A Multimodal Robust Recognition Method for Grasping Objects With Robot Flexible Grippers

Qiaokang Liang, *Senior Member, IEEE*, Wenxing Xiao, Jianyong Long, *Member, IEEE*, and Dan Zhang, *Senior Member, IEEE*

***Abstract*—In light of the critical importance of achieving robust object recognition from multimodal data in robotic operations, this article proposes a precise identification method tailored for the grasping of objects using a multiflexible gripper in scenarios characterized by multimodality, limited samples, and complex environments. Terming the BOSS-MI-ELM algorithm, this approach innovatively extracts [bag-of-SFA-symbols (BOSS)], fusion [association-based fusion (AF)], and classifies [incremental extreme learning machine (I-ELM)] features from multimodal data, facilitating an efficient recognition process. The study employs fiber Bragg grating (FBG) and inertial measurement unit (IMU) as information acquisition components, constructing a multimodal perception system and establishing a corresponding grasping dataset. Through training and testing on this dataset, empirical evidence demonstrates that even with the utilization of only 20% of the dataset, the BOSS-MI-ELM algorithm maintains a classification accuracy of 95.54%. In the presence of Gaussian noise with a mean of 0 and varying standard deviations, as well as different degrees of partial data loss, the proposed method still maintains robust recognition performance. In addition, we have validated the effectiveness of this method in identifying objects grasped at different speeds. Furthermore, comparative experiments were conducted on two publicly available multimodal tactile datasets. The results indicate that the BOSS-MI-ELM algorithm outperforms various baseline models. The extensive experiments collectively demonstrate that this system provides a viable solution for robot object recognition under multimodal tactile perception.**

Manuscript received 14 November 2023; revised 12 March 2024 and 29 April 2024; accepted 21 June 2024. This work was supported in part by the Project of the National Key Research and Development Program of China under Grant 2022YFB4703103, in part by the National Natural Science Foundation of China under Grant 62303175 and Grant 62073129, and in part by the Hunan Provincial Natural Science Foundation of China under Grant 2022JJ10020. Recommended by Technical Editor M. Rakotondrabe and Senior Editor J. Ueda. *(Corresponding author: Jianyong Long.)*

Qiaokang Liang, Wenxing Xiao, and Jianyong Long are with the National Engineering Research Center for Robot Vision Perception and Control, College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: qiaokang@hnu.edu.cn; xiaowx126@hnu.edu.cn; ljy_cv@hnu.edu.cn).

Dan Zhang is with the Department of Mechanical Engineering, Hong Kong Polytechnic University, Hong Kong (e-mail: dan.zhang@polyu.edu.hk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMECH.2024.3419420.

Digital Object Identifier 10.1109/TMECH.2024.3419420

***Index Terms*—Fiber Bragg grating (FBG), inertial measurement unit (IMU), multiflexible gripper, multimodal perception, robust object recognition.**

I. Introduction

In RECENT years, the application of flexible grasping devices in robotic manipulation tasks has steadily grown, closely associated with the emergence of soft robotics technology [1]. This trend is primarily attributed to the enhanced safety and passive adaptability exhibited by flexible grasping devices during interaction, in comparison to rigidly fixed devices [2]. Robotic grasping tasks encompass various activities, including target detection, grasp pose estimation, and object property recognition, with the majority relying on visual methods for exploration [3], [4], [5], [6]. However, visual perception has inherent limitations, necessitating the integration of tactile sensing modalities to accomplish specific tasks, such as identifying transparent objects [7] and objects with similar color and shape features [8], [9]. Tactile sensors contribute to capturing information, such as hardness, texture, roughness, and smoothness through physical contact, thereby enhancing the effectiveness and adaptability of grasping devices during task execution [10]. However, current tactile-based grasping still faces challenges due to the lack of standardized sensor platforms and extensive datasets [11]. Therefore, for flexible grippers, the integration of multimodal tactile sensing capabilities and the achievement of goal-oriented grasping and recognition in scenarios with limited samples and complex environments are challenging tasks with significant implications for the intelligence and safety of robotic operations.

In the field of rigid grasping research, Zhang et al. [12] effectively identified the hardness and elasticity of grasped objects using NumaTac sensors to capture dc pressure signals and ac pressure vibration signals. Lin et al. [13] verified that object instances can be accurately identified based on tactile information alone by applying two GelSight tactile sensors to stiff fingers. Xia et al. [14] solved the problem of pose uncertainty using a robotic hand equipped with tactile and finger joint displacement sensors. Chu et al. [15] successfully classified haptic descriptors using BioTacs sensors to collect multimodal data on objects. Bhattacharjee et al. [16] used multimodal sensors to capture variables associated with moving objects to identify properties, such as object stiffness, mobility, and material properties. These studies utilized either unimodal or multimodal tactile sensors.

By integrating information from different sensors, the robot can fully perceive its surroundings, thus improving the robustness and accuracy of the recognition.

However, grasping and recognizing objects based on flexible grippers is still a major challenge. Bai et al. [17] mounted resistive pressure sensors on a silicon-based manipulator to help recognize the roughness of an object. Jin et al. [18] utilized triboelectric nanogenerator (TENG) sensors to recognize an object by capturing continuous motion and positional information of a flexible finger. Zuo et al. [19] used ionic hydrogel-based strain and haptic sensors to achieve object recognition. Liu et al. [20] proposed a novel GelSight Fin Ray design that enhances robotic flexible fingers with high-resolution tactile sensing, object orientation estimation, and force marker tracking. Although these studies demonstrate the feasibility of flexible fixtures, their haptic feedback information is to some extent unidimensional. Kerzel et al. [21] pointed out that tactile perception can be viewed as a complex multimodal perception process. Systems that can receive information from different modalities have two main advantages over single data modality systems: performance and reliability [22].

Moreover, these works either employ electromagnetic-sensitive components for measurement, susceptible to electromagnetic interference issues with complex wiring and zero drift; or they use visually based tactile sensors with larger volumes and rigidity, making integration on flexible grasping devices challenging. In response, Lyu et al. [23] designed a modal perception system based on flexible grippers using fiber Bragg grating (FBG). In recent years, FBG has been widely used in force sensing for medical robots due to its advantages, such as flexibility, resistance to electromagnetic interference, small size, lightweight, good biocompatibility, and zero temperature drift [24], [25]. FBG provides higher accuracy [26], but it is susceptible to temperature changes and prone to breakage. This introduces noise injection and data missing in the signal feedback, posing significant challenges to the accuracy and robustness of grasping recognition by flexible grippers [27]. Simultaneously, constructing large-scale training samples is challenging for target recognition by flexible grippers, leading to issues such as decreased accuracy in network training.

Aiming at the above challenges, this article provides an in-depth analysis of the multimodal information involved in the flexible manipulator when it performs the grasping task, as shown in Fig. 1. We plan to utilize multimodal data to improve the perceptual performance of the system. While realizing general object recognition based on FBG, we combine the inertial measurement unit (IMU) to collect the motion parameters of the flexible gripper, and jointly obtain the implicit information, such as the hardness, roughness, and shape and size of the grasped object, so as to realize the recognition of the grasped object under limited samples and complex environments. And a multimodal tactile sensing method is proposed. The method is composed of the bag-of-SFA-symbols (BOSS) model, which has excellent performance in time series classification models, the association-based fusion (AF) framework, which is suitable and effective in arbitrary multimodal fusion, and the extreme learning machine (ELM) classifier, which can classify quickly and accurately. It provides a robust solution for scenarios involving multimodal perception, limited samples, as well as the presence of noise and data gaps in robotic grasping scenes. Moreover, with good scalability, the three parts of feature extraction, fusion, and classification can be replaced by more advanced methods in the future. The primary contributions of this research can be summarized as follows:

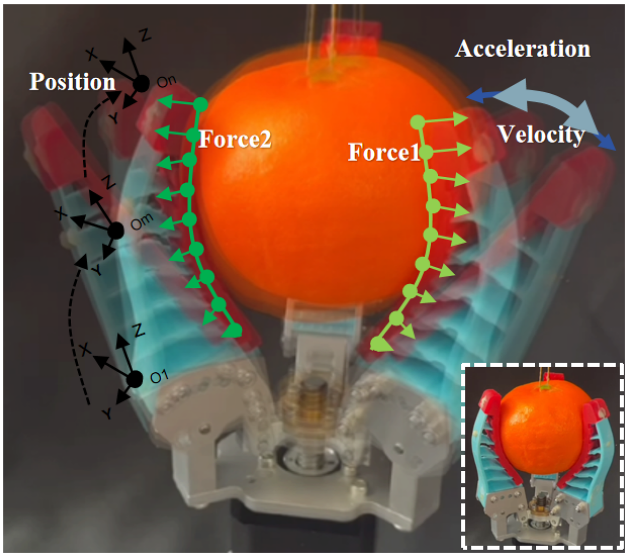
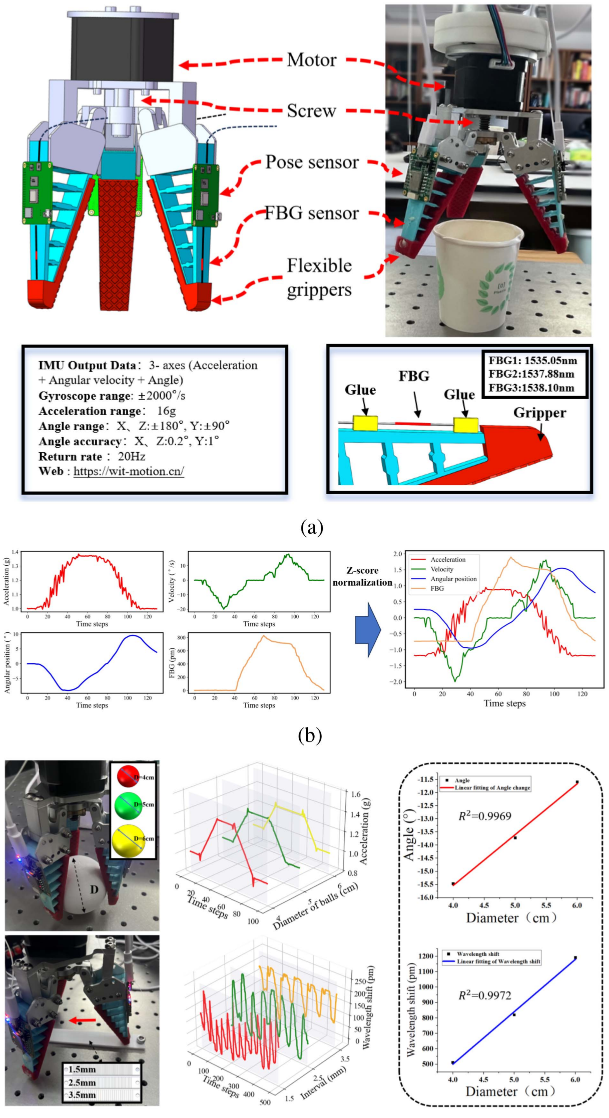


Fig. 1. Multimodal information representation of flexible grippers grasping target recognition.

1) For the first time, an integrated system of FBG and IMU on a robotic flexible finger is introduced, forming a system capable of collecting multimodal force-tactile data, as illustrated in Fig. 1. Simultaneously, a novel method named BOSS-MI-ELM was proposed for recognizing the grasping targets of multifingered robots.
2) A multimodal haptic dataset of 1694 samples built at certain gripping speeds for nine common object categories. And a small dataset built at three additional grasping velocities.
3) Performance analysis of the BOSS-MI-ELM algorithm was conducted using the self-constructed dataset, comparing it with several common baseline models, and considering the impact of data volume and gripper speed on model performance.
4) The effectiveness and robustness of the BOSS-MI-ELM algorithm were validated by injecting noise and introducing partial data loss operations into the dataset.
5) The application of this algorithm to publicly available multimodal tactile datasets yielded satisfactory results. It further substantiated the effectiveness of multimodal data and the integration model with multiple fingers. Moreover, an examination of the algorithm's real-time performance was conducted.

The rest of this article are organized as follows. Section II provides a detailed exposition of the multimodal tactile perception system. Section III delves into the recognition algorithm BOSS-MI-ELM for grasping targets with a flexible gripper.



(a)

(b)

(c)

Fig. 2. Grasping platform, multimodal tactile data, and test examples of the system. (a) Flexible claw, sensor arrangement, and parameters. (b) Captured multimodal force-tactile data. (c) The left is a test of the platform gripping a sphere and sliding on a rough surface. The center corresponds to the acceleration and strain captured by the test. The right is the result of fitting the diameter of the sphere to the most value in the collected strain and angle data, respectively.

Section IV comprises the experimental part of this study. Finally, Section V concludes this article.

II. Flexible Grasping System and Multimodal Sensing

The three-finger flexible gripper in this study is driven by a motor that moves the fingers through a screw to perform the grasping and releasing actions. In addition, the gripper was externally mounted with FBGs and IMUs to facilitate the capture of strain, angle, angular velocity, and acceleration during the grasping and releasing maneuvers. Fig. 2(a) illustrates the structural configuration of the device and the relevant parameters of the sensors, and (b) shows the effective data collected.

Since the flexible manipulator has both internal and external senses [2], IMUs were mounted on the exterior of the flexible claw to capture the external sensation after deformation more accurately. While the IMU demonstrates the capability to capture data along all three axes, it is imperative to acknowledge that the inherent constraints imposed by the finger's movement within the fixture confine data acquisition to a singular plane. Hence, the exclusively valid data comprises angle and velocity along the *X*-axis and acceleration in the context of rotation around the *Z*-axis (relative to IMU).

Furthermore, we employ FBG to monitor the strain exerted on the claw during object manipulation. In contrast to the approach undertaken by Lyu [23], who place FBG on the interior of the flexible claw, resulting in secondary deformation when the claw contacts an object, we adhere FBG to the exterior of the flexible claw. This placement ensures that the deformation observed is solely attributed to the bending of the claw itself, eliminating the influence of object contact on FBG readings. The three FBGs were packaged with a two-point encapsulation method, exhibiting initial wavelengths (nm) of 1535.05, 1537.88, and 1538.1, all consistently sampled at a frequency of 100 Hz. Their respective 3 dB bandwidths (nm) were determined as 0.33, 0.35, and 0.32, while the reflectance percentages were measured at 85, 89.86, and 83.13. And the edge suppression ratios (dB) are all 18.

During the grasping process, we set the maximum angle (Screw up to 5 revolutions) of closure of the claw to prevent excessive grasping. We set an upper limit on the strain (1500 pm FBG wavelength shift) for the FBG to ensure that the stress due to deformation is within a certain range. Finally, we relied on the current (0.55 A) of the flexible claw controller to control the maximum value of the torque. Together, these ensure that the claw does not damage the grasped object.

We recorded the strain of the FBG by sliding one of three fingers over surfaces of different roughness, as well as the acceleration in the IMUs using three balls of different diameters grasped through the entire platform, as shown in the left and center of Fig. 2(c). It can be seen that the strain obtained from different surfaces or the acceleration obtained from different balls in both tests are significantly different, which indicates that the system has the potential ability to recognize the material properties and size of objects. We fitted the strain and angle acquired by the grasping balls with different diameters, and the results are shown on the right of Fig. 2(c). The fitted results illustrate the excellent linearity of the data acquisition part of the system. Based on this, we designed the relevant perception algorithms for recognizing different objects.

III. Methodology

The flowchart for this multifinger grasp target recognition method is shown in Fig. 3. The subsequent sections provide a detailed account of the specific procedures involved in this method.

Definition: For single-finger force/tactile dataset denoted as $\{\mathbf{D}^p = (\mathbf{T}_n^p, y_n)_{n=1}^N | 1 \le p \le P\}$, where p denotes the number of distinct modalities and N signifies the total sample count.

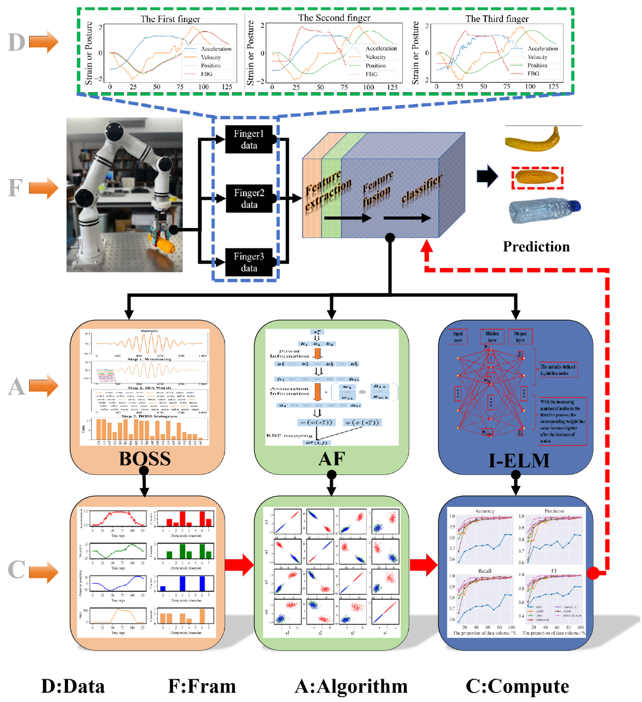


Fig. 3. overall framework of the BOSS-MI-ELM algorithm.

During the ensuing discourse, it is of consequence to highlight that, due to the preliminary processing steps of the method presented in this study not engaging in interfinger operations, the act of distinguishing between individual fingers has been deliberately omitted within the process of formulation.

A. Feature Extraction Based on BOSS

The BOSS algorithm [28] exhibits robust resistance to noise and minor variations, showcasing excellent performance in the context of time-series classification within a singular model [29]. However, its intrinsic distance classifier imposes limitations on scalability. Hence, we selectively utilize its symbolic abstract representation to extract features from force/tactile data, as follows:

Step 1. Window creation: For any sample $T^p_n = \{t^p_i\}^{n_p}_{i=1} \in \mathbf{T}^p$, where n_p is the length of the sample sequence, it can be divided into $n_p - w + 1$ subsequences of length w

$$\begin{aligned}&\text{Window}(T^p_n, w)\\ &= \left(S^p_{n,1:w}, S^p_{n,2:w}, \ldots, S^p_{n,n_p-w+1:w}\right) \qquad (1)\end{aligned}$$

where

$$S^p_{n,i:w} = \left(t^p_i, t^p_{i+1}, \ldots, t^p_{i+w-1}\right) \qquad (2)$$

which means this window starts at t^p_i and the length of it is w, each window is typically Z-score normalized to obtain offset and amplitude invariance.

Step 2. SFA transform: The Symbolic Fourier Approximation (SFA) is employed, which utilizes a finite set of characters to represent real-valued time series for low-pass filtering and string representation. Specific operation reference [30]. After the SFA transformation, each subsequence of length w in the window will be mapped to a string of length l, resulting in $n_p - w + 1$ ordered strings for a time series of length n_p, followed as follows:

$$S^p_{n,i:w} \xrightarrow{\text{SFA}} s_1, s_2, \ldots, s_l \in \sum\nolimits_a \qquad (3)$$

where $\sum_a$ represents an alphabet with a elements.

Step 3. Histogram-based statistics: To prevent an overemphasis on stable sections of a signal, numerosity reduction techniques are employed. This involves retaining the string symbolized by the previous window and discarding the current window's string if it matches the previous one. The results

Algorithm 1: BOSS-MI-ELM.

```
Input: $D$. Force/Tactile dataset for each individual finger,
  with a total of $M$ fingers;
    $w$. Window size;
    $l$. String length;
    $a$. Alphabet size;
    $L$. the highest power of data;
    $L_0$ and $L_{max}$. The initial node and upper limit count for
    each I-ELM, respectively;
    $E^*$. Expected test error;
    $\mathbf{X}^*$. Unseen instances.
    Output: $\mathbf{Y}^*$. Predicted class for $\mathbf{X}^*$
1:  for each Force/tactile Data $D$ do
2:    for $p = 1$ to $P$ do
3:      Create windows based on (1);
4:      $SFA$ transformation for each window based on (3);
5:      From $B^p$ based on $D^p$ by histogram statistics;
6:    end for
7:    for $p = 1$ to $P$ do
8:      for $n = 1$ to $N$ do
9:      Calculate $\phi_p(\mathbf{x}^p_n)$ based on (4);
10:   end for
11:   From $H^p$ based on $B^p$ according to (5);
12:   Calculate the relationship fusion matrix $R^p$ based on
      (6);
13:   for n=1 to $N$ do
14:     Calculate $\varphi(\phi_p(\mathbf{x}^p_n))$ based on (7);
15:   end for
16:   From $A^p$ based on $H^p$ according to (9);
17:   From $G^p$ based on $A^p$ according to (10);
18:   end for
19: end for
20: for $i = 1$ to $M$ do
21:   Initialize the I-ELM$i$ based on (11), (12) and (13);
22: for $h = L_0$ to $L_{max}$ or $e < E^*$ do
23:   for $i = 1$ to $M$ do
24:     Add note and train based on (14);
25:     Update training error E based on (15);
26:   end for
27:   Calculate test error $e$ by soft voting;
28: end for
29: Transform the unseen instances $\mathbf{X}^*$ from 1 to 19 and
      subsequently predicting $\mathbf{X}^*$ by MI-ELM and soft voting.
```

obtained after statistics are used as the features extracted from the sample.

The efficacy of feature extraction based on BOSS is depicted in the first figure of part A in Fig. 3, illustrating the transformation of force/haptic time series into vector features.

B. Feature Fusion Based on AF Framework

The association-based fusion method [31] is a general fusion framework, which can be embedded into various existing multimodal classification problems to improve the performance of classification problems. At the same time, it is also a completely transparent explainable fusion strategy, which can better improve the explainability of the model. The fusion strategy consists of two steps: high-level information extraction and associated information extraction. Datasets after feature extraction are defined as $B = \{B^p | 1 \leq p \leq P\}$, where the nth sample of the dataset $B = \{(\mathbf{x}_n^p, y_n) | 1 \leq n \leq N)\}$ for mode p is $\mathbf{x}_n^p = [b_1^p(\mathbf{x}_n^p), b_2^p(\mathbf{x}_n^p), \ldots, b_{m_p}^p(\mathbf{x}_n^p)]$, where m_p represents the feature dimension.

First, we employ power encoding of features into the original feature space to extract higher-order information. The original m_p-dimensional input features are mapped to the m_pL-dimensional space by A mapping $\boldsymbol{\phi}_p$

$$\boldsymbol{\phi}_p(\mathbf{x}_n^p) = [h_1^p(\mathbf{x}_n^p), h_2^p(\mathbf{x}_n^p), \ldots, h_{m_pL}^p(\mathbf{x}_n^p)] \tag{4}$$

where $h_j^p(\mathbf{x}_n^p), j = l + L(k-1)$ denotes the lth power of the value b_k^p, $1 \leq l \leq L$ and $1 \leq k \leq m_p$, where k denotes the kth feature. L is the largest power of the feature. Then, the original feature dataset $B^p \in B$ is transformed into

$$H^p = \{(\boldsymbol{\phi}_p(\mathbf{x}_n^p), y_n) | (\mathbf{x}_n^p, y_n) \in B^p\}. \tag{5}$$

Second, the relational fusion matrix R^p is defined to fuse the dependent features and the relationships between them in an explicit way. For the correlation strength between any two dimensions, h_i^p and h_j^p, within the feature space, the Pearson correlation coefficient is used to calculate

$$\begin{aligned} R^p(i,j) = \rho(h_i^p, h_j^p) &= \frac{cov(h_i^p, h_j^p)}{\sigma_{h_i^p} \cdot \sigma_{h_j^p}} \\ &= \frac{E(h_i^p - \mu_{h_i^p})(h_j^p - \mu_{h_j^p})}{\sigma_{h_i^p} \cdot \sigma_{h_j^p}}. \end{aligned} \tag{6}$$

A fusion matrix $R^p \in R^{m_pL \times m_pL}$ containing all pairwise feature relations is obtained. Finally, the m_pL-dimensional higher-order information features are mapped to the m_pL-dimensional fusion feature space through a mapping $\boldsymbol{\varphi}_p$

$$\boldsymbol{\varphi}_p(\boldsymbol{\phi}_p(\mathbf{x}_n^p)) = [a_1^p(\mathbf{x}_n^p), a_2^p(\mathbf{x}_n^p), \ldots, a_{m_pL}^p(\mathbf{x}_n^p)] \tag{7}$$

where

$$\begin{aligned} a_j^p(\mathbf{x}_n^p) &= \sum_{k=1}^{m_pL} (\omega_k R_{kj} h_k^p(\mathbf{x}_n^p)) \\ &= \boldsymbol{\phi}_p(\mathbf{x}_n^p)(\omega^T \odot R_{:j}^p) \end{aligned} \tag{8}$$

and $\omega = [\omega_1, \omega_2, \ldots, \omega_{m_pL}] = [\underbrace{\frac{1}{1!}, \frac{1}{2!}, \ldots, \frac{1}{L!}, \ldots}_{m_pL}] \in R^{m_pL}$, $\odot$ denotes multiplication of the corresponding elements and $R_{:j}^p$ denotes the jth column of the relationship fusion matrix R^p. Therefore, the boosting training set H^p is transformed into

$$A^p = \{(\boldsymbol{\varphi}_p(\boldsymbol{\phi}_p(\mathbf{x}_n^p), y_n) | (\boldsymbol{\phi}_p(\mathbf{x}_n^p), y_n) \in H^p\} \tag{9}$$

Then, the radial basis kernel function is used to map the features to the higher dimensional space:

$$\begin{aligned} G^p(i,j) &= \langle \boldsymbol{\varphi}_p(\boldsymbol{\phi}_p(\mathbf{x}_i^p), \boldsymbol{\varphi}_p(\boldsymbol{\phi}_p(\mathbf{x}_j^p)) \rangle \\ &= \exp(-\gamma \| \boldsymbol{\varphi}_p(\boldsymbol{\phi}_p(\mathbf{x}_i^p) - \boldsymbol{\varphi}_p(\boldsymbol{\phi}_p(\mathbf{x}_j^p) \|_2^2). \end{aligned} \tag{10}$$

After the above three steps, a new training set $G = \{G^p | 1 \leq p \leq P\}$ is created from the original feature set B. The process of intramodal information fusion based on the AF framework and the transformation through kernel function mapping is illustrated in the second picture of part A in Fig. 3. Following the aforementioned transformation, each sample obtained from an individual finger encompasses P feature vectors. By concatenating these features, the fusion between distinct patterns is achieved, laying the groundwork for training a single-finger classifier.

C. Multifinger I-ELM Classifier

The ELM algorithm, introduced by Huang et al. [32], offers a straightforward and efficient learning approach that significantly outperforms traditional models such as BP and SVM in terms of training speed. However, the SLFNs that exhibit extreme learning effects may require a higher number of hidden nodes to achieve optimal performance [33]. Consequently, this study adopts the incremental ELM [34], and synergistically combines it with the block increment initialization technique, wherein the initial number of nodes is deliberately set to a nonzero value. For dataset above denoted as $(\mathbf{x}, \mathbf{Y})$, standard SLFN is mathematically modeled as

$$\mathbf{H}\beta = \mathbf{Y}. \tag{11}$$

$\mathbf{H}$ is called the hidden layer output matrix of the SLFNs and $\mathbf{H} = g(\mathbf{w} * \mathbf{x} + b)$, g is the activation function, β is the output weight of the hidden layer. The minimum norm least-squares solution of the above linear system is

$$\hat{\beta} = \mathbf{H}^\dagger \mathbf{Y} \tag{12}$$

where $\mathbf{H}^\dagger$ is the Moore–Penrose generalized inverse of matrix $\mathbf{H}$. The solution method suitable for any case is singular value decomposition.

The I-ELM dynamically increases the nodes in its hidden layer during training, but this process can be slow. To address this, a novel approach combines the block increment initialization method with I-ELM. Initially, the node count is set to L_0 followed by training to derive the output weights β_0 for all nodes using (11) and (12), resulting in the current training error

$$E = \mathbf{Y} - \beta_0 H_{L_0}. \tag{13}$$

Subsequently, through iterative training, a node is incrementally added, wherein the input weight and bias are randomly initialized. The output of the newly added hidden node $\mathbf{H}_{\tilde{L}}$ is calculated and the output weight of this node is determined as

$$\beta_{\tilde{L}} = \frac{E \cdot H_{\tilde{L}}^{T}}{H_{\tilde{L}} \cdot H_{\tilde{L}}^{T}}. \tag{14}$$

Then, training error E updating by

$$E = E - \beta_{\tilde{L}} H_{\tilde{L}}. \tag{15}$$

By repeatedly adding nodes and obtaining corresponding output weights according to (14), and updating training errors according to (15) until the conditional loop is completed, a single-finger-based classifier can be built. The schematic diagram for the training of a single I-ELM is depicted in the last picture of part A in Fig. 3.

D. Integrated Decision

According to our method, a classifier needs to be trained on each finger, so the final classification needs to be made using an integrated decision, in which case a soft voting mechanism is applied, which can improve the accuracy and robustness of the overall prediction. For any I-ELM-i, for which the corresponding output is represented as $\text{pre}_i = [p_{i1}, p_{i2}, \ldots, p_{io}]$, the resulting voting outcome is outlined as follows:

$$\begin{aligned} \text{Prediction} &= \frac{1}{M} \sum_{m=1}^{M} \text{pre}_i \\ &= \frac{1}{M} \left[\sum_{m=1}^{M} \text{pre}_{m1}, \sum_{m=1}^{M} \text{pre}_{m2}, \ldots, \sum_{m=1}^{M} \text{pre}_{mo} \right] \end{aligned} \tag{16}$$

where o is the number of objects identified and M is the number of fingers. Finally, according to the principle of maximum probability, the subscript value of the object class corresponding to the number with the largest value is selected as the category label obtained by the decision. The complete procedure of BOSS-MI-ELM is shown in Algorithm 1.

IV. Experimental

A. Database and the Experimental Setting

Following the equipment overview presented in Section II, a series of grasping experiments were conducted involving nine common objects. The recorded data encompass the respective grasping angle, angular velocity, acceleration, and strains of three flexible fingers during the object-grasping process. The grasping speed parameter was set at 25 (controller speed), and due to the disparate sampling frequencies between FBG and IMU, the FBG signals underwent downsampling. Furthermore, each grasp's FBG and IMU data were subjected to two separate phase-based truncations to preserve the sensor signals during the "grasp," "hold," and "release" phases of object manipulation. A total of 1694 samples were generated, with a time step interval of 130. We set the samples of each type of object with corresponding object labels, and hope that the multimodal perception algorithm described above can recognize the corresponding object labels of each sample. For example, sample data collected by grasping a banana can be recognized as "banana."

TABLE I
Search Ranges of the Hyperparameters

Number of searches	Hyperparameter	Search range
150	w	10, 12, $\cdots$, 26
	l	2, 3, $\cdots$, 6
	a	2, 3, $\cdots$, 6
	L	1, 2, 3, 4

Fig. 4 displays the nine objects utilized in the dataset, along with the representative sample data for three of these objects. Objects are grasped standing up on the table for objects that can stand up, such as bottles and paper cups. Slender objects that cannot stand up, such as bananas, are placed flat on the table. During this time, the pose of the objects was adjusted and the grasping depth was slightly altered to obtain a wider range of haptic information. The objects are all grasped from top to bottom by the claw.

In the data preprocessing stage, we utilized the Z-score normalization method to process multidimensional time series features, ensuring comparability across different scales. For a sample data x, this process is expressed as follows:

$$x^* = \frac{x - \overline{x}}{\sigma}. \tag{17}$$

where $\overline{x}$ and σ represent the mean and standard deviation of the raw data, respectively. Through this procedure, each time series was mapped onto a standard normal distribution with a mean of 0 and a standard deviation of 1, effectively eliminating scale differences among different time series.

The dataset was systematically processed by Algorithm 1, and an extensive hyperparameter search was conducted to enhance performance. The search parameters and their specified ranges are detailed in Table I, with the parameters $L_0 = 50$, $L_{\max} = 300$ and $E^* = 10^{-3}$ held constant. After the search process, the relevant hyperparameters are retained for use in the feature extraction and fusion procedures. Subsequently, starting with an initial selection of 10% data volume from the complete dataset, the percentage was incremented by 10% in each iteration until reaching 100%. At each stage, a rigorous 10-fold cross-validation was performed.

B. Compared Methods

Our approach was subjected to comparison with several benchmark models commonly applied in object-grasping datasets, including MLP, LSTM, CNN, CNNLSTM, ConvLSTM, and D-CNN. Detailed network architectures can be referenced from the source papers of the object grasping datasets [35]. Notably, Maus et al. [35] had previously undertaken a hyperparameter exploration of these models using the grasping dataset. Consequently, we limited our hyperparameter search efforts to their dataset, adopting the same model and model parameter utilized in their comparative experiments. Moreover, the outcomes of the hyperparameter search were directly applied to both the cutting dataset [36] and the dataset we generated internally.

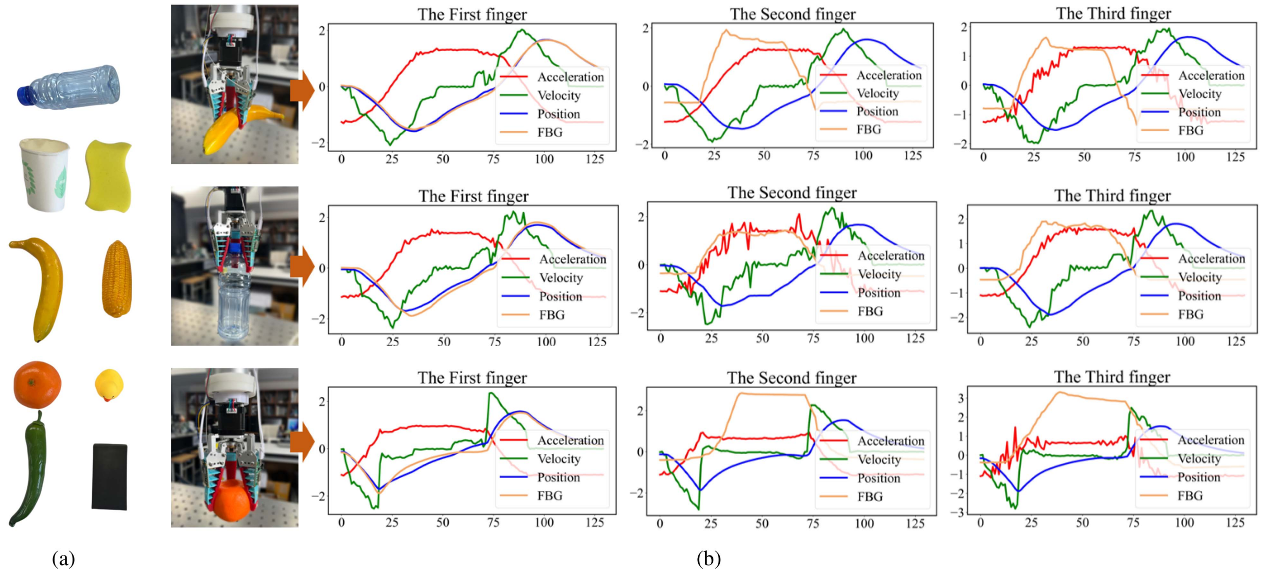


Fig. 4. (a) Nine objects used for grasping. These things are in turn: plastic bottle, paper cup, sponge, banana, corn, orange, toy duck, green pepper, and cardboard box. (b) The left three pictures are the grasped objects in our datasets and the three pictures to the right of each grasped object are its corresponding three-finger force haptic data. All data are Z-score normalized.

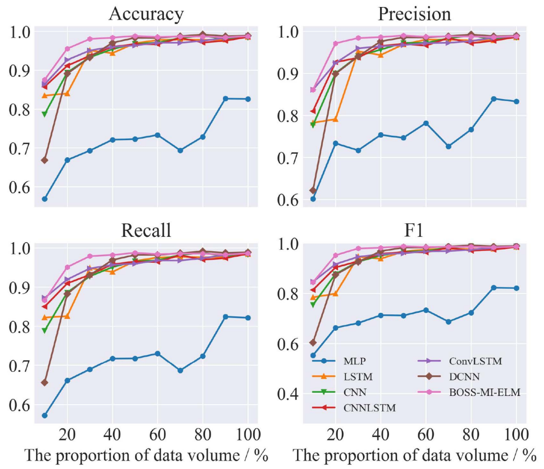


Fig. 5. Performance of various methods on different data volumes on our dataset.

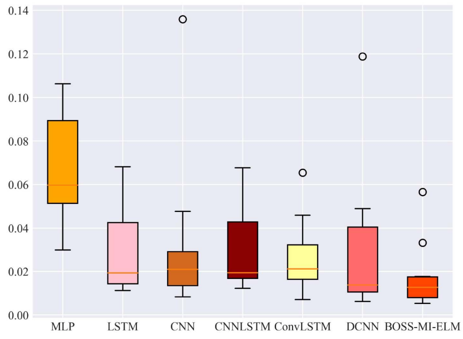


Fig. 6. Accuracy standard deviation box plot of various algorithms when using different data amounts on our dataset to do 10-fold cross-validation.

C. Performance Comparison of Multi-Modal Grasping Target Recognition

Following a 10-fold cross-validation in the grasping dataset, the results of the hyperparameter search are as follows: $w = 26$, $l = 4$, $a = 3$, and $L = 1$.

The experimental results of all models on our dataset have been encapsulated in Table II. In addition, corresponding line charts are illustrated in Fig. 5. The outcomes from the table distinctly demonstrate that our proposed method attains top rankings across the four evaluation metrics, namely accuracy, precision, recall, and F1-score, with 6, 6, 5, and 5 subsets securing the first position, respectively. Moreover, as illustrated in Fig. 5, even in cases where the method does not secure the top rank in individual performance metrics, it consistently maintains a highly competitive standing. And it can be found that BOSS-MI-ELM is the outstanding performance when the amount of data is less. These results emphasize the effectiveness of the method on haptic perception and its excellence on small datasets. It is noteworthy that with a reduction in data volume, the classification performance of the D-CNN model experiences a sharp decline. This phenomenon provides further insights into the constraints faced by deep learning models when handling small datasets.

In addition, we meticulously documented the standard deviation of classification accuracy across tenfold cross-validation iterations and conducted a box plot analysis, as illustrated in Fig. 6. The results of the box plot indicate that within the framework of multiple cross-validation iterations, the ensemble of standard deviations for BOSS-MI-ELM exhibits a narrower and more compact distribution. This pattern suggests that, in comparison to alternative algorithms, this approach demonstrates heightened stability across each training iteration.

TABLE II
EXPERIMENTAL RESULTS OF ALL ALGORITHMS UNDER DIFFERENT DATA PROPORTIONS IN OUR DATASET

Performance	Scale[1]	MLP	LSTM	CNN	CNN-LSTM	Conv-LSTM	DCNN	BOSS-MI-ELM
Accuracy	100%	0.8258	0.9852	0.9847	0.9858	0.9882	**0.9888**	0.9876
	90%	0.8268	0.9784	0.9803	0.9757	0.9830	**0.9875**	0.9843
	80%	0.7285	0.9779	0.9882	0.9712	0.9756	**0.9919**	0.9867
	70%	0.6937	0.9789	0.9831	0.9831	0.9705	**0.9873**	0.9865
	60%	0.7332	0.9774	0.9704	0.9665	0.9705	0.9833	**0.9852**
	50%	0.7227	0.9693	0.9681	0.9670	0.9634	0.9835	**0.9882**
	40%	0.7208	0.9438	0.9543	0.9601	0.9601	0.9706	**0.9837**
	30%	0.6928	0.9508	0.9332	0.9371	0.9507	0.9331	**0.9803**
	20%	0.6686	0.8400	0.8939	0.9113	0.9263	0.8908	**0.9554**
	10%	0.5688	0.8346	0.7864	0.8574	0.8636	0.6684	**0.8754**
Precision	100%	0.8332	0.9857	0.9858	0.9869	0.9882	**0.9890**	0.9883
	90%	0.8394	0.9783	0.9808	0.9772	0.9845	**0.9882**	0.9859
	80%	0.7661	0.9781	0.9895	0.9713	0.9766	**0.9922**	0.9875
	70%	0.7262	0.9800	0.9834	0.9840	0.9724	**0.9882**	0.9879
	60%	0.7817	0.9800	0.9721	0.9659	0.9715	0.9852	**0.9867**
	50%	0.7465	0.9687	0.9720	0.9696	0.9672	0.9853	**0.9899**
	40%	0.7535	0.9434	0.9563	0.9653	0.9640	0.9760	**0.9859**
	30%	0.7168	0.9523	0.9410	0.9369	0.9594	0.9400	**0.9840**
	20%	0.7333	0.7909	0.9001	0.9263	0.9244	0.8989	**0.9712**
	10%	0.6015	0.7822	0.7761	0.8100	**0.8607**	0.6216	**0.8607**
Recall	100%	0.8218	0.9839	0.9833	0.9849	0.9875	**0.9887**	0.9861
	90%	0.8242	0.9767	0.9792	0.9740	0.9819	**0.9872**	0.9827
	80%	0.7239	0.9760	0.9874	0.9697	0.9735	**0.9911**	0.9854
	70%	0.6865	0.9772	0.9824	0.9815	0.9677	**0.9863**	0.9852
	60%	0.7305	0.9753	0.9681	0.9641	0.9681	0.9821	**0.9839**
	50%	0.7181	0.9665	0.9660	0.9649	0.9595	0.9819	**0.9873**
	40%	0.7177	0.9385	0.9509	0.9579	0.9567	0.9692	**0.9819**
	30%	0.6896	0.9485	0.9274	0.9315	0.9474	0.9304	**0.9789**
	20%	0.6611	0.8259	0.8861	0.9093	0.9194	0.8824	**0.9509**
	10%	0.5722	0.8222	0.7889	0.8500	**0.8722**	0.6556	0.8667
F1	100%	0.8208	0.9842	0.9840	0.9849	0.9876	**0.9882**	0.9867
	90%	0.8228	0.9765	0.9793	0.9740	0.9817	**0.9869**	0.9833
	80%	0.7228	0.9761	0.9879	0.9696	0.9734	**0.9912**	0.9860
	70%	0.6871	0.9772	0.9824	0.9818	0.9684	**0.9863**	0.9858
	60%	0.7333	0.9747	0.9686	0.9633	0.9677	0.9823	**0.9846**
	50%	0.7115	0.9666	0.9663	0.9644	0.9595	0.9823	**0.9875**
	40%	0.7127	0.9376	0.9504	0.9573	0.9560	0.9680	**0.9816**
	30%	0.6814	0.9460	0.9265	0.9277	0.9455	0.9251	**0.9790**
	20%	0.6627	0.7984	0.8770	0.9033	0.9163	0.8738	**0.9517**
	10%	0.5527	0.7845	0.7543	0.8138	**0.8460**	0.6037	0.8442

[1] This ”Scale” represents the percentage of the overall dataset used for training and testing.
The bold formatting to highlight the optimal results.

TABLE III
10-FOLD CROSS-VALIDATION RESULTS WITH INTRODUCED GAUSSIAN NOISE

Performance	SD[1]	MLP	LSTM	CNN	CNN-LSTM	Conv-LSTM	DCNN	BOSS-MI-ELM
Accuracy	0.05	0.4050	0.8311	0.9529	0.8762	0.9112	0.9084	**0.9617**
	0.1	0.3608	0.8047	0.9470	0.8727	0.9228	0.8665	**0.9587**
	0.15	0.3342	0.8367	0.9114	0.8786	0.9322	0.8817	**0.9558**
Precision	0.05	0.3986	0.7967	0.9585	0.8667	0.8974	0.9062	**0.9707**
	0.1	0.2961	0.8006	0.9583	0.8463	0.9049	0.9002	**0.9691**
	0.15	0.2916	0.8156	0.9247	0.8608	0.9230	0.8659	**0.9648**
Recall	0.05	0.3963	0.8167	0.9491	0.8620	0.9009	0.8972	**0.9611**
	0.1	0.3556	0.7880	0.9407	0.8630	0.9093	0.8630	**0.9565**
	0.15	0.3250	0.8278	0.9046	0.8685	0.9194	0.8759	**0.9537**
F1	0.05	0.3657	0.7982	0.9488	0.8490	0.8911	0.8916	**0.9599**
	0.1	0.2946	0.7737	0.9404	0.8468	0.8993	0.8511	**0.9572**
	0.15	0.2829	0.8049	0.9027	0.8538	0.9138	0.8579	**0.9530**

[1] This ”SD” represents the standard deviation of the added Gaussian noise.

TABLE IV
10-FOLD CROSS-VALIDATION RESULTS WITH 40% DATA ANOMALIES

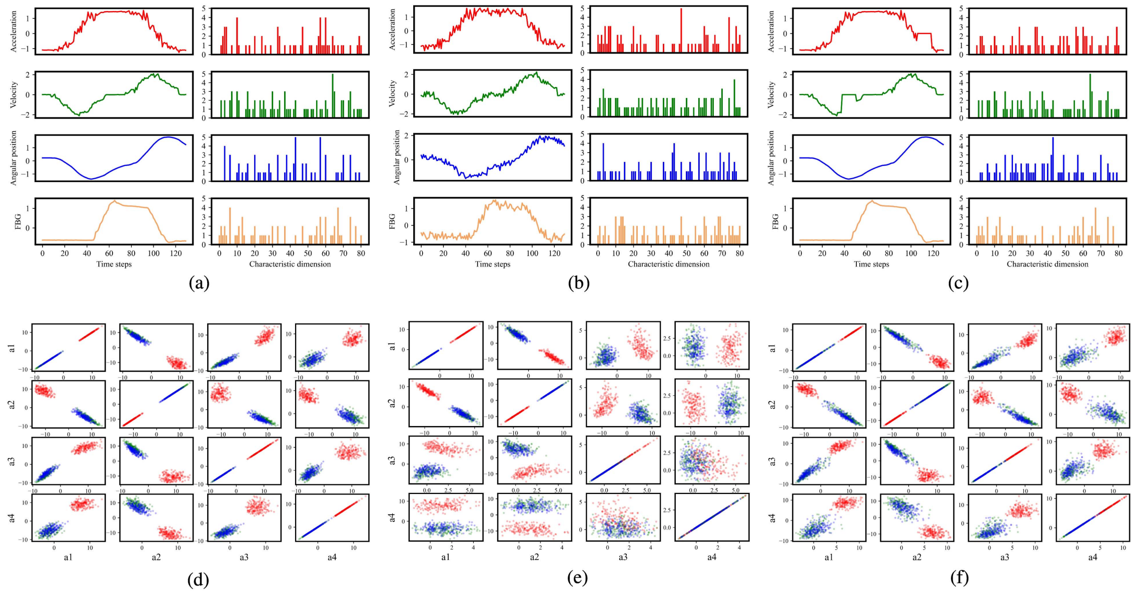
Performance	Scale[1]	MLP	LSTM	CNN	CNN-LSTM	Conv-LSTM	DCNN	BOSS-MI-ELM
Accuracy	2%	0.5590	0.8522	0.9499	0.8702	0.9170	0.8575	**0.9556**
	5%	0.5444	0.8078	0.9291	0.8816	0.9112	0.8344	**0.9318**
	10%	0.6004	0.7987	0.9024	0.8312	0.9053	0.8430	**0.9111**
Precision	2%	0.5710	0.8105	0.9615	0.8604	0.9030	0.8735	**0.9671**
	5%	0.5670	0.7680	0.9481	0.8544	0.9032	0.8802	**0.9536**
	10%	0.6187	0.7734	0.9087	0.8260	0.9069	0.8501	**0.9223**
Recall	2%	0.5463	0.8389	0.9454	0.8565	0.9056	0.8500	**0.9546**
	5%	0.5278	0.7935	0.9250	0.8676	0.9019	0.8287	**0.9287**
	10%	0.5833	0.7907	0.8935	0.8176	**0.8981**	0.8315	**0.8981**
F1	2%	0.5165	0.8115	0.9444	0.8407	0.8948	0.8363	**0.9526**
	5%	0.4995	0.7618	0.9219	0.8517	0.8958	0.8216	**0.9265**
	10%	0.5599	0.7648	0.8877	0.8079	**0.8950**	0.8172	0.8908

[1] This ”Scale” pertains to the proportion of time steps set to zero in the selected time series randomly sampled from the dataset, relative to the total number of time steps.
The bold formatting to highlight the optimal results.

D. Algorithm Robustness Analysis

To assess the algorithm’s robustness, further experiments were conducted on our dataset. A subset comprising 20% of the data volume was extracted to create a small dataset. The small dataset underwent both noise injection and data missing procedures to assess the algorithm’s resilience in coping with perturbations within constrained data environments. In the noise experiment, Gaussian noise was applied three times with a mean of zero and varying standard deviations. In the data missing experiment, 40% of the time series within the small dataset had specific-length subsequences zeroed, repeated three times with variations in the selected lengths. The results of these experiments are delineated in Tables III and IV. The experimental results show that our method performs best with both data anomalies, topping all performance metrics in almost every experiment. This demonstrates the robustness of the method in solving data anomalies in robot grasping target recognition.

To examine the effect of noise and data loss on the extracted and fused features, we selected a sample to show its feature extraction results under raw, noise injection, and data loss. Three types of recognition types in the dataset are selected to count their distribution relationship between the first four features, two by two after all sample features are fused. As shown in Fig. 7. From the BOSS feature maps, it is evident that, in both anomalous scenarios, only a small subset of features undergo numerical changes. In the AF feature space, the overlapping of points with different colors indicates that the corresponding recognition types cannot be distinguished under these two features. In both types of anomalies, the feature distributions become more dispersed and these features are shifted after noise injection. However, the clustering relationship of points of the same color does not change significantly. It indicates that the samples that are recognizable in the original data can still be recognized when the data are anomalous. Even some points of different colors transition from highly clustered state



(a) (b) (c)

(d) (e) (f)

Fig. 7. (a), (b), and (c) shows the data and these features under normal conditions and two abnormal conditions. (d), (e), and (f) shows the two-by-two distribution of the first four features (a1, a2, a3, a4) in AF space for all samples of the three recognition types. The parameters of the two abnormal cases are SD=0.15 and Scale=15%, respectively. (a) The raw data and BOSS features. (b) The adding noise data and BOSS features. (c) The missing data and BOSS features. (d) Normal conditions. (e) Noise injections. (f) Data missing.

TABLE V
FRIEDMAN TEST RESULTS ON FOUR TYPES OF PERFORMANCE

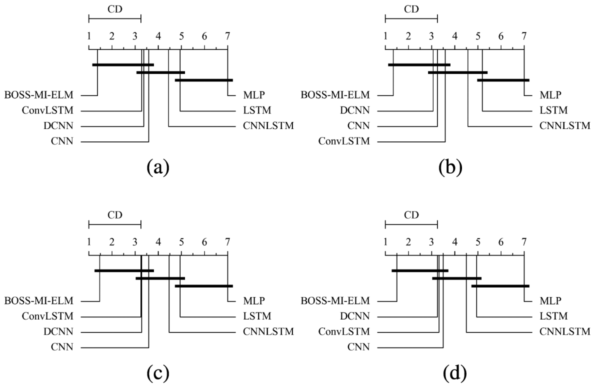
Performance	τ_F	Critical value ($\alpha = 0.05$)
Accuracy	27.7006	2.6572
Precision	33	
Recall	25.8977	
F1	25.6349	

to distinguishable state due to data anomalies. This phenomenon is likely to be the reason for the improved classification performance under anomalies and explains the use of data anomalies as a data enhancement method in many tactile perception studies.

To further assess the statistical significance of the classification performance differences, the Friedman test [37] was employed. Considering varied data quantities and the treatment of datasets manipulated with noise injection and data missing as distinct entities, it is evident that $k_g = 7$ and $N_D = 16$. Detailed results of the analysis can be found in Table V.

Significant differences exist for τ_F greater than the critical value. The results demonstrate notable disparities in the four performances exhibited by each model. Therefore, to gain a deeper understanding of the comparative performance of these algorithms, we also conducted the Nemenyi post hoc test (shown in Fig. 8).

In this context, the position of each algorithm's leading line corresponds to its ranking. A higher position in the ranking indicates the superior performance of the algorithm. The findings of the detection results unveil the outstanding performance of the proposed methodology in this study, securing the top position in all four performance metrics, albeit not uniformly surpassing other algorithms in a statistically significant manner.



(a) (b)

(c) (d)

Fig. 8. Four categories of performance indicators on the CD chart. Smaller numbers indicate higher rankings. There is no significant difference between algorithms that are traversed by horizontal lines. (a) Accuracy. (b) Precision. (c) Recall. (d) F1.

E. Generalized Performance Verification

To verify whether the recognition performance remains good under different grasping velocities. We collected a small dataset of 654 samples at three additional grasping velocities. Ten-fold cross-validations of our model were performed. The results are shown in Fig. 9. It can be seen that our algorithm still performs well with all four performances above 90%.

Furthermore, the analysis incorporated two publicly accessible force-tactile datasets. Grasping dataset [35]: It includes force and hand grip position information to recognize objects. A total of 2000 samples with a time step of 50. In addition, it incorporates an independent validation set consisting of 120 samples collected by various experimenters to reflect real-world data scenarios. Cutting dataset [36]: It contains force, position,

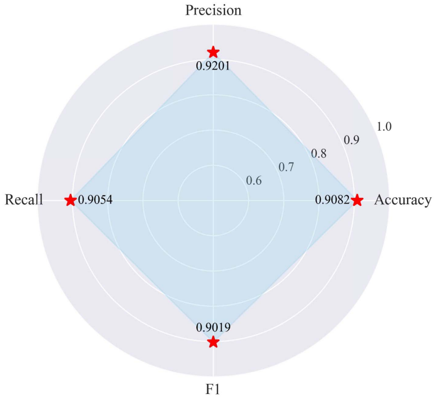


Fig. 9. Recognition performance of BOSS-MI-ELM at different grasping velocities.

TABLE VI
EXPERIMENTAL RESULTS OF ALL ALGORITHMS UNDER DIFFERENT DATA PROPORTIONS IN GRASPING DATASET AND CUTTING DATASET

Dataset	Scale[1]	MLP	LSTM	CNN	CNN-LSTM	Conv-LSTM	DCNN	BOSS-MI-ELM
Grasping dataset	100%	0.6500	0.6675	0.6617	0.6733	0.6625	0.6867	**0.7117**
	90%	0.6550	0.6658	0.6742	0.6742	0.6717	0.6817	**0.7100**
	80%	0.6567	0.6642	0.6508	0.6600	0.6658	0.6758	**0.7000**
	70%	0.6458	0.6683	0.6608	0.6550	0.6567	0.6692	**0.7000**
	60%	0.6558	0.6558	0.6375	0.6558	0.6592	0.6617	**0.6825**
	50%	0.6392	0.6517	0.6667	0.6425	0.6575	0.6650	**0.6750**
	40%	0.6133	0.6242	0.6367	0.6308	0.6367	0.5958	**0.6492**
	30%	0.6183	0.6067	0.6042	0.6150	0.6117	0.4975	**0.6642**
	20%	0.6000	0.5792	0.6108	0.6117	0.6150	0.4400	**0.6292**
	10%	0.5725	0.5225	0.5692	0.5717	0.5700	0.3900	**0.5975**
Cutting dataset	100%	**0.8775**	0.8747	0.8770	**0.8775**	**0.8775**	0.8753	**0.8775**
	90%	**0.8777**	0.8764	0.8770	**0.8777**	0.8752	0.8739	**0.8777**
	80%	**0.8778**	0.8764	0.8736	**0.8778**	0.8750	0.8764	**0.8778**
	70%	**0.8772**	**0.8772**	0.8724	**0.8772**	**0.8772**	0.8764	**0.8772**
	60%	**0.8773**	0.8764	0.8727	**0.8773**	0.8727	0.8680	**0.8773**
	50%	0.8775	**0.8787**	0.8764	0.8775	**0.8787**	0.8764	0.8775
	40%	**0.8778**	0.8764	0.8764	**0.8778**	**0.8778**	0.8666	**0.8778**
	30%	0.8783	**0.8802**	0.8671	0.8783	**0.8802**	0.8690	0.8783
	20%	0.8765	**0.8821**	0.8737	0.8765	0.8595	0.8453	0.8765
	10%	0.8709	0.8542	0.8709	**0.8765**	0.8595	0.7876	**0.8765**

[1] This ”Scale” represents the percentage of the overall dataset used for training and testing.
The bold formatting to highlight the optimal results.

and velocity control input information for predicting the cut state. The raw data are subdivided into 50 time steps per frame for a total of 1780 samples.

Subsequently, we conducted experiments with incremental data volumes on the two publicly available datasets, and the classification accuracy for both experiments is presented in Table VI. The results indicate that in the 10 experiments on the grasping dataset, our method consistently claimed the top position. On the cutting dataset, while the numerical differences in classification accuracy among various models were marginal, our algorithm still outperformed others in seven experiments. These findings affirm the excellent performance and robust generalization of our method on both datasets.

However, the classification performance on the two public datasets is significantly lower than on our dataset. For the grasping dataset and the cutting dataset, we attribute this to the result of having little modal data and no multifinger data, respectively. To verify the conjecture, we add two experiments: one is to use only some of the four modalities on our dataset, and the other is to use only single-finger data for training on our dataset. The results of the experiments are shown in Fig. 10, which confirms that our idea is correct. It can also be found that with the gradual addition of the three types of signals (P, V, A) captured by the IMU, the final perceptual performance of the system rises robustly, further illustrating the effectiveness of combining the IMU with the force-tactile sensing that we have implemented using the FBG.

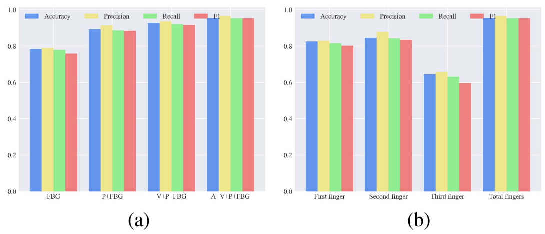


Fig. 10. Experimental results of BOSS-MI-ELM with different combinations of modal data and different finger data. The dataset uses 20% of our original dataset with Gaussian noise added with a standard deviation of 0.15. In the figure on the left, P represents angle, V represents angular velocity, and A represents acceleration. (a) Classification performance with in different combinations of modal data. (b) Classification performance with different finger data.

TABLE VII
ALGORITHM REAL-TIME VALIDATION

Model	MLP	LSTM	CNN	CNN-LSTM	Conv-LSTM	DCNN	BOSS-MI-ELM
Time (s)	0.0459	0.2206	0.0601	0.2522	0.1595	0.1132	0.6837

Finally, the real-time performance of the various methods was verified. For the 10 cross-validations on 20% of our original data, the time taken for the test data to go from the input to the model to the model outputting the corresponding labels was calculated, and the average time for the last nine validations was calculated (the time for the first validation usually includes the time to start the computer equipment). We used the Python language and ran it on Intel Core i7-13700KF. The results are shown in Table VII. The real-time performance of our method is 0.6837 s, which can be used for the process of human–computer interaction. Of course, this time will be shorter when using GPU or switching to C++ language.

From sliding and grasping balls of different sizes on different surfaces to considering the effect of data volume versus grasping speed on the perceptual ability of the model, as well as the robustness of the model in the presence of data anomalies, and the final validation on a public dataset, these results reveal that our system can improve haptic-based human–computer interaction.

V. CONCLUSION

This study establishes a robotic hand platform, marking the first integration of FBG and IMU on a flexible robotic hand. A method for identifying the grasping targets of a multifingered robotic hand is proposed, which includes feature extraction based on the BOSS algorithm, intramodal feature fusion based on the AF framework, and the application of a multifinger I-ELM classifier. A dataset is constructed based on the platform, and through extensive experiments, the reliability and robustness of the proposed method in situations with limited samples and data anomalies are demonstrated. In addition, the method is validated for recognizing grasped objects under different hand speeds. After obtaining satisfactory validation results on a public dataset, the experiments are extended based on these results, further confirming the effectiveness of multimodal data and the integration of multifinger models. Finally, the real-time performance of the algorithm is verified. Comprehensive experiments validate the feasibility and robustness of the proposed method in robotic grasping target recognition, contributing to the improvement of touch-based human–machine interaction.

In future work, our foremost objective entails the conversion of strain data from FBG into specific contact force values through rigorous calibration. Subsequently, we aim to incorporate additional tactile sensing modalities to enhance the richness of tactile information during object grasping. Last, our strategic plan encompasses the design and integration of a miniature visual sensor, culminating in the realization of a comprehensive vision-tactile fusion-based robotic flexible gripper for robust object recognition during manipulation.

References

[1] D. Rus and M. T. Tolley, "Design, fabrication and control of soft robots," *Nature*, vol. 521, no. 7553, pp. 467–475, 2015.

[2] H. Wang, M. Totaro, and L. Beccai, "Toward perceptive soft robots: Progress and challenges," *Adv. Sci.*, vol. 5, no. 9, 2018.

[3] S. Yu, D.-H. Zhai, and Y. Xia, "Robotic grasp detection based on category-level object pose estimation with self-supervised learning," *IEEE/ASME Trans. Mechatron.*, vol. 29, no. 1, pp. 625–635, Feb. 2024.

[4] S. Yu, D.-H. Zhai, and Y. Xia, "CGNet: Robotic grasp detection in heavily cluttered scenes," *IEEE/ASME Trans. Mechatron.*, vol. 28, no. 2, pp. 884–894, Apr. 2023.

[5] K. Song, J. Wang, Y. Bao, L. Huang, and Y. Yan, "A novel visible-depth-thermal image dataset of salient object detection for robotic visual perception," *IEEE/ASME Trans. Mechatron.*, vol. 28, no. 3, pp. 1558–1569, Jun. 2023.

[6] H. Cao, G. Chen, Z. Li, Q. Feng, J. Lin, and A. Knoll, "Efficient grasp detection network with gaussian-based grasp representation for robotic manipulation," *IEEE/ASME Trans. Mechatron.*, vol. 28, no. 3, pp. 1384–1394, Jun. 2023.

[7] J. Jiang, G. Cao, A. Butterworth, D. Thanh-Toan, and S. Luo, "Where shall I touch? vision-guided tactile poking for transparent object grasping," *IEEE/ASME Trans. Mechatronics*, vol. 28, no. 1, pp. 233–244, Feb. 2023.

[8] L. Wang, Q. Li, J. Lam, and Z. Wang, "Tactual recognition of soft objects from deformation cues," *IEEE Robot. Automat. Lett.*, vol. 7, no. 1, pp. 96–103, Jan. 2022.

[9] P. Xiong, K. He, E. Q. Wu, L. -M. Zhu, A. Song, and P. X. Liu, "Human-exploratory-Procedure-Based hybrid measurement fusion for material recognition," *IEEE/ASME Trans. Mechatron.*, vol. 27, no. 2, pp. 1093–1104, Apr. 2022.

[10] R. S. Johansson and J. R. Flanagan, "Coding and use of tactile signals from the fingertips in object manipulation tasks," *Nat. Rev. Neurosci.*, vol. 10, no. 5, pp. 345–359, 2009.

[11] S. Sundaram et al., "Learning the signatures of the human grasp using a scalable tactile glove," *Nature*, vol. 569, no. 7758, pp. 698–702, 2019.

[12] P. Zhang, G. Yu, D. Shan, Z. Chen, and X. Wang, "Identifying the strength level of objects' tactile attributes using a multi-scale convolutional neural network," *Sensors*, vol. 22, no. 5, 2022, Art. no. 1908.

[13] J. Lin, R. Calandra, and S. Levine, "Learning to identify object instances by touch: Tactile recognition via multimodal matching," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 3644–3650.

[14] Y. Xia et al., "Beta mixture model for the uncertainties in robotic haptic object identification," *IEEE/ASME Trans. Mechatron.*, vol. 27, no. 4, pp. 1955–1963, Aug. 2022.

[15] V. Chu et al., "Robotic learning of haptic adjectives through physical interaction," *Robot. Auton. Syst.*, vol. 63, pp. 279–292, 2015.

[16] T. Bhattacharjee, H. M. Clever, J. Wade, and C. C. Kemp, "Multimodal tactile perception of objects in a real home," *IEEE Robot. Automat. Lett.*, vol. 3, no. 3, pp. 2523–2530, Jul. 2018.

[17] J. Bai, B. Li, H. Wang, and Y. Guo, "Tactile perception information recognition of prosthetic hand based on DNN-LSTM," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 2513310.

[18] T. Jin et al., "Triboelectric nanogenerator sensors for soft robotics aiming at digital twin applications," *Nat. Commun.*, vol. 11, no. 1, 2020, Art. no. 5381.

[19] R. Zuo, Z. Zhou, B. Ying, and X. Liu, "A soft robotic gripper with anti-freezing ionic hydrogel-based sensors for learning-based object recognition," in *Proc IEEE Int. Conf. Robot. Autom.*, 2021, pp. 12164–12169.

[20] S. Q. Liu and E. H. Adelson, "GelSight fin ray: Incorporating tactile sensing into a soft compliant robotic gripper," in *Proc. IEEE Int. Conf. Soft Robot.*, 2022, pp. 925–931.

[21] M. Kerzel et al., "Neuro-robotic haptic object classification by active exploration on a novel dataset," in *Proc. Int. Joint Conf. Neural Netw*, 2019, pp. 1–8.

[22] R. P. Babadian, K. Faez, M. Amiri, and E. Falotico, "Fusion of tactile and visual information in deep learning models for object recognition," *Inf. Fusion*, vol. 92, pp. 313–325, 2023.

[23] C. Lyu et al., "Three-fingers FBG tactile sensing system based on squeeze-and-excitation LSTM for object classification," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 7004611.

[24] Q. Liang et al., "Multi-component FBG-Based force sensing systems by comparison with other sensing technologies: A review," *IEEE Sens. J.*, vol. 18, no. 18, pp. 7345–7357, Sep. 2018.

[25] J. Long, Q. Liang, W. Sun, Y. Wang, and D. Zhang, "Ultrathin three-axis FBG wrist force sensor for collaborative robots," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 3519115.

[26] T. G. Thuruthel, B. Shih, C. Laschi, and M. T. Tolley, "Soft robot perception using embedded soft sensors and recurrent neural networks," *Sci. Robot.*, vol. 4, no. 26, 2019, Art. no. eaav1488.

[27] H. Soleimani, J. Hensman, and S. Saria, "Scalable joint models for reliable uncertainty-aware event prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1948–1963, Aug. 2018.

[28] P. Schaefer, "The BOSS is concerned with time series classification in the presence of noise," *Data Mining Knowl. Discov.*, vol. 29, no. 6, pp. 1505–1530, 2015.

[29] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh, "The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances," *Data Mining Knowl. Discov.*, vol. 31, no. 3, pp. 606–660, 2017.

[30] P. Schäfer and M. Högqvist, "SFA: A symbolic fourier approximation and index for similarity search in high dimensional datasets," in *Proc. Int. Conf. Extending Database Technol.*, 2012, pp. 516–527.

[31] X. Liang, Y. Qian, Q. Guo, H. Cheng, and J. Liang, "AF: An association-based fusion method for multi-modal classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9236–9254, Dec. 2022.

[32] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks," in *Proc. Int. Joint Conf. Neural Netw.*, 2004, pp. 985–990.

[33] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1/3, pp. 489–501, 2006.

[34] G.-B. Huang, L. Chen, and C.-K. Siew, "Universal approximation using incremental constructive feedforward networks with random hidden nodes," *IEEE Trans. Neural Netw.*, vol. 17, no. 4, pp. 879–892, Jul. 2006.

[35] P. Maus, J. Kim, O. Nocentini, M. Z. Bashir, and F. Cavallo, "The impact of data augmentation on tactile-based object classification using deep learning approach," *IEEE Sens. J.*, vol. 22, no. 14, pp. 14574–14583, Jul. 2022.

[36] I. Mitsioni et al., "Interpretability in contact-rich manipulation via kinodynamic images," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 10175–10181.

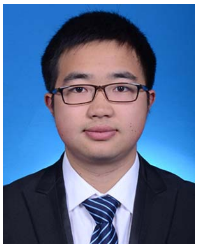
[37] J. Demar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.

Qiaokang Liang (Senior Member, IEEE) received the Ph.D. degree in control science and engineering from the University of Science and Technology of China, Hefei, China, in 2011.

He is currently a Professor with the College of Electrical and Information Engineering, Hunan University, Changsha, China, where he is the Vice Director with the Hunan Key Laboratory of Intelligent Robot Technology in Electronic Manufacturing. His research interests include robotics and mechatronics, biomimetic sensing, advanced robot technology, and human–computer interaction.

Wenxing Xiao received the B.S degree in automation, in 2022, from Huazhong Agricultural University, Wuhan, China, where he is currently working toward the M.S degree in electronic information with the College of Electrical and Information Engineering, Hunan University, Changsha, China.

His research interests include machine vision and robotics systems.



Jianyong Long (Member, IEEE) received the M.S. degree in agricultural informatization from the College of Information and Computer Engineering, Northeast Forestry University, Harbin, China, in 2017, and the Ph.D. degree in control science and engineering from Hunan University, Changsha, China, in 2022.

His main research interests are in machine learning, robotics, and mechatronics.

Dan Zhang (Senior Member, IEEE) received the Ph.D. degree in mechanical engineering from Laval University, Quebec City, QC, Canada, in June 2000.

He is a Chair Professor of intelligent robotics and automation with The Hong Kong Polytechnic University, Hong Kong. His research interests include robotics and mechatronics, high-performance parallel robotic machine development, micromanipulation/ nanomanipulation and MEMS devices, and rehabilitation robots and rescue robots. Dr. Zhang is a fellow of the Canadian Academy of Engineering (CAE), the Engineering Institute of Canada (EIC), the American Society of Mechanical Engineers (ASME), and the Canadian Society for Mechanical Engineering (CSME), and a Senior Member of SME