Boost	XGB	LGBM	CatBoost
Accuracy	0.87	0.46	0.89	0.73	0.83	0.61	0.53	0.89	0.93	0.94	0.89
Accuracy Optimized Hyperparameter	0.89	0.82	0.97	0.85	0.84	0.77	0.79	0.90	0.89	0.94	0.89
F1-score	0.88	0.46	0.91	0.74	0.86	0.65	0.62	0.90	0.93	0.94	0.89
Recall	0.90	0.44	0.88	0.69	0.89	0.58	0.53	0.87	0.97	0.99	0.85
Precision	0.87	0.49	0.94	0.79	0.83	0.73	0.75	0.93	0.89	0.89	0.93

The results shown in the Tables show that, for the classification techniques, the Random Forest achieved the highest precision in Male Flexibility, whose value is 0.97, indicating a more significant adjustment of the estimated prediction. For the f1-score metrics, the result was 0.97, in the case of recall 0.95.

The choice of hyperparameters has a critical impact on the performance of machine learning models. In this case, the optimized parameters allowed Random Forest to achieve outstanding results in classifying flexibility data in male schoolchildren. The precision, F1-Score, and recall metrics indicate a well-fitted model with adequate generalization capacity.

The Random Forest algorithm obtained the best results for most of the tests. The following Fig. 3 to Fig. 5 show the ROC-AUC curves generated by the Random Forest algorithm for the Physical Fitness tests evaluated on male schoolchildren.

Fig. 3 shows the ROC-AUC curve for the flexibility test where the High class (AUC = 0.88): The model distinguishes between the "High" category and the other categories well. Low (AUC = 0.84), the model also distinguishes between the “Low” category and the different categories. Normal (AUC = 0.58), the model performs significantly lower in distinguishing between the “Normal” category and the others, indicating that the model is ineffective in this classification.

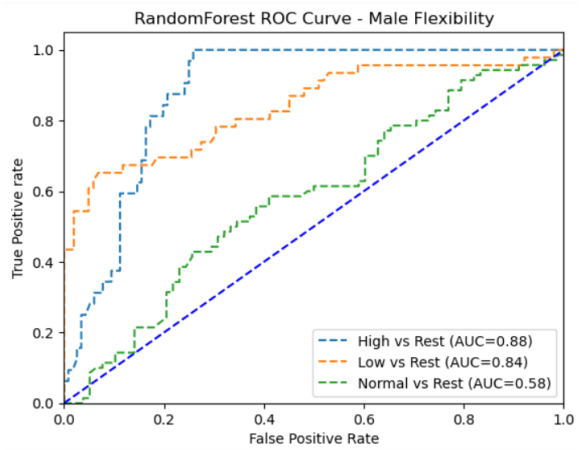

Fig. 3. Random forest ROC-AUC curve for the male flexibility test.

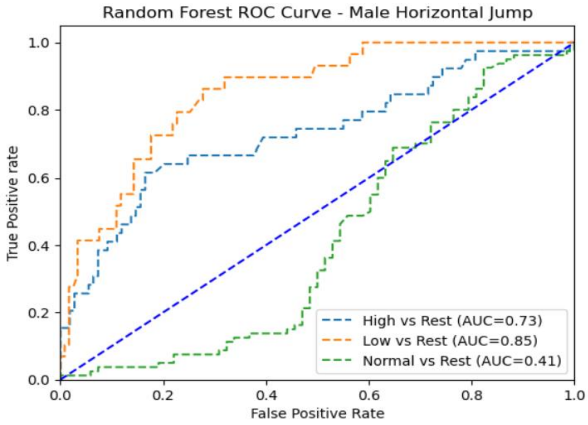

Fig. 4. Random forest ROC-AUC curve for the male horizontal jump test.

Fig. 4 shows the ROC-AUC curve for the horizontal jump test. The "High" class has an AUC of 0.73, indicating that the model performs moderately in distinguishing between this category and the others. The "Low" class has an AUC of 0.85, demonstrating the model's good performance in differentiating this category from the rest. In contrast, the "Normal" class has an AUC of 0.41, reflecting a relatively poor performance of the model in distinguishing between this category and the others.

Fig. 5 shows the ROC-AUC curve for the speed 20m test. For the High class (AUC = 0.83), the model performs well in distinguishing this category from the others. The model performs moderately for the Low class (AUC = 0.74). However, the model's performance is low for the Normal class (AUC = 0.53), indicating that it is ineffective for this classification.

In summary, from the tests of male schoolchildren, the model performs well for the "High" categories in all tests, with AUCs of 0.88, 0.73, and 0.83, respectively. The model has excellent performance for the "Low" category in flexibility and horizontal jump, with AUCs of 0.84 and 0.85, respectively, and moderate performance in speed with an AUC of 0.74. The model shows overall poor performance for the "Normal" category in all tests, with AUCs of 0.58, 0.41, and 0.53, respectively, indicating that the model's predictive ability is limited.

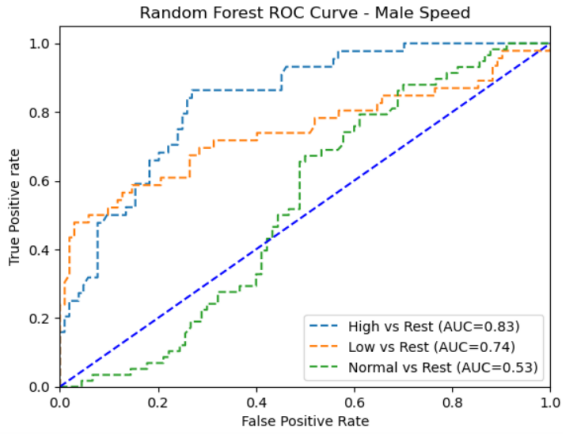

Fig. 5. Random forest ROC-AUC curve for the men's 20m speed test.

In conclusion, the Random Forest model performs well in classifying the “High” and “Low” categories in most tests but struggles to classify the “Normal” category correctly.

Fig. 6 to Fig. 8 show the ROC-AUC curves generated by the Random Forest algorithm for the Physical Fitness tests evaluated on female schoolchildren.

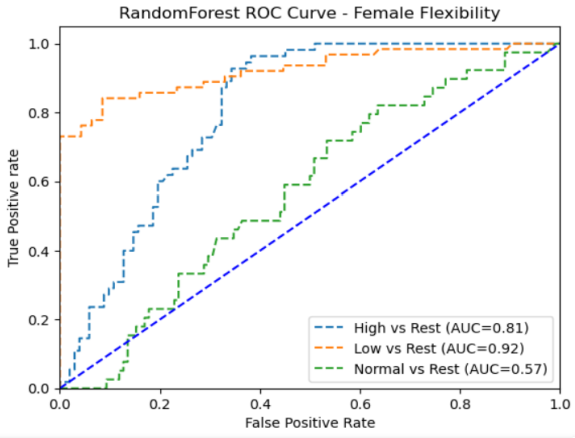

Fig. 6. Random forest ROC-AUC curve for the female flexibility test.

Fig. 6 shows the ROC-AUC curve for the flexibility test in female schoolchildren. The "High" class has an AUC of 0.81, indicating that the model performs well in distinguishing this category from the others. The "Low" class has an AUC of 0.92, demonstrating excellent model performance for this category. On the other hand, the "Normal" class has an AUC of 0.57, suggesting that the model's performance is moderately poor in distinguishing this category from the others.

The ROC-AUC curve for the "horizontal jump" test is shown in Fig. 7. The "High" class has an AUC of 0.71, indicating that the model has moderate performance in distinguishing between the "High" category and the others. The "Low" class has an AUC of 0.73 and moderately performs in distinguishing between the "Low" category and the others. However, for the "Normal" class, the AUC is 0.39, suggesting that the model performs poorly in distinguishing between the "Normal" category and the others, indicating that it is ineffective in this classification.

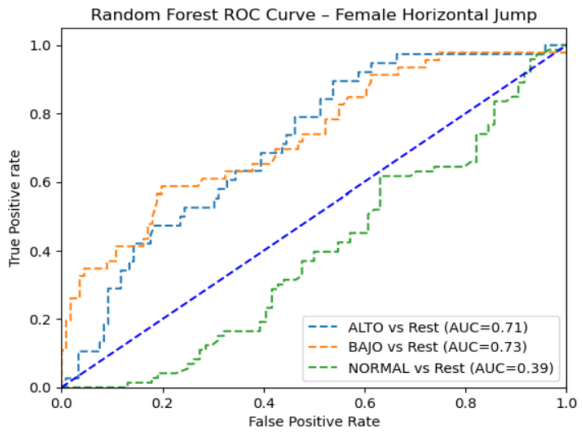

Fig. 7. Random forest ROC-AUC curve for the female horizontal jump test.

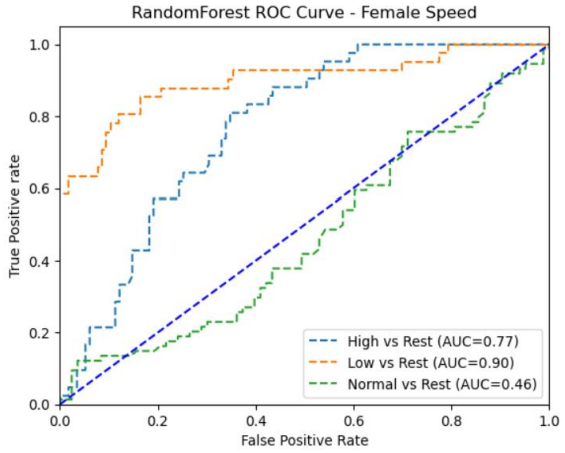

Fig. 8. Random forest ROC-AUC curve for the women's 20m speed test.

Fig. 8 shows the ROC-AUC curve for the Speed 20m High test (AUC = 0.77). The model distinguishes between the "High" category and the other categories. Low (AUC = 0.90): The model distinguishes between the “Low” category and the others. Normal (AUC = 0.46): The model performs poorly distinguishing between the “Normal” category and the others.

In summary, from the results of the tests of female schoolchildren, the model generally shows good performance for the "High" categories in all tests, with AUCs of 0.81, 0.71, and 0.77, respectively. The model has excellent performance for the "Low" category in flexibility and speed, with AUCs of 0.92 and 0.90, respectively, and moderate performance in horizontal jump with an AUC of 0.73. The model shows overall poor performance for the "Normal" category in all tests, with AUCs of 0.57, 0.39, and 0.46, respectively, indicating the model's limited predictive ability.

In conclusion, the Random Forest model performs well in classifying the “High” and “Low” categories in most tests but struggles to classify the “Normal” category correctly.

V. Discussion

This research used machine learning techniques to analyze the accuracy and effectiveness in classifying data related to male and female schoolchildren's flexibility, horizontal jump, and 20m speed. Algorithms evaluated included Decision Trees, Random Forest, Support Vector Machine (SVM), Naive Bayes, Logistic Regression, K-Nearest Neighbor (KNN), Multi-Layer Perceptron (MLP), Gradient Boosting, XGBoost, LightGBM and CatBoost [27]. The data was divided into 80% for training and 20% for testing. Hyperparameter optimization techniques were used to improve the accuracy of the models, as seen in Tables VI to XI. For example, the hyperparameters configured for Gradient Boost included criteria such as 'entropy,' 'max_depth,' and 'n_estimators,' which were manually tuned before model training.

The comparative results of the traditional and enhanced techniques with hyperparameters showed significant variations in accuracy, F1 score, recall, and precision for male and female schoolchildren. For the male group, the optimized Random Forest model showed outstanding accuracy in classifying flexibility, with an accuracy of 0.97, an F1-score of 0.97, and a recall of 0.95. This suggests the model can correctly predict physical activity classes with a low false positive rate. In contrast, the analysis of the female group revealed that the Random Forest model was also highly effective in classifying flexibility with an AUC of 0.92, indicating high sensitivity and specificity.

The ROC-AUC curves presented in Fig. 3 to Fig. 8 show the effectiveness of Random Forest models in classifying physical activities in schoolchildren, like studies of children aged 6 to 12 years [28]. Specifically, for male flexibility, the ROC-AUC curve showed a significant increase towards the upper left corner with an AUC of 0.88, confirming the high sensitivity of the model to detect the 'High' classification of physical activity. For female flexibility, the AUC was 0.92 for the 'Low' class, indicating high sensitivity and a low false positive rate.

Comparison tables showed that hyperparameter-optimized models significantly improved accuracy and other critical metrics compared to their non-optimized versions. For example, the SVM increased accuracy from 0.46 to 0.82 for the female group in the horizontal jump test. Similarly, the Gradient Boost model showed substantial improvements in accuracy and F1-score across multiple tests.

Comparing our results with those of other recent studies, we observed a congruence in the effectiveness of the Random Forest [29] and Gradient Boosting [30] algorithms. In our research and previous studies, these algorithms have repeatedly demonstrated their superiority in precision and recall in physical activity classification. This suggests that using advanced hyperparameter optimization techniques can significantly improve the accuracy of machine-learning models.

The consistency in results across multiple studies reinforces the validity of our findings and underscores the importance of selecting and optimizing appropriate algorithms for specific data classification tasks.

Limitations in the quality and quantity of data available for training machine learning models can significantly affect their performance. The challenge will be obtaining a large and representative data set, which can be difficult, especially in specific studies such as classifying physical activities in

schoolchildren. Another limitation is that incorrect selection of relevant features can decrease the effectiveness of the models. Determining which features are most informative requires deep domain knowledge and sometimes a thorough process of trial and error. Although hyperparameter optimization can significantly improve model performance, it is a resource-intensive process that requires time and computational power. The challenge is identifying the optimal combination of hyperparameters for each algorithm, which can be complicated and requires advanced search and cross-validation techniques.

VI. CONCLUSION

It is demonstrated that the application of machine learning algorithms, together with the optimization of hyperparameters, is an effective strategy for classifying students' physical condition in educational centers. Using the Knowledge Discovery in Databases (KDD) process and collecting anthropometric data and physical fitness tests could accurately assess the physical fitness of a representative sample of students.

The results indicated that the Random Forest and Gradient Boosting algorithms were particularly effective in classifying physical activities with high levels of precision, F1 score, recall, and specificity. These models' ability to differentiate between levels of physical fitness (below average, average, and above average) with a low false positive rate suggests their practical applicability in monitoring and evaluating students' physical health.

Future work could explore integrating these models into educational and health platforms and evaluating their long-term impact on student health. Other parameters and physical tests could also be considered for a more complete evaluation. The continued evolution of machine learning techniques and the availability of more granular data promise to further improve the accuracy and usefulness of these models.

ACKNOWLEDGMENT

To the 'Universidad Nacional de San Agustín de Arequipa', who has financed the project «Propuesta Normativa para valorar los niveles de actividad física de los escolares de la provincia de Arequipa», with contract number 15-2016-UNSA.

REFERENCES

[1] Kufel, J., Bargieł-Łączek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., ... & Gruszczyńska, K. (2023). What is machine learning, artificial neural networks and deep learning?—Examples of practical applications in medicine. Diagnostics, 13(15), 2582.

[2] Aized Amin Soofi and Arshad Awan, "Classification Techniques in Machine Learning: Applications and Issues," Journal of Basic & Applied Sciences, vol. 13, pp. 459–465, Jan. 2017, doi: 10.6000/1927-5129.2017.13.76.

[3] J. Sulla-Torres et al., "Quantification of the Number of Steps in a School Recess by Means of Smart Bands: Proposal of Referential Values for Children and Adolescents," Children, vol. 10, no. 6, p. 915, May 2023, doi: 10.3390/children10060915.

[4] WHO, "Physical activity," WHO. [Online]. Available: https://www.who.int/es/news-room/fact-sheets/detail/physical-activity

[5] H. N. C. Betancur, L. G. P. Canqui, Y. Y. R. Yapuchura, K. Pérez, S. Chura, and W. W. C. Castillo, "Obesidad infantil en estudiantes de educación primaria en Puno, Perú," Retos: nuevas tendencias en educación física, deporte y recreación, no. 54, pp. 466–477, 2024.

[6] "América Latina y el Caribe: Más de 4 millones de niños y niñas menores de 5 tienen sobrepeso." Accessed: Jun. 19, 2024. [Online]. Available: https://www.unicef.org/lac/comunicados-prensa/america-latina-caribe-mas-4-millones-ninos-ninas-menores-5-sobrepeso

[7] S. Andermo et al., "School-related physical activity interventions and mental health among children: a systematic review and meta-analysis," Sports Med Open, vol. 6, no. 1, p. 25, Dec. 2020, doi: 10.1186/s40798-020-00254-x.

[8] P. Trejo Ortiz, S. Jasso Chairez, F. Mollinedo Montaño, and L. Lugo Balderas, "Relación entre actividad física y obesidad en escolares," Revista Cubana de Medicina General Integral, vol. 28, no. 1, 2012.

[9] A. Rosa Guillamón, E. García-Cantó, P. L. Rodríguez García, J. J. Pérez Soto, M. L. Tárraga Marcos, and P. J. Tárraga López, "Physical activity, physical fitness and quality of diet in schoolchildren from 8 to 12 years," Nutr Hosp, vol. 34, no. 6, 2017.

[10] M. Zhou, Y. Fukuoka, K. Goldberg, E. Vittinghoff, and A. Aswani, "Applying machine learning to predict future adherence to physical activity programs," BMC Med Inform Decis Mak, vol. 19, no. 1, p. 169, Dec. 2019, doi: 10.1186/s12911-019-0890-0.

[11] S. A. Alsareii et al., "Physical Activity Monitoring and Classification Using Machine Learning Techniques," Life, vol. 12, no. 8, p. 1103, Jul. 2022, doi: 10.3390/life12081103.

[12] M. N. Ahmadi, T. G. Pavey, and S. G. Trost, "Machine Learning Models for Classifying Physical Activity in Free-Living Preschool Children," Sensors, vol. 20, no. 16, p. 4364, Aug. 2020, doi: 10.3390/s20164364.

[13] J. Sulla-Torres, A. Calla Gamboa, C. Avendaño Llanque, J. Angulo Osorio, and M. Zúñiga Carnero, "Classification of Motor Competence in Schoolchildren Using Wearable Technology and Machine Learning with Hyperparameter Optimization," Applied Sciences, vol. 14, no. 2, p. 707, Jan. 2024, doi: 10.3390/app14020707.

[14] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AI Mag, vol. 17, no. 3, 1996.

[15] J. R. Fernández, D. T. Redden, A. Pietrobelli, and D. B. Allison, "Waist circumference percentiles in nationally representative samples of African-American, European-American, and Mexican-American children and adolescents," Journal of Pediatrics, vol. 145, no. 4, 2004, doi: 10.1016/j.jpeds.2004.06.044.

[16] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, 2011.

[17] J. Sulla-Torres, G. Luna-Luza, D. Ccama-Yana, J. Gallegos-Valdivia, and M. Cossio-Bolaños, "Neuro-fuzzy System with Particle Swarm Optimization for Classification of Physical Fitness in School Children," International Journal of Advanced Computer Science and Applications, vol. 11, no. 6, 2020, doi: 10.14569/IJACSA.2020.0110663.

[18] E. Engür and B. Soylu, "A linear multivariate decision tree with branch-and-bound components," Neurocomputing, vol. 576, 2024, doi: 10.1016/j.neucom.2024.127354.

[19] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," Stata Journal, vol. 20, no. 1, 2020, doi: 10.1177/1536867X20909688.

[20] H. Kamel, D. Abdulah and J. M. Al-Tuwaijari, "Cancer Classification Using Gaussian Naive Bayes Algorithm," *2019 International Engineering Conference (IEC)*, Erbil, Iraq, 2019, pp. 165-170, doi: 10.1109/IEC47844.2019.8950650.

[21] JavaTpoint, "Support Vector Machine Algorithm," JavaTpoint, 2021.

[22] C. El Morr, M. Jammal, H. Ali-Hassan, and W. El-Hallak, "Logistic Regression," in International Series in Operations Research and Management Science, vol. 334, 2022. doi: 10.1007/978-3-031-16990-8_7.

[23] Y. chen Wu and J. wen Feng, "Development and Application of Artificial Neural Network," Wirel Pers Commun, vol. 102, no. 2, 2018, doi: 10.1007/s11277-017-5224-x.

[24] P. Cunningham and S. J. Delany, "K-Nearest Neighbour Classifiers-A Tutorial," ACM Computing Surveys, vol. 54, no. 6. 2021. doi: 10.1145/3459665.

[25] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. doi: 10.1145/2939672.2939785.

[26] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," Front Neurorobot, vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.

[27] C. Milanese, M. Sandri, V. Cavedon, and C. Zancanaro, "The role of age, sex, anthropometry, and body composition as determinants of physical fitness in nonobese children aged 6-12," PeerJ, vol. 2020, no. 3, 2020, doi: 10.7717/peerj.8657.

[28] S. R. Shakya, C. Zhang, and Z. Zhou, "Comparative study of machine learning and deep learning architecture for human activity recognition using accelerometer data," Int J Mach Learn Comput, vol. 8, no. 6, 2018, doi: 10.18178/ijmlc.2018.8.6.748.

[29] P. Probst, M. N. Wright, and A.-L. Boulesteix, 'Hyperparameters and tuning strategies for random forest', Wiley Interdisciplinary Reviews: data mining and knowledge discovery, vol. 9, no. 3, p. e1301, 2019.

[30] J. Guo et al., "An XGBoost-based physical fitness evaluation model using advanced feature selection and Bayesian hyper-parameter optimization for wearable running monitoring," Computer Networks, vol. 151, 2019, doi: 10.1016/j.comnet.2019.01.026.