Shaochen Z & Jiaxing Y. (2025) ATHLETE HEALTH DATA MANAGEMENT FROM THE PERSPECTIVE OF PRIVACY PROTECTION. Revista Internacional de Medicina y Ciencias de la Actividad Física y el Deporte vol. 25 (100) pp. 35-53.
DOI: https://doi.org/10.15366/rimcafd2025.100.003

ORIGINAL

ATHLETE HEALTH DATA MANAGEMENT FROM THE PERSPECTIVE OF PRIVACY PROTECTION

Zhou Shaochen[1*], Yao Jiaxing [2]

[1] School of Physical Education Shanxi University, Taiyuan, Shanxi, 030006, China
[2] Yangquan Vocational and Technical College, Yangquan, Shanxi, 045000, China
E-mail: zsc19920427@163.com

Recibido 03 de junio de 2024 **Received** June 03, 2024
Aceptado 03 de diciembre de **2024 Accepted** December 03, 2024

ABSTRACT

In response to the problems of insufficient privacy protection and model performance in traditional athlete health data analysis models, this paper explored distributed model training methods based on the framework of federated learning. The paper first divided athlete data into time segmentation, body index segmentation, sports item segmentation, and environmental condition segmentation, and used transport layer security protocols and homomorphic encryption to protect data computation. When training the local model, a lightweight decision tree was selected for training; the dynamic weighted learning was used to aggregate the model; finally, differential privacy technology was applied to protect data privacy by adding Gaussian noise, and some optimization methods were used to improve model performance. In the third experiment of model performance, the precision of the model in this paper reached 98.45%. This indicated that the model had extremely high accuracy and reliability in classification and regression tasks. In the speedup ratio experiment, when the synchronization interval was 50 and the number of clients was 200, the speedup ratio of the model in this paper was 4.01, reflecting that the model can effectively improve training efficiency with the participation of multiple clients. In the privacy leakage risk test, the success rate of the model in this paper was the lowest under attack, at 1.5%, 2.3%, and 1.4%, respectively. Finally, in the model loss test, the model in this paper experienced the fastest decline in the initial stage, with a final convergence value of 0.3, which was the smallest among the tested models. The data showed that the model studied in this paper had good performance and privacy protection ability.

KEYWORDS: Federated Learning, Data Privacy Protection, Decision Tree, Model Aggregation, Differential Privacy Technology.

1. INTRODUCTION

Today, the collection, storage, and analysis of athlete health data are very important. It is an integral part of improving athlete competitiveness and maintaining health. The era of big data has helped sports professionals effectively analyze athlete health data, but it has also brought about new problems: multiple data sources lead to model performance degradation; traditional centralized models lack personal privacy protection for athlete health data; data is prone to theft and tampering. These issues constrain the performance and security of data models, thereby affecting the training effectiveness and competitive performance of athletes (Sun et al., 2023). In order to protect data privacy during model training, scholars have conducted extensive research. The content of privacy protection has been extensively discussed in various models of machine learning (Kaissis et al., 2020; So et al., 2021; Tan & Zhang, 2020). Niu C et al. used decomposition and specific mathematical methods to make predictions without revealing key parameters (Niu et al., 2020). They extensively evaluated the performance and low cost by using support vector machines on the short message service dataset. Kawamura A et al. used encrypted-compressed images to protect privacy, and verified the effectiveness of this scheme in the face recognition experiment (Kawamura et al., 2020). Zhu L et al. utilized partial homomorphic encryption to implement multiple privacy preserving training protocols in aggregated scenarios, capable of dealing with collusion threats (Zhu et al., 2021). Strict security analysis and experimental verification have demonstrated its effectiveness and privacy preserving ability. Li X et al. proposed a prediction scheme for edge enhanced human physical systems, and achieved good results in security and privacy protection (X. Li et al., 2021). Arachchige P C M et al. applied PriMod Chain, which combined multiple techniques to protect data (Arachchige et al., 2020). Gupta R et al. proposed a model for classifying services in a cloud environment (Gupta & Singh, 2022).

The privacy of data and classifiers was protected through communication protocols. Experimental results have shown that this model has high accuracy and privacy protection ability on large-scale datasets. Li T et al. proposed a server-assisted framework that supports learning tasks without the involvement of data owners and significantly reduces communication overhead (Li et al., 2020). The above research includes the exploration of technologies such as multi-party machine learning and differential privacy, with the aim of solving the problem of protecting data privacy in distributed environments. However, there is a lack of research on privacy protection of athlete data, and the security issue of model training by different data holders without sharing data has not been addressed. Federal learning enables multiple organizations to use data in a manner that protects user privacy, ensures data security, and complies with regulations. Federated learning conducts model training in a distributed data environment with multiple participants, while protecting data privacy and security (Khan et al., 2021; Liu et al., 2022; Nguyen et al., 2021). Nguyen H T et al. proposed a fast convergent federated learning algorithm that optimizes the convergence speed and stability of model training through intelligent sampling and weight update device updates, in order to reduce communication and computational costs (Nguyen et al., 2020). Zhang T et al. discussed the opportunities and challenges of federated learning in IoT platforms and proposed methods to address these challenges, providing useful guidance for implementing various IoT applications (Zhang et al., 2022). Yu R et al. proposed a federal learning management-based approach to realize the flexible and efficient use of resources (Yu & Li, 2021). Gafni T et al. pointed out the importance of edge device data privacy protection and federated learning, and applied specialized solutions in signal processing and communication to address the challenges in federated learning (Gafni et al., 2022). Rieke N et al. considered the issues of medical data privacy and data silos, and explored potential solutions for federated learning in the future of digital health (Rieke et al., 2020). Pfitzner B et al. explored in depth the applicability of federated learning to confidential medical datasets (Pfitzner et al., 2021). Zhang Y et al. explored existing federated learning methods and proposed future directions (Zhan et al., 2021). These studies explored various directions of federated learning in model training, providing useful ideas and methods for solving distributed model training methods that protect the privacy of athlete health data. This paper improved the privacy protection and performance of traditional athlete data models through a federated learning framework. Firstly, the data was segmented and privacy was protected using security protocols and homomorphic encryption. A lightweight decision tree was used to train the local model, while dynamic weighted learning was used for model aggregation and privacy protection was enhanced by adding Gaussian noise.

2. Privacy Protection Distributed Model Training Method

2.1. Data Segmentation and Encryption

The health data of athletes contains numerous categories, and the model processing is cumbersome, resulting in a long processing time. Therefore, data segmentation is required when processing athlete health data. Data segmentation refers to dividing logically unified data into separately managed physical units for storage, increasing the efficiency of the model in data indexing and scanning. Figure 1 shows the segmentation types for athlete health data in this paper:

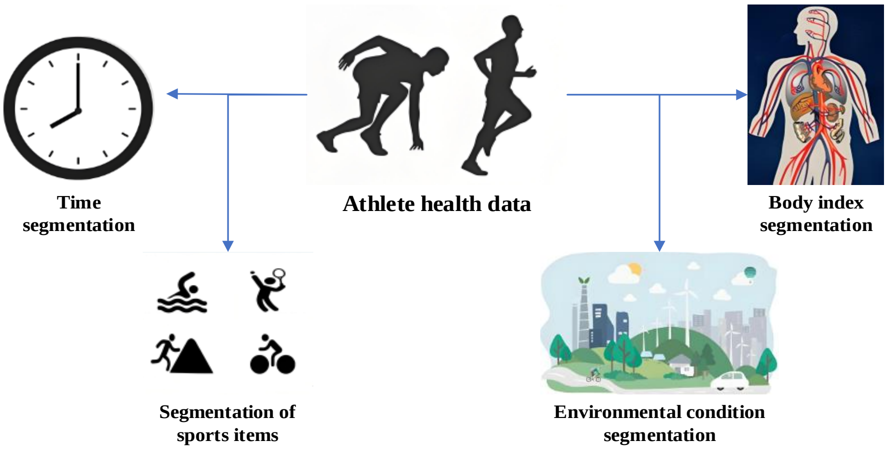

Time segmentation
Athlete health data
Body index segmentation
Segmentation of sports items
Environmental condition segmentation

Figure 1: Data Segmentation Types

Figure 1 shows the data segmentation types used in this paper, including time segmentation, body index segmentation, sports item segmentation, and environmental condition segmentation. Time segmentation divides the health data of athletes into time periods, dividing them by day, week, or month according to different needs. Time segmentation is suitable for tracking the long-term health data of athletes, discovering their periodic changes in the body, and helping athletes develop long-term training or rehabilitation plans. Body index segmentation is commonly used in medical analysis. These indicators provide personalized health management and rehabilitation training directions for athletes. Sports item segmentation is the process of categorizing data according to different sports, which is divided into football, basketball, badminton, short distance running, and long-distance running in this paper. Different sports require the same type of data to be collected, but there are also some differences. In terms of heart rate, in football, heart rate exhibits periodic fluctuations because running, walking, and standing occur alternately during matches, and the game center rate is influenced by the game situation and personal mood. In sprint events, the heart rate exhibits sustained high intensity. By segmenting data based on sports events, it is possible to more precisely analyze the specific impact of each sport on the physical and skill development of athletes, and conduct targeted training and adjustments. Environmental condition segmentation classifies data according to different sports environments, and in this paper, it is divided into indoor, outdoor, sunny, rainy, and altitude. These environmental conditions affect factors such as temperature, humidity, wind speed, and light, and can be used to study the physical reactions and adaptation mechanisms of athletes in different environments. Health data has high sensitivity and privacy, especially for athletes. If health data of other athletes can be obtained, competitors may obtain their physical condition and training effectiveness, which seriously affects the fairness of competitive sports. The encryption of data during transmission and computation needs to be considered. When transmitting data, this paper chooses Transport Layer Security (TLS) protocol (Akbar & Iqbal, 2022) as the encryption technology during transmission. TLS protocol establishes a secure connection by using public key encryption algorithms to ensure data is protected during transmission. When performing calculations, this paper uses homomorphic encryption (Wang et al., 2020; Yan et al., 2020) technology for encryption, which provides higher protection for the privacy of athlete health data. Through homomorphic encryption, encrypted data can be computed without exposing plaintext data.

2.2. Construction of Federated Learning Models

The federated learning model makes the health data of each athlete need not leave the local, only need to pass the specific federated training algorithm and parameter exchange mechanism, and finally build a global sharing model. Federated learning can be divided into three types: horizontal federated learning, vertical federated learning, and federated transfer learning (ZHOU et al., 2021). Horizontal federated learning makes judgments based on features and is more suitable when the sample feature repetition rate is high and the sample repetition rate is low; vertical federated learning, on the other hand, is more suitable when the sample feature repetition rate is low and the sample repetition rate is high; federated transfer learning is different from the above two methods and is more suitable when the sample feature repetition rate and sample repetition rate are low. This paper chooses to use the method of horizontal federated learning (Hammoud et al., 2022). Because the health data of athletes has highly repetitive features, such as heart rate, step count, body temperature, etc., the repetition rate of sample features is high. However, when it comes to data from different athletes, each athlete is a unique individual, so the sample repetition rate is very low. The optimization objective of federated learning is similar to other machine learning algorithms:

$$F_x = \frac{1}{n}\sum_{i=1}^{n} f(S_i) \tag{1}$$

Among them, F_x represents the loss function of the federated learning model; n is the number of samples; S_i represents the i-th sample individual; $f(S_i)$ represents the loss function of the model on S_i. Figure 2 shows the training process of the federated learning model:

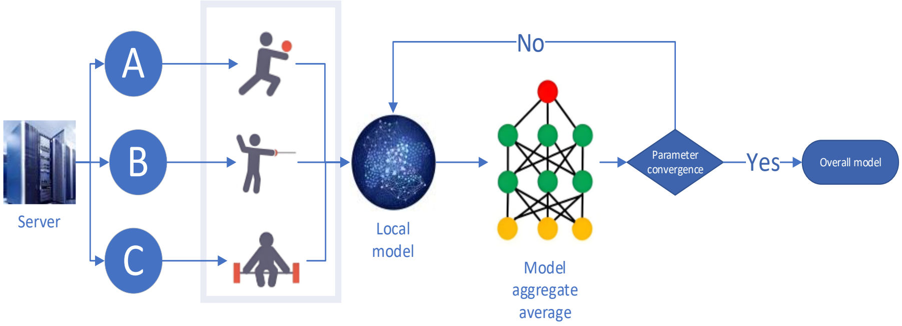


Figure 2: Training Process of Federated Learning Model

Figure 2 shows the training process of the federated learning model. The training participants of different sports projects obtain their respective global models from the same server, namely A, B, and C in the figure, which have the same training objectives and standards. Subsequently, each training participant uses their local data for model training. The server aggregates and averages the trained models, and verifies the parameters of the aggregated models. If the validation parameters do not converge, they are returned to their respective local models for further training, and the models are re aggregated. If the validation parameters converge, the overall model construction is complete.

2.3. Local Model Training

Preprocessing is required before training local data, which includes handling missing and outliers in the data, as well as standardizing the data. This paper chooses the linear interpolation method (Fenglei et al., 2020) in interpolation to fill in missing values. Among them, linear interpolation only requires constructing a straight line through the coordinates of two data points, and an unknown point between these two data points can be estimated through linear interpolation. If the known data points are (x_1, y_1) and (x_2, y_2), then the linear interpolation formula is:

$$y_0 = y_1 + \frac{(x_0 - x_1)}{(x_2 - x_1)} \times (y_2 - y_1) \tag{2}$$

According to the above formula, (x_0, y_0) is the data point that fills the gap position with linear interpolation. The method for handling outliers in this paper uses the box plot method (Aghighi et al., 2022). Figure 3 is an example of a box plot.

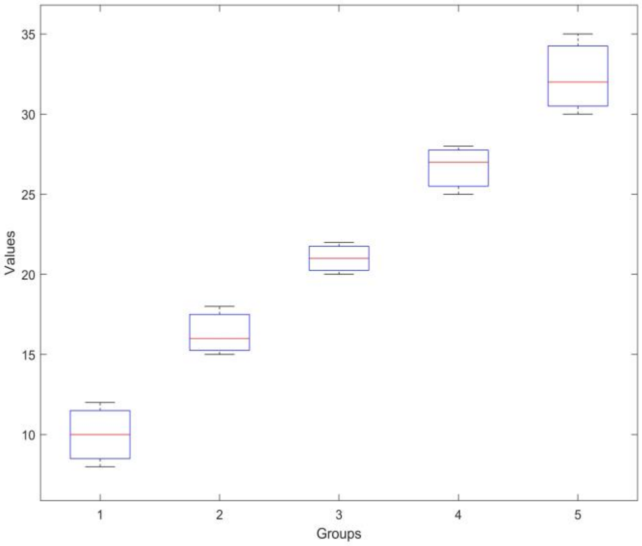


Figure 3: Example of Box Line Diagram

In Figure 3, this paper selects five sets of data (Aghighi et al., 2022; Akbar & Iqbal, 2022; Alrmali et al., 2023; Arachchige et al., 2020; Hammoud et al., 2022; Li et al., 2020; Liu et al., 2022; Rieke et al., 2020; Wang et al., 2020; Yu & Li, 2021; Zhan et al., 2021; Zhang et al., 2022), and (Chen et al., 2021; Gao et al., 2021) to draw box plots. The length of the box represents the range of variation for each set of data, and it can be seen that the box length of the fifth set of data (Chen et al., 2021; Gao et al., 2021) is significantly higher than that of the third set of data (Akbar & Iqbal, 2022; Wang et al., 2020; Zhan et al., 2021). The horizontal line in the box represents the median in each set of data, and it can be seen that the median position in the first set of data is 10, corresponding to the median 10 in the first set of data. The line segments above and below the box are called tentacles, which are often determined by multiplying the quartile distance by a constant. When there is data outside the range of the tentacles, it is considered an outlier and removed or corrected. This paper uses the method of range standardization for data standardization. Scope standardization adjusts athlete health data to a specific range. If athlete heart rate changes are standardized, range standardization can be used to reduce the data to a reasonable heart rate range, such as [60, 200]. The calculation formula for range standardization is:

$$\ddot{x} = a + \frac{b-a}{x_{max}-x_{min}} \times (x - x_{min}) \tag{3}$$

Among them, $\ddot{x}$ is the data that has been standardized and processed; [a, b] is the standardized range that needs to be scaled down; x_{max} and x_{min} are the maximum and minimum values in the original dataset. The advantage of range standardization is that it is simple and easy to understand, but it is highly sensitive to data. Whenever there are outliers, they stretch the maximum and minimum values, leading to distorted distribution of subsequent data. Therefore, before using range standardization, it is important to handle the issue of outliers. When training local models, the computing power of each participating device may vary, and some devices may have limitations in computing power. Therefore, it is necessary to use lightweight machine learning models as much as possible. Taking all factors into consideration, this paper chooses to use the decision tree (Alrmali et al., 2023) model for local model training. The training process of decision trees is simple, with features compared and information gain calculated each time the nodes are split. The cost is low, and it has high interpretability and strong resistance to overfitting. Therefore, it is very suitable for local model training of athlete health data.

2.4. Model Aggregation

In the process of federated learning, model aggregation involves updating the models of the training participants onto the overall model. However, each athlete’s health data has different characteristics and cannot be measured using a unified weight. Therefore, this paper uses the method of dynamic weighted learning (Jing et al., 2023) to weight the datasets of each training participant. This method includes a local update module and an overall update module, and its framework is shown in Figure 4:

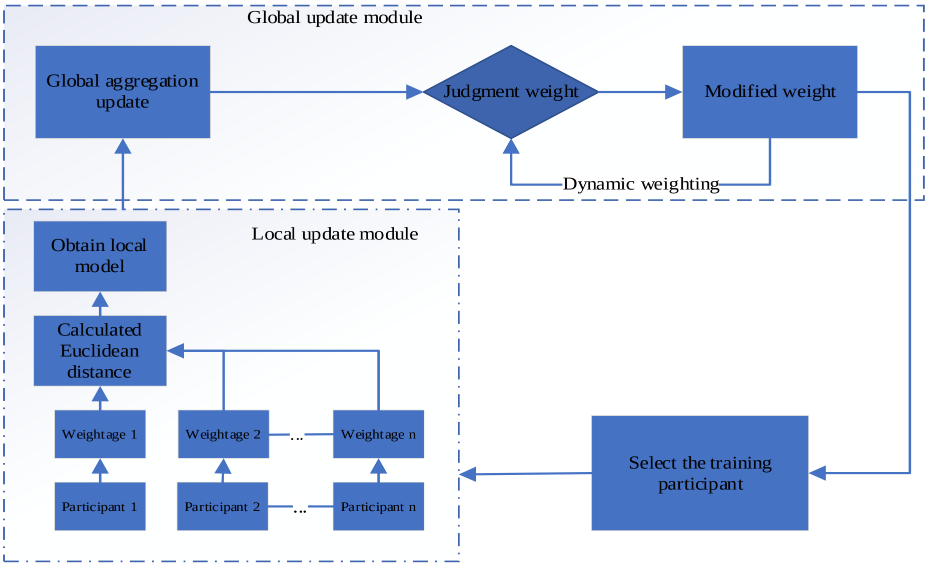


Figure 4: Framework of Dynamic Weighted Learning

As shown in Figure 4, in the local update module, when the model completes one round of training, the weight of the previous round is retained, and the weight of the new training participants is updated. The Euclidean distance between the weight of this round and the weight of the previous round is calculated, and the degree of offset of each local model is judged to obtain a new local model, which is then comprehensively aggregated. In the overall update module, the overall model is compared in weight for each round, and weights with smaller loss functions are selected for modification. After such dynamic weighting adjustments, the overall model of athlete health data is continuously optimized. The Euclidean distance is calculated. Firstly, the mean $\bar{y}$ and standard deviation s of the data samples from each participant are obtained. Standardization is utilized to convert the components of each dimension into a unified standard, namely:

$$\dot{Y} = \frac{Y-\bar{y}}{s} \tag{4}$$

Among them, Y is the original set of data samples, and $\dot{Y}$ is the standardized set of data samples. The Euclidean distance D between the weights of each sample is calculated after standardization, which is:

$$D = \sqrt{\sum_{n=1}^{n}\left(\frac{\omega_n-\omega_{n-1}}{s_n}\right)} \tag{5}$$

Among them, ω_n is the current weight; ω_{n-1} represents the weight of the previous round; s_n is the variance of the current data sample. When analyzing athlete health data, there may be some extreme situations: the sample size of badminton athletes accounts for 90% of the total sample, while the sample size of long-distance runners only accounts for 10%. At this point, the overall model is being optimized towards the direction of badminton player data during the update process, which may overlook some features of long-distance runners. Therefore, it is necessary to maintain dynamic weighting in each training cycle. When selecting weights, it is necessary to minimize their loss function values, namely:

$$F_{min}(\omega) = \sum_{i=1}^{i} P_i \times f_i(\omega) \tag{6}$$

Among them, $F_{min}(\omega)$ is the minimum value of the loss function for the overall model weight; i is the number of training participants; P_i is the probability that the i-th participant is selected for training; $f_i(\omega)$ is the minimum value of the local model weight loss function for the i-th participant. As the number of participants increases, the loss function value of the overall model weight is constantly changing, which implements a dynamic weighting process and helps the model obtain the global optimal solution.

2.5. Differential Privacy Protection

When training the overall model, federated learning is susceptible to inference attacks (Gao et al., 2021), resulting in athlete data leakage and affecting competitive fairness and personal privacy. In addition to encrypting the data, this paper uses differential privacy to ensure that the identity of athletes is not recognized, and to ensure that the dataset can be used for statistical analysis. Differential privacy (Chen et al., 2021; H. Li et al., 2021) is a cryptographic technique that can reduce identification operations when querying databases while improving query accuracy. In differential privacy, random noise is added to the data to protect privacy. This paper chooses to add Gaussian noise to achieve differential privacy protection. Assuming the probability density function of the Gaussian distribution is $G(\sigma)$, its expression is:

$$G(\sigma) = \frac{1}{\sqrt{2\pi}} \exp(\frac{1}{2\sigma^2}) \tag{7}$$

Among them, σ is the Gaussian distribution N $(0,\sigma^2)$. The sensitivity of Gaussian noise calculation is calculated using the L2 norm method in this paper. The sensitivity calculation method for the overall model is:

$$S = \left\| F(\theta) - F(\tilde{\theta}) \right\|_2 \tag{8}$$

Among them, S represents sensitivity; θ is the parameter of the overall model; $F(\theta)$ is the corresponding loss function; $F(\tilde{\theta})$ is the loss function after adding random noise. By calculating the L2 norm of the parameter change vector, the sensitivity of the overall model in the parameter space, that is, the degree of response to parameter changes, is evaluated. Figure 5 shows the situation of adding Gaussian noise during overall model training:

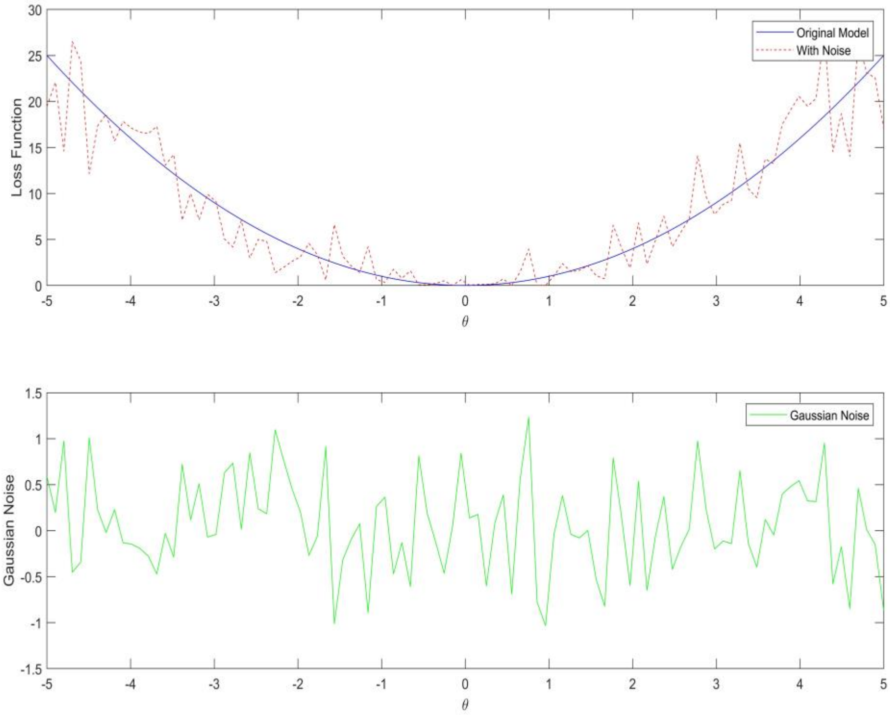


Figure 5: Schematic Diagram of Adding Gaussian Noise

The upper part of Figure 5 shows the original loss function of the overall model and the loss function with added noise. It can be seen that after adding noise, the loss function deviates from its original position on most nodes. The application of noise changes the path of model optimization and also increases the difficulty of identifying athlete identities in the original data. The following figure in Figure 5 shows the amplitude of Gaussian noise. After adding Gaussian noise to the overall model function, if certain specific distributions are ignored, the privacy loss is directly added up, resulting in a loose privacy boundary. The solution of privacy loss involves quantifying the degree of privacy leakage and evaluating it through privacy loss measurement. This paper uses conditional mutual information to measure privacy loss. Conditional mutual information is a method in information theory that measures the coexistence or dependency relationship between two random variables. Assuming there are three random variables X, Y, and Z, the conditional mutual information $I(X;Y|Z)$ is defined as:

$$I(X;Y|Z) = \sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} P(x,y,z)\log\frac{P(x,y|z)}{P(x|z)P(y|z)} \qquad (9)$$

Among them, $P(x,y|z)$ represents the probability of X=x and Y=y under the premise of Z. $P(x|z)$ and $P(y|z)$ represent the marginal probability distribution of X and Y at this time. The smaller $I(X;Y|Z)$, the lower the correlation between X and Y, and the smaller the privacy loss.

2.6. Model Optimization

When training distributed models, it is not only necessary to ensure model convergence, but also to consider the performance of the model. This paper adopts certain technical means to optimize the model. In order to reduce the communication overhead from the local model to the overall model, this paper uses compression techniques to compress the parameters of each local model. The integrity of athlete health data is crucial for the understanding of the model. This paper chooses Huffman encoding in lossless compression to compress the data. Huffman encoding marks data bytes by constructing a Huffman tree and encodes them based on the frequency of byte occurrence. Assuming O_i is the frequency of the i-th byte and L_i is its length, then its Huffman encoding length $L_{Huffman}$ is:

$$L_{Huffman} = \sum_{i=1}^{m} O_i \times L_i \tag{10}$$

Among them, m is the number of characters. The adjustment of learning rate is very important, and this paper chooses to apply a dynamic learning rate adjustment mechanism to improve the convergence speed and performance of the model. RMSprop (Root Mean Square Propagation) is an algorithm that uses the exponential decay average of historical gradients to adaptively adjust learning rates. It solves the problem of premature decay in some dynamic adjustment learning rate algorithms. The learning rate update rule in RMSprop is:

$$R_{t+1} = \frac{R_t}{\sqrt{E[g^2]t+C}} \tag{11}$$

Among them, R_t is the learning rate at time t; R_{t+1} is the learning rate for the next moment; $E[g^2]t$ is the weighted average of historical gradients; C is a constant, taken between 10^{-8} and 10^{-6}. To prevent overfitting in the model, L1 regularization is added to the loss function in this paper. L1 regularization term is added to $F(\theta)$ in the loss function:

$$J(\theta) = F(\theta) + \tau \sum_{i=1}^{i} \left|\theta_i\right| \tag{12}$$

Among them, $J(\theta)$ is the modified loss function; $F(\theta)$ is the original loss function; θ is the parameter of the overall model; i is the number of parameters. For the model selection of each training participant, the overall model prioritizes selecting local model data upload parameters with fast data update frequency and high data quality. This ensures that the overall server obtains high-quality information and improves the performance of the overall model.

3. Evaluation of Privacy Protection Effectiveness and Model Performance

3.1. Datasets

In the study, real athlete data involves personal privacy and sensitive information. Therefore, in this study, some health data datasets are used instead of athlete data. PhysioNet is a database that provides physiological signals that can be used to analyze health conditions. Fitness Tracker Data is a dataset generated by the fitness tracker, which includes information such as step count, heart rate, sleep, etc. Electronic Health Records is a medical record dataset that includes diagnostic, treatment, and medication information. The federated learning model is trained and tested using the above dataset.

3.2. Model Performance Tests

In order to test the performance of the federated learning distributed model in this paper, 1000 labeled data samples are selected in the dataset for testing the model's classification and regression metrics. The test indicators include accuracy A, precision P, recall R, and F1 value. Meanwhile, models established by K-Means Algorithm (KMA), PageRank Algorithm, and Support Vector Machine (SVM) are applied for comparative experiments. The experiment is conducted three times, and the results obtained are shown in Table 1:

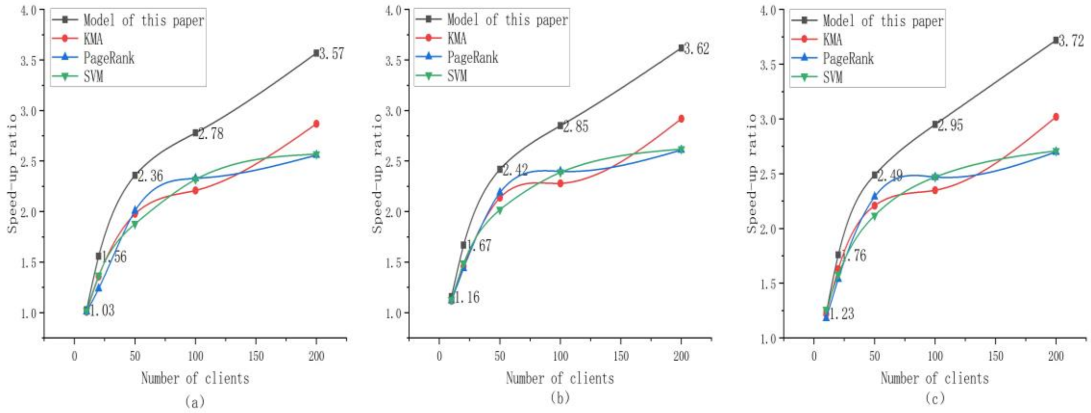
Table 1: Performance Tests of Each Model

INDEX	A (%)	P (%)	R (%)	F1 SCORE
EXPERIMENT 1				
MODEL OF THIS PAPER	90	90.58	88.31	0.89
KMA	86	83.79	86.33	0.85
PAGERANK	86.5	82.74	86.77	0.85
SVM	89.1	89.14	87.01	0.88
EXPERIMENT 2				
MODEL OF THIS PAPER	93.8	91.99	94.44	0.93
KMA	75.2	85.27	60.56	0.71
PAGERANK	86	85.84	83.22	0.85
SVM	86.5	84.18	86.93	0.86
EXPERIMENT 3				
MODEL OF THIS PAPER	96.55	98.45	93.47	0.96
KMA	84	84.36	81.14	0.83
PAGERANK	88.5	87.75	87.17	0.87
SVM	86	86.03	83.83	0.85

Table 1 shows the performance indicators of each model in three experiments. Overall, the model of this paper performs better than other models in various indicators. In the third experiment, the accuracy, precision, recall, and F1 values reach 96.55%, 98.45%, 93.47%, and 0.96 respectively, which are the highest among all test results. In the second experiment, the KMA model performs poorly with an accuracy of 75.2% and a recall rate of only 60.56%. The performance index is very low, and it differs significantly from its performance in the other two experiments due to overfitting issues that occurred during the process, resulting in poor performance. The experimental results demonstrate that the model proposed in this paper performs well in regression and classification when processing athlete health data.

3.3. Speedup Ratio Experiment

In distributed model training, the synchronization interval refers to the time interval between the local model parameters of the training participants after the overall model is updated. There is a client for each participant's local model, and during the synchronization interval, the client updates the operation. When updating, the client uploads the updated model parameters to the central server for overall model updates. The synchronization interval affects the overall performance of the model. Six different synchronization intervals are set in the experiment, namely 1, 5, 10, 15, 20, and 50. When the synchronization interval is 1, it means that each client immediately synchronizes the model parameters with the server after local updates. This has the advantage of accelerating the overall model convergence, but it increases communication overhead because each client needs to communicate frequently with the server. As the synchronization interval increases, the convergence speed of the overall model slows down, and the communication overhead also decreases. The experiment tests the speedup ratio of four models in a distributed structure with different synchronization intervals as the number of clients changes. The speedup ratio is defined as the ratio of the convergence time of the distributed structure model to the benchmark convergence time. The results obtained are shown in Figure 6:



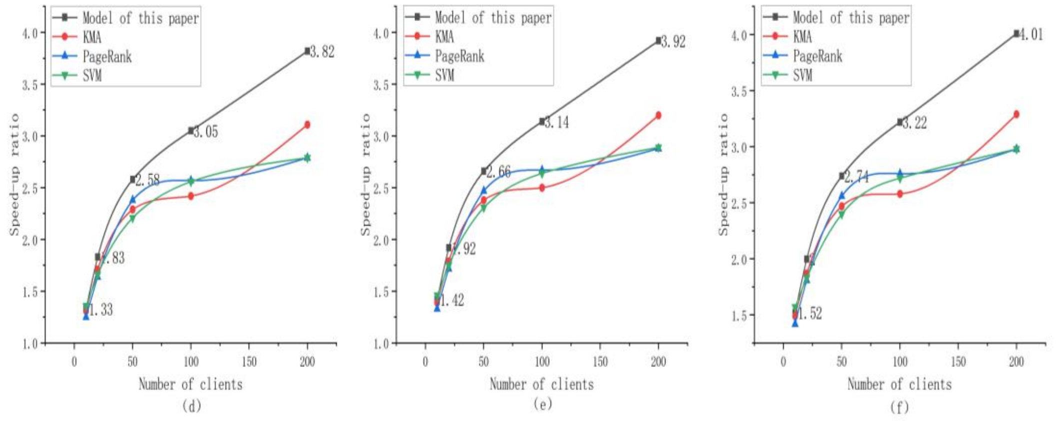


Figure 6: Statistical Results of Speedup Ratio Experiment

Figures (a) to (f) in Figure 6 represent the speedup ratios of each model as the number of clients varies with synchronization intervals of 1, 5, 10, 15, 20, and 50, respectively. Observing Figure 6, it can be seen that under the same synchronization interval conditions, the speedup ratio of each model increases rapidly with the increase of clients. This is because more clients participate in training, and the overall convergence speed of the model increases. When the number of clients increases to a certain extent, the increase in speedup ratio becomes slow because adding too many clients also increases communication overhead and synchronization time, offsetting the overall computational power improvement of the model.

Meanwhile, increasing the synchronization interval can also slightly increase the speedup ratio. The speedup ratio of this model increases the most with the increase of clients. When the synchronization interval is 1 and the number of clients is 200, the speedup ratio of this model is 3.57. When the synchronization interval is 50 and the number of clients is 200, the speedup ratio of this model is 4.01. The experimental results demonstrate that the model in this paper still has good convergence speed when there are a large number of clients and a large synchronization interval. When calculating athlete health data, it can reduce the number of communications and reduce the risk of privacy leakage.

3.4. Privacy Leakage Risk Testing

To evaluate the effectiveness of federated learning in protecting athlete health data privacy, this study tests the effectiveness of the model in this paper by simulating three types of attacks. The three types of attacks are: member inference attack, model reverse engineering attack, and data leakage attack. Member inference attacks attempt to infer specific athlete data from the dataset; model reverse engineering attacks attempt to infer the physiological characteristics of athletes from pre trained models; data leakage attacks attempt to intercept data during the transmission process of the model. The experimental setup involves 1000 attacks of three types on each of the four models, and calculates the success rate and average attack time of the attacks. The obtained results are shown in Table 2:

Table 2: Test Results for Various Types of Attacks

TYPES OF MODELS	ATTACK TYPES	NUMBER OF SUCCESSFUL ATTACKS	NUMBER OF FAILED ATTACKS	ATTACK SUCCESS RATE (%)	MEAN ATTACK TIME(S)
MODEL OF THIS PAPER	A	15	97	1.5	85.2
	B	23	95	2.3	90.5
	C	14	86	1.4	100.8
KMA	A	124	88	12.4	55.4
	B	139	87	13.9	45.9
	C	151	85	15.1	50.6
PAGERANK	A	213	79	21.3	49.3
	B	118	89	11.8	34.6
	C	137	87	13.7	57.4
SVM	A	129	88	12.9	67.1
	B	172	83	17.2	72.4
	C	134	87	13.4	66.7

Table 2 shows three types of attacks: member inference attack, model reverse engineering attack, and data leakage attack, represented by A, B, and C, respectively. The model in this paper performs best in resisting three types of attacks, with the least number of successful attacks. The success rates of the three attacks are 1.5%, 2.3%, and 1.4%.

The PageRank model performs the worst in resisting member inference attacks, with a success rate of 21.3%. The average attack time represents the efficiency of attackers in obtaining information, and the longer the time, the better the security of the model. The average attack time of the model in this paper is the longest, especially when resisting data leakage attacks, with an average attack time of 100.8 seconds. The experimental results demonstrate the excellent performance of this paper in protecting the privacy of athlete health data, providing reliable protection for the data.

3.5. Model Loss Experiment

In order to test the training convergence performance of the model in this paper, four algorithms are set up under the same conditions to record the changes in the loss function as the training period increases. The values of the loss function are recorded, and the results obtained are shown in Figure 7:

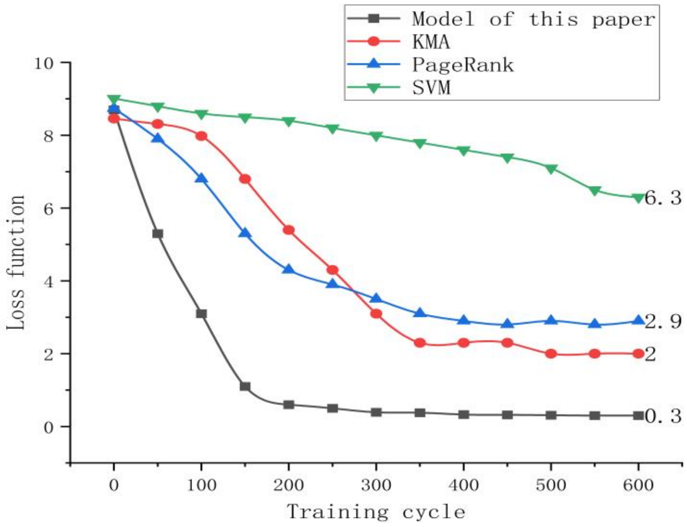


Figure 7: Changes in Loss Function

Figure 7 shows the variation of the loss functions of each model with the training period of the model. As the training period increases, the loss function first rapidly decreases, then slowly decreases, and finally converges. It can be seen that the model in this paper learns the features of the data very quickly, so the loss function decreases the fastest in the initial stage, followed by the PageRank model, and the SVM model decreases the slowest. When the model converges, the convergence value of the model in this paper is also the smallest, with a convergence value of 0.3. Although the descent speed of the KMA model is slightly slower than that of the PageRank model, the final convergence value is 2, which is lower than the PageRank model's 2.9. The experiment proves that the model in this paper has a fast convergence speed and low convergence value during training, reflecting good convergence performance.

4. Conclusions

This paper was based on the federated learning framework and investigated a distributed model training method for athlete health data, with the aim of addressing the shortcomings of traditional athlete health data analysis models in terms of privacy protection and model performance. By segmenting athlete health data, the model processing efficiency can be improved. The encryption technology was used to protect data transmission and computing security. Lightweight models were chosen to process data and train local models. During the model aggregation phase, dynamic weighted learning methods were used to optimize model updates. The differential privacy technology was applied, and the Gaussian noise was added to protect data privacy and improve query accuracy. The experimental results showed that the model proposed in this paper outperformed models such as KMA, PageRank, and SVM in performance indicators such as accuracy, precision, recall, and F1 value. At the same time, it exhibited high efficiency and security in speedup ratio experiments and privacy leakage risk tests. Slightly insufficient is that this paper did not study the specific computational cost of the model, and the computational performance without considering cost resulted in a lack of comprehensive consideration of the model. It is hoped that in the future, more expert suggestions can be combined to improve the research gap in this area.

REFERENCES

Aghighi, F., Ebadati E, O. M., & Aghighi, H. (2022). SVM-CRF Method and Box Plot Technique for Outlier Detection of Lidar Point Cloud. *Iranian Journal of Remote Sensing & GIS*, *14*(2), 91-109.

Akbar, B. R., & Iqbal, M. (2022). Design and Build Data Transfer Process Security on File Transfer Protocol (Ftp) Servers Using Transport Layer Security (Tls) Protocol. *Jurnal Mantik*, *6*(2), 2664-2675.

Alrmali, A., Stuhr, S., Saleh, M. H., Latimer, J., Kan, J., Tarnow, D. P., & Wang, H. L. (2023). A decision-making tree for evaluating an esthetically compromised single dental implant. *Journal of Esthetic and Restorative Dentistry*, *35*(8), 1239-1248.

Arachchige, P. C. M., Bertok, P., Khalil, I., Liu, D., Camtepe, S., & Atiquzzaman, M. (2020). A trustworthy privacy preserving framework for machine learning in industrial IoT systems. *IEEE Transactions on Industrial Informatics*, *16*(9), 6092-6102.

Chen, S., Fu, A., Su, M., & Sun, H. (2021). Trajectory privacy protection scheme based on differential privacy. *Tongxin Xuebao*, *42*(9).

Fenglei, T., Zhaojun, Z., Xingquan, W., Guangbin, W., & Hongzhong, M. (2020). Application of prediction accuracy interpolation method based on support vector machine optimization in transformer oil temperature preprocessing. *Modern Electric Power*, *37*(6), 591-597.

Gafni, T., Shlezinger, N., Cohen, K., Eldar, Y. C., & Poor, H. V. (2022). Federated learning: A signal processing perspective. *IEEE Signal Processing Magazine*, *39*(3), 14-41.

Gao, J., Hou, B., Guo, X., Liu, Z., Zhang, Y., Chen, K., & Li, J. (2021). Secure aggregation is insecure: Category inference attack on federated learning. *IEEE Transactions on Dependable and Secure Computing*, *20*(1), 147-160.

Gupta, R., & Singh, A. K. (2022). A differential approach for data and classification service-based privacy-preserving machine learning model in cloud environment. *New Generation Computing*, *40*(3), 737-764.

Hammoud, A., Otrok, H., Mourad, A., & Dziong, Z. (2022). On demand fog federations for horizontal federated learning in IoV. *IEEE Transactions on Network and Service Management*, *19*(3), 3062-3075.

Jing, L., Jiahao, Z., Runmeng, Y., & Haipeng, J. (2023). Federated dynamic weighted learning method based on non-independent and identically distributed industrial big data. *Computer Integrated Manufacturing System*, *29*(5), 1602.

Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, *2*(6), 305-311.

Kawamura, A., Kinoshita, Y., Nakachi, T., Shiota, S., & Kiya, H. (2020). A privacy-preserving machine learning scheme using etc images. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, *103*(12), 1571-1578.

Khan, L. U., Saad, W., Han, Z., Hossain, E., & Hong, C. S. (2021). Federated learning for internet of things: Recent advances, taxonomy, and open challenges. *IEEE Communications Surveys & Tutorials*, *23*(3), 1759-1799.

Li, H., Ren, X., Wang, J., & Ma, J. (2021). Continuous location privacy protection mechanism based on differential privacy. *Journal on Communications*, *42*(8), 102-110.

Li, T., Li, J., Chen, X., Liu, Z., Lou, W., & Hou, Y. T. (2020). NPMML: A framework for non-interactive privacy-preserving multi-party machine learning. *IEEE Transactions on Dependable and Secure Computing*, *18*(6), 2969-2982.

Li, X., He, J., Vijayakumar, P., Zhang, X., & Chang, V. (2021). A verifiable privacy-preserving machine learning prediction scheme for edge-enhanced HCPSs. *IEEE Transactions on Industrial Informatics*, *18*(8), 5494-5503.

Liu, J., Huang, J., Zhou, Y., Li, X., Ji, S., Xiong, H., & Dou, D. (2022). From distributed machine learning to federated learning: A survey. *Knowledge and Information Systems*, *64*(4), 885-917.

Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., Li, J., & Poor, H. V. (2021). Federated learning for internet of things: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, *23*(3), 1622-1658.

Nguyen, H. T., Sehwag, V., Hosseinalipour, S., Brinton, C. G., Chiang, M., & Poor, H. V. (2020). Fast-convergent federated learning. *IEEE Journal on Selected Areas in Communications*, *39*(1), 201-218.

Niu, C., Wu, F., Tang, S., Ma, S., & Chen, G. (2020). Toward verifiable and privacy preserving machine learning prediction. *IEEE Transactions on Dependable and Secure Computing*, *19*(3), 1703-1721.

Pfitzner, B., Steckhan, N., & Arnrich, B. (2021). Federated learning in a medical context: a systematic literature review. *ACM Transactions on Internet Technology (TOIT)*, *21*(2), 1-31.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., & Maier-Hein, K. (2020). The future of digital health with federated learning. *NPJ digital medicine*, *3*(1), 1-7.

So, J., Güler, B., & Avestimehr, A. S. (2021). CodedPrivateML: A fast and privacy-preserving framework for distributed machine learning. *IEEE Journal on Selected Areas in Information Theory*, *2*(1), 441-451.

Sun, G., Xu, J., & Zuo, M. (2023). The role of multi-layer spiral CT based perfusion imaging in lung cancer radiotherapy assessment in athletic patients. *Revista multidisciplinar de las Ciencias del Deporte*, *23*(89).

Tan, Z. W., & Zhang, L. F. (2020). Survey on privacy preserving techniques for machine learning. *Journal of Software*, *31*(7), 2127-2156.

Wang, R., Tang, Y., Zhang, W., & ZHANG, F. (2020). Privacy protection scheme for internet of vehicles based on homomorphic encryption and block chain technology. *Chin. J. Netw. Inf. Secur*, *6*(01), 46-53.

Yan, L., Xianhe, L., & Shaojing, P. (2020). Encryption algorithm based on ECC and homomorphic encryption [J]. *Computer Engineering and Design*, *41*(05), 1243-1247.

Yu, R., & Li, P. (2021). Toward resource-efficient federated learning in mobile edge computing. *IEEE Network*, *35*(1), 148-155.

Zhan, Y., Zhang, J., Hong, Z., Wu, L., Li, P., & Guo, S. (2021). A survey of incentive mechanism design for federated learning. *IEEE Transactions on Emerging Topics in Computing*, *10*(2), 1035-1044.

Zhang, T., Gao, L., He, C., Zhang, M., Krishnamachari, B., & Avestimehr, A. S. (2022). Federated learning for the internet of things: Applications, challenges, and opportunities. *IEEE Internet of Things Magazine*, *5*(1), 24-29.

ZHOU, C., SUN, Y., WANG, D., & GE, H. (2021). Top Read Articles. *Chinese Journal of Network and Information Security*, *7*(5), 77-92.

Zhu, L., Tang, X., Shen, M., Gao, F., Zhang, J., & Du, X. (2021). Privacy-preserving machine learning training in IoT aggregation scenarios. *IEEE Internet of Things Journal*, *8*(15), 12106-12118.