Granular Ball Computing-Based Large Margin Distribution Machine

1st Guangming Lang[a,b], 2nd Tingyao Di[a,b], 3rd Jie Zhou[a,b], 4th Qimei Xiao[a,b,*]
[a]School of Mathematics and Statistics, Changsha University of Science and Technology, Changsha, Hunan 410114, P.R. China
[b]Hunan Provincial Key Laboratory of Mathematical Modeling and Analysis in Engineering,
Changsha University of Science and Technology, Changsha, Hunan 410114, P.R. China
*Corresponding author E-mail address: qimeixiao@sohu.com

***Abstract*—The Large Margin Distribution Machine (LDM) and its variants have received considerable attention due to their superior generalization performance. However, they encounter significant hurdles regarding computational efficiency and noise robustness when handling complex, large-scale data. To overcome these limitations and further enhance model performance, this paper proposes the Granular Ball Large Margin Distribution Machine (GBLDM). First, we develop a dual-strategy granular ball generation mechanism designed to reconcile label consistency with geometric compactness. Specifically, a coarse-grained strategy extracts the topological backbone using high-purity "Important Granular Balls," while a fine-grained strategy refines complex boundaries via "Edge Granular Balls." Second, we construct an adaptive weighted optimization model integrating granular ball geometric information, effectively extending margin distribution theory into the granular ball space. By dynamically allocating weights based on the scale and type of granular balls, GBLDM explicitly optimizes global margin statistics at the granular ball level. Finally, extensive experiments on eight UCI benchmark datasets demonstrate that GBLDM significantly outperforms four baselines in both classification accuracy and training efficiency, exhibiting exceptional robustness particularly in noisy scenarios.**

***Index Terms*—Granular Computing, Granular Ball Computing, Large Margin Distribution Machine, Pattern Classification**

I. Introduction

Classification remains a pivotal cornerstone of pattern recognition and machine learning, serving as a fundamental component across a myriad of real-world applications [1]. Over the past few decades, the Support Vector Machine (SVM) [2], grounded in statistical learning theory, has achieved remarkable success in handling small-sample, nonlinear, and high-dimensional tasks, primarily due to its adherence to the Structural Risk Minimization (SRM) principle [3]. However, the rapid evolution of information technology has ushered in an era of explosive data growth, often characterized by high-dimensional noise and complex distributional structures. Confronted with such massive and complex data streams, traditional algorithms relying on fine-grained "point-based" inputs face severe challenges [4]. Specifically, the sheer volume of samples induces a dramatic surge in computational complexity, creating prohibitive bottlenecks in training efficiency. Furthermore, pervasive noise and outliers inevitably destabilize decision boundaries, thereby compromising model robustness. Consequently, designing a classification paradigm that reconciles high computational efficiency with strong noise robustness has emerged as a critical imperative in current academic research.

The classical Support Vector Machine (SVM) constructs an optimal separating hyperplane by maximizing the minimum margin between classes [5]. While theoretically sound, its decision boundary relies exclusively on a sparse set of support vectors, thereby neglecting the global distributional information embedded in the vast majority of non-support vectors. This characteristic inevitably constrains its generalization capability under complex distributions. To mitigate this deficiency, Zhang et al. [6] proposed the Large Margin Distribution Machine (LDM). Distinct from SVM, LDM explicitly incorporates the margin mean and margin variance into the optimization objective, significantly enhancing performance by leveraging the second-order statistical information of the data. Motivated by this merit, a variety of LDM variants have subsequently been developed to further bolster model performance [7], [8]. However, LDM remains constrained by inherent limitations [9]. Firstly, the optimization of margin variance necessitates processing the dense covariance matrix of all samples, incurring prohibitive computational costs on large-scale datasets. Secondly, strictly adhering to a fine-grained "point-based" paradigm, LDM remains highly sensitive to local noise, thereby lacking the ability to capture macroscopic data structures from a multi-granularity perspective.

To address the efficiency and robustness constraints inherent in point-based models, Zadeh [10] introduced Granular Computing (GrC) as a paradigm for handling complex, massive problems. Building upon this foundation and incorporating the "global precedence" cognitive mechanism proposed by Chen [11]. Xia et al. [12], [13] formulated the theory of Granular Ball Computing (GBC) and the Granular Ball Support Vector Machine (GBSVM) framework. By compressing massive datasets into a concise set of granular balls, GBSVM significantly accelerates the training process. Additionally, it effectively mitigates local noise through an intra-granule majority voting mechanism. Recognizing that the geometric quality of granular balls determines the model's performance ceiling, recent studies have increasingly focused on optimizing their generation mechanisms. For instance, Xie et al. [14]

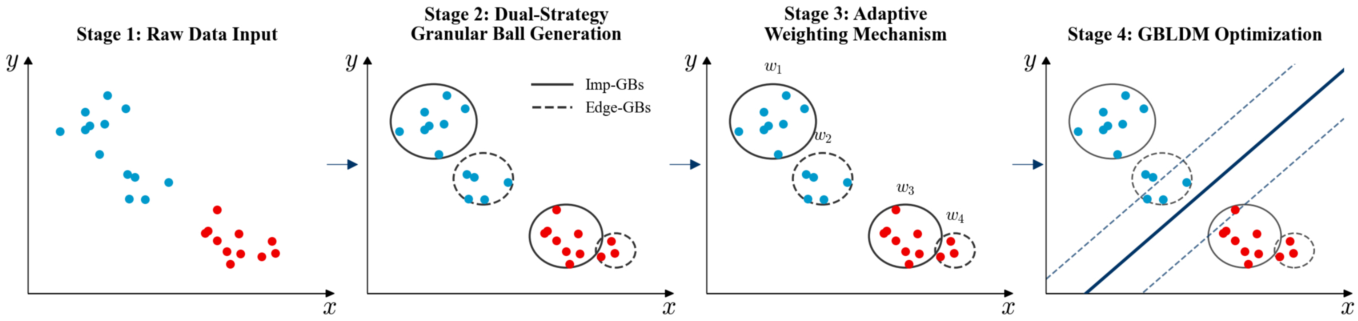

Fig. 1. Architectural schematic of GBLDM.

introduced attention mechanisms to enhance the stability of granular ball center selection; Xie et al. [15] proposed a weighted distribution metric to circumvent the dependency on fixed purity thresholds. Concurrently, diverse generation strategies have been explored from multiple perspectives to further optimize data representation [16], [17].

Motivated by the aforementioned analysis, this paper proposes a novel framework termed the Granular Ball Large Margin Distribution Machine (GBLDM). Anchored in the principles of "divide-and-conquer" and "distributional optimization," GBLDM orchestrates a dual-strategy generation mechanism: the first stage extracts the topological backbone using high-purity "Important Granular Balls," while the second stage meticulously refines complex boundaries via "Edge Granular Balls." Departing from traditional point-based approaches, GBLDM elevates the optimization landscape from "points" to "granular balls," leveraging geometric centers and radii to explicitly optimize the margin mean and variance, thereby achieving an optimal fit to the global data structure. The architectural schematic is shown in Fig. 1.

The main contributions of this paper are summarized as follows:

- A novel dual-strategy granular ball generation method is designed. To address the boundary ambiguity and reliance on fixed purity thresholds in traditional methods, we introduce the dual criteria of "label consistency" and "geometric compactness."
- A general GBLDM classification framework is proposed. By integrating Granular Ball Computing into Large Margin Distribution Machine theory, we shift the statistical optimization from the instance space to the granular ball space.
- Comprehensive experiments demonstrate the model's superiority. Extensive evaluations on UCI benchmark datasets validate that GBLDM achieves statistically significant improvements in accuracy, noise robustness, and training efficiency compared to four baselines.

II. PRELIMINARIES

This section reviews the theoretical foundations of the Large Margin Distribution Machine and Granular Ball Computing.

A. Large Margin Distribution Machine

The Large Margin Distribution Machine was introduced by Zhang et al. [6] as a solution to SVM's inability to capture the global distribution of data. Distinguishing itself from SVM, which prioritizes only boundary support vectors, LDM builds upon the Structural Risk Minimization theory and emphasizes the critical role of margin distribution in generalization performance. Consequently, the model aims to maximize the margin mean and minimize the margin variance simultaneously, while also retaining the maximization of the minimum margin.

To formally introduce the philosophy of LDM, let the functional margin of a training sample (x_i, y_i) with respect to the hyperplane be denoted as $\gamma_i = y_i(\omega^T x_i + b)$. On this basis, LDM defines the margin mean $\bar{\gamma}$ and margin variance $\hat{\gamma}$ over the entire dataset as follows:

$$\bar{\gamma} = \frac{1}{N}\sum_{i=1}^{N}\gamma_i, \quad \hat{\gamma} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\gamma_i - \gamma_j)^2, \tag{1}$$

where N represents the total number of samples. Here, $\bar{\gamma}$ quantifies the average deviation of the sample set from the hyperplane, while $\hat{\gamma}$ measures the dispersion of the margin distribution.

Integrating the principle of SRM, the ultimate optimization objective of LDM under the assumption of ideal linear separability is formulated as:

$$\begin{aligned} &\min_{\omega,b} \frac{1}{2}||\omega||^2 + \lambda_1\hat{\gamma} - \lambda_2\bar{\gamma}, \\ &\text{s.t. } y_i(\omega^T x_i + b) \geq 1, \;\; i = 1, \ldots, N. \end{aligned} \tag{2}$$

However, in practical scenarios where data are often non-linearly separable or contaminated with noise, LDM incorporates slack variables ξ_i, analogous to soft-margin SVM, to enhance robustness. Consequently, the objective is extended to minimize the empirical risk while optimizing margin distribution statistics. This constitutes the standard form of LDM, serving as the foundation for the improved model proposed in this paper:

$$\begin{aligned} &\min_{\omega,b,\xi} \frac{1}{2}||\omega||^2 + \lambda_1\hat{\gamma} - \lambda_2\bar{\gamma} + C\sum_{i=1}^{N}\xi_i, \\ &\text{s.t. } y_i(\omega^T x_i + b) \geq 1 - \xi_i, \;\; \xi_i \geq 0, \;\; i = 1, \ldots, N. \end{aligned} \tag{3}$$

It is noteworthy that while soft-margin LDM theoretically surpasses SVM, its training procedure remains grounded in individual sample points. As the data scale expands, calculating the mean and variance of the margin necessitates traversing all sample pairs, leading to prohibitive computational complexity.

B. Granular Ball Computing

The paradigm of Granular Ball Computing was introduced by Xia et al. [12] as a solution to the inefficiency and noise sensitivity inherent in traditional point-based models during large-scale data processing. Its core philosophy simulates the "global precedence" mechanism of human cognition by employing coarse-grained granular balls to cover and characterize local sample clusters within the original data space, which facilitates efficient data abstraction.

1) Geometric Definition: In the GBC framework, a granular ball is defined as a hypersphere encapsulating a subset of sample points. For a data subset $D \in U$ containing m sample points, the corresponding granular ball $gb = (c, r)$ is uniquely determined by its center c and radius r:

$$c = \frac{1}{m}\sum_{i=1}^{m} x_i, \quad r = \frac{1}{m}\sum_{i=1}^{m} ||x_i - c||. \tag{4}$$

In contrast to the strategy adopted by Peng et al. [18], which utilizes the maximum distance $r = \max_{x_i \in D} \|x_i - c\|$ to enhance granular ball coverage, this paper employs the average distance for the radius to enhance the representation's robustness against noise.

2) Generation Strategy: The generation of granular balls typically follows a top-down, purity-driven iterative splitting strategy. To guide this process, the label and purity of a granular ball are defined as follows:

$$\begin{aligned} label(gb) &= \operatorname*{argmax}_{y \in \{+1,-1\}} |\{(x_i, y_i) \in D \mid y = y_i\}|, \\ purity(gb) &= \frac{|\{(x_i, y_i) \in D \mid y_i = label(gb)\}|}{|D|}. \end{aligned} \tag{5}$$

The generation process begins by conceptualizing the entire dataset as a single or a few large granular balls. If the purity of a granular ball falls below a predefined threshold, it is split into finer-grained child balls. This fission process repeats recursively until the purity of all granular balls satisfies the requirement or a minimum size constraint is reached. Ultimately, the massive dataset U is compressed into a multi-granularity granular ball set $GB = \{gb_1, gb_2, \ldots, gb_k\}$, where $k \ll N$.

III. Granular Ball Large Margin Distribution Machine

This section proposes the Granular Ball Large Margin Distribution Machine. For convenience, the main notations used throughout this paper are summarized in Table I.

TABLE I
TABLE OF NOTATIONS

Notation	Description
U	The training dataset $\{(x_i, y_i)\}_{i=1}^{N}$
N	The total number of samples in U
d	The dimensionality of the feature space
x_i, y_i	The i-th sample vector and its class label
ω, b	The weight vector and bias of the hyperplane
GB	The set of granular balls generated from U
k	The total number of granular balls in GB
gb_i	The i-th granular ball, $gb_i \in GB$
c_i, r_i	The center and radius of gb_i
n_i	The number of samples contained in gb_i
GB_{imp}	The subset of Important Granular Balls
GB_{edge}	The subset of Edge Granular Balls
DM	The Distribution Metric for compactness
α	The Granularity Regulation Coefficient
s_i	The adaptive comprehensive weight of gb_i
ζ_i	The functional margin of granular ball gb_i
$\bar{\zeta}, \hat{\zeta}$	The margin mean and variance of granular balls
ρ	The radius penalty coefficient
λ_1, λ_2	Regularization parameters for margin distribution

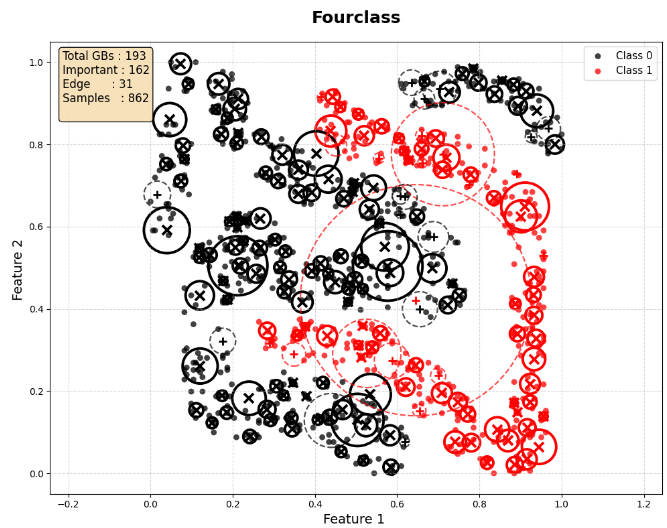

Fig. 2. Visualization of Granular Balls on the Fourclass dataset. Important Granular Balls are marked with solid lines, and Edge Granular Balls with dashed lines.

A. Granular Ball Generation Mechanism

To precisely characterize the global topological structure and local decision boundaries of the data, we propose a dual-strategy generation mechanism. The visualization of granular balls generated on the Fourclass dataset is presented in Fig. 2.

Definition 1 (Important Granular Ball and Edge Granular Ball). Given a dataset $U = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in R^d$ is the input vector and $y_i \in \{+1, -1\}$ is the class label. Let $GB = \{gb_1, gb_2, \ldots, gb_k\}$ be the family of granular balls generated on U. Based on the distribution characteristics and functions of the granular balls in the data space, we define the granular ball family GB as the union of two disjoint subsets:

$$GB = GB_{imp} \cup GB_{edge}. \tag{6}$$

Definition 2 (Distribution Metric). Given a granular ball gb_i containing n_i samples on a subset $D \in U$, with center c_i, the Distribution Metric (DM) is defined by Xie et al. [15] as the average distance from internal samples to the center:

$$DM_i = \frac{1}{n_i}\sum_{j=1}^{n_i} ||x_j - c_i||. \tag{7}$$

Geometrically, a smaller DM indicates a more compact data distribution and higher representation quality.

Xie calculates the Weighted Distribution Metric DM_w to evaluate the geometric optimization during iterative splitting. This metric represents the weighted average compactness of the child granular balls. Assuming a parent ball gb_i is partitioned into two child balls gb_{i1} and gb_{i2} with sample sizes n_{i1} and n_{i2} respectively, DM_w is calculated as:

$$DM_w = \frac{n_{i1}}{n_i} DM(gb_{i1}) + \frac{n_{i2}}{n_i} DM(gb_{i2}). \tag{8}$$

Building upon this comparison, we propose a robust criterion to control the splitting process.

Definition 3 (Granularity Regulation Coefficient). The Granularity Regulation Coefficient, denoted as α, is defined as a scaling factor that modulates the stringency of the geometric splitting criterion. A splitting operation is deemed valid if and only if it satisfies the following inequality:

$$DM_w < \alpha \cdot DM_i, \tag{9}$$

where α quantifies the acceptable tolerance for the trade-off between the compactness gain and the number of generated granular balls.

B. Dual-Strategy Granular Ball Generation Mechanism

We propose a coarse-to-fine Dual-Strategy Granular Ball Generation Mechanism to accurately capture both the global topological backbone and local decision boundaries.

1) Stage I: Generation of Important Granular Balls

The primary objective of this stage is to extract high-level label features using a minimal number of granular balls, thereby reducing computational complexity. We employ a top-down iterative splitting strategy governed by dual criteria:

- **Label Consistency:** A granular ball is deemed "impure" and must undergo splitting if its purity falls below a predefined threshold p_{th}.
- **Geometric Compactness:** To prevent loose structures, the splitting validity is determined by the Granularity Regulation Coefficient α. Only splits satisfying $DM_w < \alpha \cdot DM_{parent}$ are executed.

Post-iteration, a rigorous filtering step retains only those granular balls satisfying both criteria as Important Granular Balls. Granular balls failing these requirements are dissolved, and their internal samples, which typically reside in class overlap regions or act as outliers, are released back to the unassigned sample pool for secondary processing.

2) Stage II: Generation of Edge Granular Balls

The second stage focuses on the residual sample set $U_{res} = U \setminus S_{covered}$, where $S_{covered}$ denotes samples in GB_{imp}, aiming to guarantee complete coverage of the feature space. As these samples often represent complex boundary noise or outliers, we adopt a "Fine-Grained Isolation" strategy:

- **Size Relaxation:** The minimum size constraint is relaxed to 1, allowing even isolated outliers to form independent granular balls.
- **Purity Enforcement:** Strict purity requirements are maintained to meticulously separate intermingled samples from different classes.

This strategy generates refined Edge Granular Balls to patch boundary details. By employing this generation module, the subsequent optimization process can be effectively shifted from the point space to the granular ball space. Consequently, the prohibitive time complexity of the traditional LDM is significantly reduced from $\mathcal{O}(N^3+N^2d)$ to $\mathcal{O}(N \cdot d \cdot t+k^3+k^2d)$, where $k \ll N$. The comprehensive generation process is summarized in Algorithm 1.

Algorithm 1 Dual-Strategy Granular Ball Generation Mechanism

Input: Dataset U, Purity threshold p_{th}, Minimum size m_{size}, Granularity regulation coefficient α.
Output: Granular ball set $GB = GB_{imp} \cup GB_{edge}$.

Initialize $GB \leftarrow \{$Initial ball containing all $U\}$
repeat
  $GB_{new} \leftarrow \emptyset$
  for each $gb \in GB$ **do**
    Calculate $purity(gb)$ and $DM(gb)$
    if $purity(gb) < p_{th}$ **or** $|gb| > 2 \times m_{size}$ **then**
      Perform geometric splitting on gb to obtain children gb_{c1}, gb_{c2}
      Calculate Weighted Distribution Metric DM_w
      if $DM_w < \alpha \times DM(gb)$ **then**
        $GB_{new} \leftarrow GB_{new} \cup \{gb_{c1}, gb_{c2}\}$
      else
        $GB_{new} \leftarrow GB_{new} \cup \{gb\}$
      end if
    else
      $GB_{new} \leftarrow GB_{new} \cup \{gb\}$
    end if
  end for
  $GB \leftarrow GB_{new}$
until $GB_{new} = GB$
$GB_{imp} \leftarrow \emptyset$
for each $gb \in GB$ **do**
  if $purity(gb) \geq p_{th}$ **then**
    Mark gb as **Important**
    $GB_{imp} \leftarrow GB_{imp} \cup \{gb\}$
  end if
end for
$S_{covered} \leftarrow \bigcup_{gb \in GB_{imp}}$ Samples(gb)
$S_{res} \leftarrow U \setminus S_{covered}$
if $S_{res} \neq \emptyset$ **then**
  Generate GB_{edge} on S_{res} with criteria
  Mark balls in GB_{edge} as **Edge**
end if
return $GB \leftarrow GB_{imp} \cup GB_{edge}$

C. Construction of GBLDM Classification Model

Based on the granular ball set $GB = GB_{imp} \cup GB_{edge}$ obtained from the dual-strategy generation mechanism, GBLDM aims to learn an optimal separating hyperplane $f(x) = \omega^T x + b = 0$.

Definition 4 (Granular Ball Margin). Given a granular ball gb_i characterized by center c_i, radius r_i, and label y_i, its functional margin ζ_i with respect to the hyperplane is defined as:

$$\zeta_i = y_i(\omega^T c_i + b) - \rho \cdot r_i, \tag{10}$$

where $\rho \geq 0$ is the radius penalty coefficient. This penalty explicitly captures the geometric characteristics of the granular ball. By extending the margin constraint from an isolated center point to the entire granular ball space, it effectively bounds the spatial uncertainty of the encapsulated samples.

To further elevate generalization performance, GBLDM goes beyond merely satisfying the minimum margin constraint; it explicitly optimizes the global distribution of the granular ball margins.

Definition 5 (Margin Mean and Variance). Given the granular ball set $GB = \{gb_1, \ldots, gb_k\}$, the margin mean $\bar{\zeta}$ and margin variance $\hat{\zeta}$ are defined as:

$$\bar{\zeta} = \frac{1}{k}\sum_{i=1}^{k}\zeta_i, \quad \hat{\zeta} = \frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k}(\zeta_i - \zeta_j)^2. \tag{11}$$

Unlike point-based models, granular balls exhibit distinct variations in sample capacity and label roles. Consequently, a uniform weighting scheme is insufficient. To accurately reflect these differences, we introduce an adaptive weighting mechanism.

Definition 6 (Adaptive Granular Weight). For a granular ball gb_i containing n_i samples, its adaptive comprehensive weight s_i is formulated as:

$$s_i = n_i \times \eta_{type} \times \eta_{balance}, \tag{12}$$

where n_i represents the number of samples contained within the granular ball gb_i. η_{type} acts as the type factor, assigning lower weights to Edge Granular Balls to dampen their potential interference with gradient optimization. $\eta_{balance}$ is the class balance factor, defined as:

$$\eta_{balance} = \frac{\text{number of minority samples}}{\text{number of majority samples}} = \frac{N_{min}}{N_{maj}}. \tag{13}$$

Synthesizing these components and adhering to the Structural Risk Minimization principle, the ultimate optimization objective of GBLDM is formulated to minimize the weighted empirical risk while optimizing the margin distribution.

Definition 7 (GBLDM Objective Function). Given the granular ball set $GB = \{gb_1, \ldots, gb_k\}$, the primal optimization problem of GBLDM is formulated as:

$$\begin{aligned} &\min_{\omega,b,\xi} \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{k} s_i\xi_i + \lambda_1\hat{\zeta} - \lambda_2\bar{\zeta} \\ &\text{s.t. } y_i(\omega^T c_i + b) - \rho r_i \geq 1 - \xi_i, \\ &\qquad \xi_i \geq 0, \ \ i = 1, \ldots, k. \end{aligned} \tag{14}$$

Since the objective function in (14) is strictly convex, upon solving it to obtain the optimal weight vector ω^* and bias b^*, the classification rule can be established.

Definition 8 (Decision Function). For an arbitrary test sample x, the decision function is defined as:

$$f(x) = \text{sgn}(\omega^* x + b^*), \tag{15}$$

where $\text{sgn}(\cdot)$ denotes the sign function.

Remarkably, the proposed GBLDM framework serves as a theoretical generalization of the existing GBSVM. Specifically, by setting the margin distribution hyperparameters λ_1 and λ_2 to zero and fixing the adaptive weights s_i to 1, the objective function in Definition 7 mathematically reduces to the standard soft-margin GBSVM. The complete algorithmic logic of GBLDM is summarized in Algorithm 2.

Algorithm 2 The General Framework of GBLDM

Input: Training set U_{train}, Test set U_{test},
Hyperparameters: $p_{th}, m_{size}, \alpha, C, \lambda_1, \lambda_2, \rho$.
Output: Predicted labels Y_{pred} for U_{test}.

```
Initialize GB ← ∅
Construct Important Granular Balls GB_imp and Edge Granular Balls GB_edge on U_train using Algorithm 1
Set GB ← GB_imp ∪ GB_edge
for each gb_i ∈ GB do
    Calculate sample count n_i and identify granular type
    Compute adaptive comprehensive weight s_i via Eq. 12
end for
Initialize weight vector w ← 0 and bias b ← 0
repeat
    for each gb_i ∈ GB do
        Calculate functional margin considering radius penalty ρ
        Compute gradients for Margin Mean, Margin Variance, and Empirical Risk
        Update w and b to minimize the objective function
    end for
until convergence
Obtain optimal parameters w* and b*
Initialize Y_pred ← ∅
for each test sample x_k ∈ U_test do
    Predict label y_k ← f(x_k) according to Definition 8
    Y_pred ← Y_pred ∪ {y_k}
end for
return Y_pred
```

IV. Numerical Experiments

A. Experimental Environment

Comprehensive performance validation of the proposed dual-strategy granular ball generation mechanism and the GBLDM classifier is conducted by evaluating GBLDM against four baselines: SVM [19], LDM [6], GBSVM [20] and ORI-GBLDM, on eight UCI benchmark datasets. Specifically, ORI-GBLDM serves as an ablation model where the proposed granular ball generation module is replaced by a K-means-based method. The statistical characteristics of these datasets are summarized in Table II. Datasets were randomly partitioned into 80% training and 20% testing sets. To assess robustness, label noise was introduced at ratios of 5%, 10%, and 15%. We report the average accuracy and training time over four independent trials under optimal parameters. All experiments

TABLE II
Statistical Characteristics of the Benchmark Datasets

Datasets	Samples	Features	Classes
Diabetes	768	7	2
Fourclass	862	2	2
Gamma	19020	10	2
German.numer	1000	23	2
Heart	270	12	2
HTRU	17898	8	2
Phoneme	5404	5	2
WDBC	569	30	2

TABLE III

OPTIMAL PARAMETERS OF FIVE MODELS ON UCI DATASETS WITHOUT NOISE

Model	Diabetes	Fourclass	Gamma	German.numer	Heart	HTRU	Phoneme	WDBC
SVM (C)	2^1	2^{-5}	2^{-1}	2^{-1}	2^{-5}	2^4	2^0	2^2
LDM (C, λ_1, λ_2)	2^{-2}, 2^{-5}, 2^1	2^{-4}, 2^0, 2^4	2^{-2}, 2^{-5}, 2^{-5}	2^{-3}, 2^{-4}, 2^4	2^{-2}, 2^0, 2^3	2^{-2}, 2^{-5}, 2^{-5}	2^4, 2^{-3}, 2^0	2^2, 2^2, 2^4
GBSVM (C)	2^0	2^4	2^2	2^{-1}	2^{-4}	2^1	2^4	2^3
ORI-GBLDM (C, λ_1, λ_2)	2^{-2}, 2^4, 2^1	2^0, 2^2, 2^1	2^3, 2^4, 2^0	2^3, 2^4, 2^2	2^3, 2^{-5}, 2^{-4}	2^{-4}, 2^4, 2^{-1}	2^2, 2^2, 2^{-1}	2^{-4}, 2^3, 2^3
GBLDM (C, λ_1, λ_2)	2^0, 2^4, 2^3	2^3, 2^{-4}, 2^{-1}	2^{-2}, 2^4, 2^3	2^4, 2^3, 2^4	2^4, 2^1, 2^3	2^3, 2^4, 2^{-3}	2^4, 2^3, 2^{-3}	2^4, 2^3, 2^{-1}

TABLE IV

COMPARISON OF ACCURACY BETWEEN DIFFERENT METHODS WITH DIFFERENT LEVELS OF CLASS NOISE

Dataset		Diabetes	Fourclass	Gamma	German.numer	Heart	HTRU	Phoneme	WDBC
0%	SVM	0.7662	0.7919	0.7911	0.7938	0.8287	**0.9797**	0.7803	0.9803
	LDM	0.7706	0.8069	0.7923	0.7725	0.8642	0.9696	0.7730	0.9854
	GBSVM	0.8084	0.7905	0.7899	0.7300	0.8750	0.9788	0.7849	0.9868
	ORI-GBLDM	0.7711	0.7977	0.7780	0.7525	0.8611	0.9608	0.7428	0.9693
	GBLDM	**0.8166**	**0.8136**	**0.8029**	**0.8075**	**0.9167**	0.9727	**0.8027**	**0.9912**
5%	SVM	0.7662	0.7775	0.7890	0.7750	0.8380	0.9719	0.7738	0.9627
	LDM	0.7695	0.8050	0.7905	0.7675	0.8457	0.9632	0.7710	0.9737
	GBSVM	0.7955	0.7384	0.7907	0.7225	0.8287	**0.9752**	**0.7828**	0.9737
	ORI-GBLDM	0.7695	0.7962	0.7535	0.7312	0.8426	0.9580	0.7259	0.9627
	GBLDM	**0.8019**	**0.8092**	**0.7992**	**0.7888**	**0.9167**	0.9707	0.7692	**0.9825**
10%	SVM	0.7646	0.7702	0.7874	0.7600	0.8333	0.9664	0.7680	0.9605
	LDM	0.7532	0.7722	0.7879	0.7625	0.8549	0.9626	0.7537	0.9649
	GBSVM	**0.7873**	0.7803	0.7896	0.7000	0.8380	**0.9743**	**0.7842**	0.9583
	ORI-GBLDM	0.7614	0.7934	0.7673	0.7337	0.8565	0.9544	0.7216	0.9561
	GBLDM	**0.7873**	**0.8035**	**0.7979**	**0.7838**	**0.8935**	0.9724	0.7528	**0.9759**
15%	SVM	0.7581	0.7659	0.7835	0.7575	0.8287	0.9613	0.7662	0.9496
	LDM	0.7457	0.7577	0.7847	0.7367	0.8302	0.9573	0.7517	0.9547
	GBSVM	0.7776	0.7746	0.7890	0.6962	0.8472	**0.9744**	**0.7782**	**0.9627**
	ORI-GBLDM	0.7532	0.7890	0.7666	0.7100	0.8519	0.9562	0.7185	0.9605
	GBLDM	**0.7906**	**0.8064**	**0.7971**	**0.7863**	**0.8981**	0.9729	0.7523	**0.9627**

were conducted in a Python 3.11 environment on a computing platform equipped with an Intel Core i7-12700 CPU @ 2.10 GHz and 32 GB RAM.

B. Parameter Settings

To ensure a fair comparison, consistent search spaces were strictly maintained for parameters sharing equivalent physical semantics. Specifically, The regularization parameter C across all models, alongside the margin distribution hyperparameters λ_1 and λ_2 specifically for LDM, ORI-GBLDM and GBLDM, were optimized within the range $\{2^i \mid i \in \{-5, \ldots, 5\}\}$. Regarding the granular ball generation mechanism in ORI-GBLDM and GBLDM, the purity threshold was selected from $\{0.90, \ldots, 1.0\}$, the granularity regulation coefficient α within $[0.8, 1.2]$, and the minimum granular ball size from the discrete set $\{2, \ldots, 5\}$. A 10-fold cross-validation scheme was employed to identify the optimal parameter combination. The detailed configurations for each model are listed in Table III.

C. Results Analysis

Table IV summarizes the detailed statistical results under varying noise levels. In noise-free benchmarks, GBLDM demonstrates superior generalization, achieving the highest accuracy on $7/8$ datasets. Furthermore, GBLDM consistently surpasses the ablation model ORI-GBLDM across all datasets, explicitly validating the superiority of the proposed dual-strategy generation mechanism over standard K-means clustering. By incorporating global margin distribution optimization, it effectively transcends the geometric-margin limitation of GBSVM and the computational bottlenecks of traditional LDM. Crucially, in noisy scenarios, GBLDM exhibits robust stability. While point-based models suffer sharp performance degradation with increasing noise ratios, GBLDM consistently outperforms baselines. This immunity arises from the "synergistic effect" of the granular structure and margin optimization: discrete noise is effectively encapsulated within edge granular balls and down-weighted, while fine-grained distribution constraints prevent overfitting to noise-affected regions, thereby ensuring a robust decision boundary.

Table V illustrates the computational efficiency analysis among LDM, ORI-GBLDM and GBLDM. On small-scale datasets, both granular ball-based models incur a marginal fixed overhead due to the granular ball generation process. However, as the dataset scale and complexity increase, the granular ball-based models demonstrate overwhelming computational superiority over LDM. Notably, while the ablation model ORI-GBLDM exhibits the lowest running time on large datasets owing to its simpler K-means clustering logic,

TABLE V
COMPARISION OF THE RUNNING TIME AND SPEEDUP BETWEEN LDM AND GBLDM WITHOUUT NOISE

Dataset	Diabetes	Fourclass	Gamma	German.number	Heart	HTRU	Phoneme	WDBC
LDM	**0.1516**	**0.1211**	494.4192	**0.1899**	**0.0200**	661.7615	33.9084	**0.0643**
ORI-GBLDM	1.2901	0.5651	**22.3940**	3.8662	0.6416	**1.4468**	**4.8550**	0.5609
GBLDM	2.7734	1.8182	50.5759	5.5663	1.4039	9.5144	13.1979	1.1184

GBLDM remains highly efficient and achieves orders-of-magnitude speedups compared to the traditional LDM. This slightly higher time cost of GBLDM over ORI-GBLDM is a highly acceptable trade-off for the superior classification accuracy and noise robustness introduced by the proposed dual-strategy initialization. This scalability stems from the granular ball-computing paradigm, which breaks the performance bottleneck of point-based models by aggregating massive raw samples into a significantly reduced set of granular balls.

D. Statistical Test

To further validate the performance superiority of the proposed algorithm, non-parametric statistical tests were employed for significance analysis. The Friedman test was conducted across all noise levels to evaluate the overall differences among the five algorithms. The results yield a p-value far below 0.05, indicating statistically significant differences in performance. Specifically, the overall average rankings of all evaluated models are: GBLDM (1.47), GBSVM (2.73), LDM (3.31), SVM (3.45), and ORI-GBLDM (4.03). This confirms that GBLDM not only achieves the highest average accuracy but also possesses statistically significant robustness and generalization capability.

V. CONCLUSION

In this paper, we proposed the Granular Ball Large Margin Distribution Machine (GBLDM), a novel framework that integrated granular ball computing with large margin distribution theory. By employing a dual-strategy granular ball generation mechanism and an adaptive weighted optimization model, GBLDM effectively captured global topological structures and explicitly optimized margin statistics within the granular ball space. This mechanism strategically extracts the data's topological backbone using Important Granular Balls while meticulously refining complex decision boundaries via Edge Granular Balls. Extensive experiments across eight UCI benchmark datasets demonstrated that this approach simultaneously resolved the computational bottlenecks of traditional LDM and the geometric-margin limitations of GBSVM. Consequently, GBLDM achieves statistically significant improvements in superior accuracy and robustness, particularly in scenarios with significant noise. In future work, we plan to introduce kernel methods to enhance the model's non-linear mapping capability for high-dimensional data and explore advanced acceleration strategies to further improve computational efficiency.

REFERENCES

[1] Babita and D. R. Nayak, "Histoformer: A frequency domain-based convolutional transformer for improved breast cancer classification," *IEEE Trans. Artif. Intell.*, pp. 1-12, 2026.

[2] M. Bernardini, L. Romeo, P. Misericordia, and E. Frontoni, "Discovering the type 2 diabetes in electronic health records using the sparse balanced support vector machine," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 1, pp. 235-246, Jan. 2020.

[3] J. Yang *et al.*, "Three-way outlier detection based on shadowed granular-balls," *IEEE Trans. Fuzzy Syst.*, vol. 34, no. 1, pp. 101-113, Jan. 2026.

[4] W. J. Jia, X. D. Liu, Y. G. Wang, W. Pedrycz, and J. X. Zhou, "Semisupervised learning via axiomatic fuzzy set theory and SVM," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 4661-4674, Jun. 2022.

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273-297, Sep. 1995.

[6] T. Zhang and Z. H. Zhou, "Large margin distribution machine," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, New York, NY, USA, 2014, pp. 313-322.

[7] F. Y. Cheng, J. Zhang, and C. H. Wen, "Cost-sensitive large margin distribution machine for classification of imbalanced data," *Pattern Recognit. Lett.*, vol. 80, pp. 107-112, 2016.

[8] S. Abe, "Unconstrained large margin distribution machines," *Pattern Recognit. Lett.*, vol. 98, pp. 96-102, 2017.

[9] L. B. Zhang, D. H. Dong, L. Y. Luo, and D. Liu, "A novel fuzzy large margin distribution machine with unified pinball loss," *IEEE Trans. Fuzzy Syst.*, vol. 32, no. 4, pp. 1782-1795, 2024.

[10] L. A. Zadeh, "Fuzzy sets and information granularity," in *Advances in Fuzzy Set Theory and Applications*, N. Gupta, R. Ragade, and R. Yager, Eds. Amsterdam: North-Holland, 1979, pp. 3-18.

[11] L. Chen, "Topological structure in visual perception," *Science*, vol. 218, pp. 699-700, Nov. 1982.

[12] S. Y. Xia, Y. S. Liu, X. Ding, G. Y. Wang, H. Yu, and Y. G. Luo, "Granular ball computing classifiers for efficient, scalable and robust learning," *Inf. Sci.*, vol. 483, pp. 136-152, May 2019.

[13] S. Y. Xia, X. C. Dai, G. Y. Wang, X. B. Gao, and E. Giem, "An efficient and adaptive granular-ball generation method in classification problem," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 4, pp. 5319-5331, Apr. 2024.

[14] Q. Xie, Q. H. Zhang, S. Y. Xia, F. Zhao, C. Y. Wu, G. Y. Wang, and W. P. Ding, "GBG++: A fast and stable granular ball generation method for classification," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 2, pp. 2022-2036, Apr. 2024.

[15] J. Xie, W. Y. Kong, S. Y. Xia, G. Y. Wang, and X. B. Gao, "An efficient spectral clustering algorithm based on granular-ball," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 9, pp. 9743-9753, Sep. 2023.

[16] F. Liu, Q. H. Zhang, S. Y. Xia, Q. Xie, W. Liao, and S. Y. Zhang, "A granular-ball generation method based on local density for classification," *Inf. Sci.*, vol. 717, Art. no. 122295, Nov. 2025.

[17] J. L. Pan, G. M. Lang, Q. M. Xiao, and T. Yang, "A framework of granular-ball generation for classification via granularity tuning," *Appl. Intell.*, vol. 55, no. 1, Art. no. 63, Dec. 2025.

[18] X. L. Peng, P. Wang, S. Y. Xia, C. Wang, and W. Q. Chen, "VPGB: A granular-ball based model for attribute reduction and classification with label noise," *Inf. Sci.*, vol. 611, pp. 504-521, Sep. 2022.

[19] J. C. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Microsoft Research, Redmond, WA, USA, Tech. Rep. MSR-TR-98-14, 1998.

[20] S. Y. Xia, X. Y. Lian, G. Y. Wang, X. B. Gao, J. C. Chen, and X. L. Peng, "GBSVM: An efficient and robust support vector machine framework via granular-ball computing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 5, pp. 9253-9267, May 2025.