Available online at www.HighTechJournal.org

HighTech and Innovation Journal

ISSN: 2723-9535

Vol. 5, No. 3, September, 2024

A Novel Classification Model Based on Hybrid K-Means and Neural Network for Classification Problems

Cui Chenghu [1], Arit Thammano [1*]

[1] *Computational Intelligence Laboratory, School of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand.*

Received 16 May 2024; Revised 11 August 2024; Accepted 17 August 2024; Published 01 September 2024

Abstract

We propose a new classification model—a new classification model for clustering overlapping problems based on K-Means and neural networks. K-means clustering algorithm belongs to unsupervised learning. It is a classic algorithm for solving clustering problems. Since this algorithm calculates its categories based on distance, the results tend to converge to the local optimal solution and have poor boundary clustering properties. The K-Means classification algorithm defines clusters by the distance between the cluster center value and the target object, and the optimal result is obtained through continuous iteration. Therefore, clustering results are overlapped, and there are often outliers that do not belong to the current cluster, resulting in unsatisfactory clustering results. Our model offers a new method to segment non-ideal data in overlapping regions. Since clustering algorithms cannot effectively identify and classify this part of the data, we split this part of the data and train it using a neural network. The results are then integrated into the clustered data. In the experiment, the k-fold cross-validation method ensures the model stability of the results. We used the accuracy to evaluate the quality of the model, and we used standard deviation and mean deviation to detect clustering results. Five sets of experimental data from the cross-experiment show that compared with the K-Means classification model, the accuracy of our model is effectively improved.

Keywords: Overlapping Clustering; K-Means Classification; Neural Network; Machine Learning.

1. Introduction

The k-means classification algorithm and its solutions remain important for practical applications such as data mining, image processing, and text analysis [1, 2]. The random seed is used as the initial centroid, and the optimal solution of the model is determined through continuous iteration, this process is usually called model training [3, 4]. The parameter values of the optimal solution are used to achieve classification results for a specific dataset. The model usually uses a loss function to measure the performance of the model [5]. Due to algorithm advantages, the K-Means algorithm is widely used in various fields [6-8].

As researchers continue to improve the performance of the k-means classification model, the traditional solutions have become insufficient [9]. For decades, one of the most popular topics in K-means has been cluster overlapping problems [10]. The K-means classification model is a classic unsupervised learning algorithm. This algorithm uses the

* Corresponding author: arit@it.kmitl.ac.th

http://dx.doi.org/10.28991/HIJ-2024-05-03-012

inherent similarity of the data as an important indicator for classification. It groups unlabeled datasets into different clusters. However, since clustering is based on the center value to measure similarity, the central area would have higher accuracy. In contrast, the accuracy of the remote area of the cluster has a lower accuracy [11]. The overlapping issue of K-means clustering is caused by its characteristics [12, 13].

This is determined by its characteristics and cannot be changed. At present, the main methods to improve classification performance are hierarchical clustering and hybrid models. Hierarchical clustering establishes a hierarchical nested clustering tree based on the similarity of data points. The top layer of the tree is the cluster root node, and then it is continuously divided into two clusters until the optimal solution is reached. This is a Divisive analysis (DIANA), which is carried out from top to bottom. Another hierarchical clustering (AGNES) is carried out from bottom to top. This method is exactly the opposite of DIANA.

Although DIANA and AGNES can avoid clustering to the local optimal solution, this method also has great disadvantages, such as complex algorithm complexity and difficulty in controlling singular values.

In recent years, hybrid models have received widespread attention due to their advantages in processing complex data structures, data adaptability, and flexibility. "Hybrid algorithm" refers to a solution that combines algorithms and methods to solve complex problems. It can integrate the advantages of multiple algorithms and achieve optimal results in a shorter time frame. In another research paper, the authors proposed an improved K-means algorithm, this algorithm has the advantages of K-means clustering and avoids the problem of local optimality [14, 15].

Although some effective solutions have been proposed to solve the problem, we believe that our model could offer a new method for solving this problem. Therefore, this study proposes a novel classification model. The important contributions of our solution are as follows. First, our model retains the advantages of the K-means Model. Second, we propose a method to process only the overlapping area data without affecting other valid data.

2. Literature Review

Clustering algorithms, particularly k-means, have been extensively studied and applied across various domains. This review highlights several significant contributions in the field of clustering techniques and related applications.

Gan & Ng (2017) and Wang et al. (2019) introduce innovative approaches to enhance traditional k-means clustering by addressing the challenge of unclear cluster boundaries. Both studies propose three-way k-means algorithms that utilize overlap clustering to identify core and fringe regions. Perturbation analysis is then employed to separate these regions, resulting in an improved clustering structure. Evaluation of UCI and USPS datasets demonstrates the effectiveness of these methods in enhancing clustering accuracy [16, 17].

Lu et al. (2024) pointed out that the solution to the overlapping areas of clusters is that the traditional clustering method assumes binary classification [18], and this algorithm needs to be based on specific conditions to achieve its task. In contrast, the method they use does not have this problem; that is, it does not need to be based on the type of relationship of inclusion or exclusion of objects. Therefore, the three-way clustering method is used as an effective alternative to solve overlapping clustering [19].

Dai et al. (2024) pointed out that when there are imbalanced and overlapping categories in the dataset, the performance of traditional classifiers will decline. The majority and minority classes have an obvious relationship in the binary classification methodology. The classifier's overall performance can be greatly improved by using an area of overlapping categories. However, in multi-class imbalance issues, the relationship between clusters becomes very complicated [20]. In their study, a new solution was proposed that integrates the advantages of genetic algorithms. The experimental results of 19 public datasets show that this integrated genetic algorithm method is effective [21].

Zhu et al. (2019) propose a novel approach to multi-viewpoint set registration by framing it as a clustering problem and employing k-means clustering. Their method demonstrates effectiveness and robustness approaches, as evaluated against benchmark datasets [22, 23]. Lücke & Forster (2019) [24] introduce a novel perspective linking k-means clustering with Gaussian mixture models (GMM) using truncated variational EM. This approach offers a promising avenue for future theoretical and empirical research in clustering algorithms [25].

Sieranoja et al. (2024) present a new Model (OMOS) to solve the problem of overlapping problems in clusters. They believe that intra-class imbalance and sensitivity to parameter settings are the main problems. So, they proposed a solution to use the mean shift algorithm to identify minority classes. This way, the distribution characteristics of minority class clusters can be captured. The experimental results come from 20 imbalanced data sets and four different classifiers. The results show that the OMOS algorithm is effective [15].

Vuttipittayamongkol et al. (2018) address the challenge of imbalanced data classification by introducing a novel under-sampling technique. This method is mainly used to eliminate erroneous instances in overlapping areas, leading to significant improvements in classification performance across various datasets [26]. Huang et al. (2021) propose a new k-means classification algorithm to improve clustering performance by extracting hierarchical representations. The deep structure captures complex hierarchical information, leading to significant performance gains over classical and state-of-the-art methods across benchmark datasets [27]. Saputra et al. (2020) investigate the effects of different distance algorithms on the performance of the k-means model. Their study provided in-depth insights into the impact of distance metrics on clustering results [28].

These studies collectively contribute to advancing clustering techniques and their applications across various domains, addressing challenges such as unclear cluster boundaries, optimal parameter selection, and efficient data representation. Figure 1 shows the flowchart of the research methodology through which the objectives of this study were achieved. The contribution of this research:

- The proposed novel classification model framework was designed to solve clusters overlapping of the K-Means clusters overlapping problem of linear decision boundary and improve classification performance.
- An implemented ensemble classification model was tested for its performance in terms of standard deviation and mean deviation. The standard deviation can detect the size of cluster boundaries, while the mean deviation can detect the compactness of data within clusters.

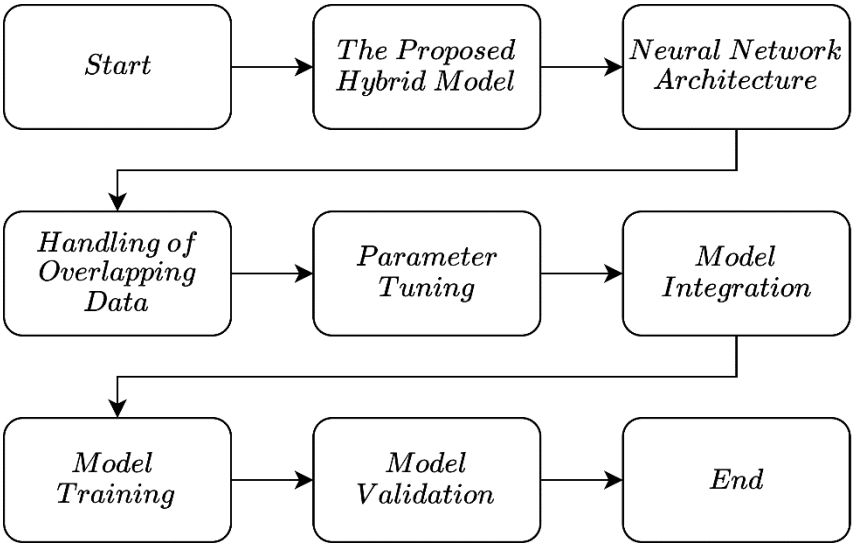

Figure 1. The Process of the methodology

3. Research Methodology

K-Means is widely recognized in the domain of machine learning and primarily utilized for classifying unlabeled data. Conversely, neural network algorithms operate within the realm of supervised learning, where models are trained using existing data labels for classification purposes. This paper introduces a novel model designed specifically for classifying unsupervised data within clustered overlapping regions. Notably, this new model combines features from both the neural network and K-Means algorithms. In the subsequent section, we will provide an overview of the fundamental K-Means algorithm and basic concepts of neural networks and elaborate on our innovative classification model.

3.1. K-Means Architecture

The K-Means classification model [8, 29] relies on distance metrics to assess similarity and categorize data points into clusters. Each data point is assigned to the cluster with the closest centroid, determined by distance calculation. The process involves iterative adjustments of centroids by the average values of the data within each cluster. The performance of the K-Means classification model is typically evaluated using the sum of squares. Below is the workflow implementation of the K-Means algorithm, which is depicted in Figure 2:

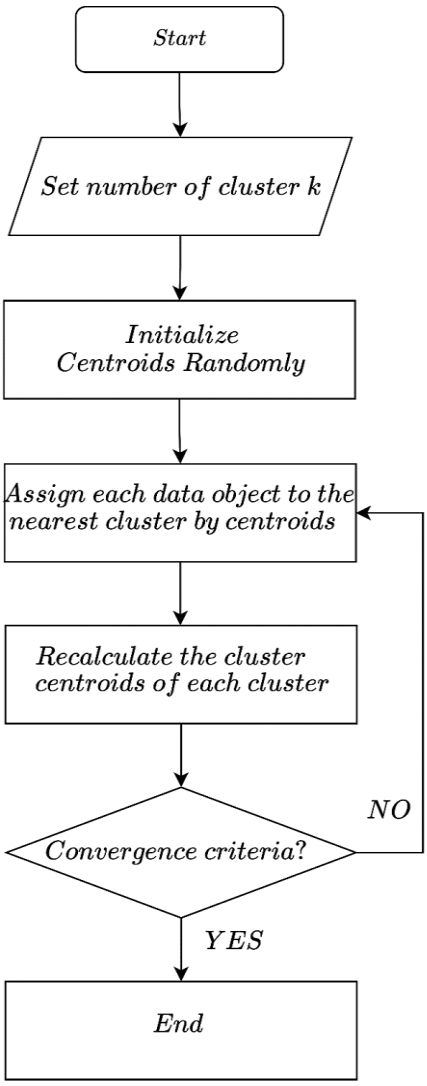

Figure 2. K-Means Classification Model Architecture

Perform K-Means Steps:

- Euclidean distance: It is calculated in high-dimensional space as defined by Equation 1. *D*: distance between 2 points.

$$D(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (1)$$

- The sum of squares: It is used for the line of best fit will minimize this value as defined by Equation 2. Where y_i is the observed value, $\hat{y}_i$ is estimated by the regression line.

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

3.2. Feedforward Neural Network

The sigmoid function is a logistic function characterized by an "S"-shaped curve [30]. In neural network models, sigmoid is commonly utilized as an activation function [31]. The forward propagation neural network involves the process of input data and output results; the network architecture consists of an input layer, a hidden layer, and an output layer. The derivative of the sigmoid function is essential for backpropagation in neural network models. Backpropagation utilizes formulas to adjust weights from the output layer to hidden layers, as defined by equations, and to adjust weights from hidden layers to input layers [32]. Mean squared error [33] serves as a loss function, frequently employed in regression problems, where the output consists of continuous values or a vector of values. The neural network functions are represented in the Figure 3.

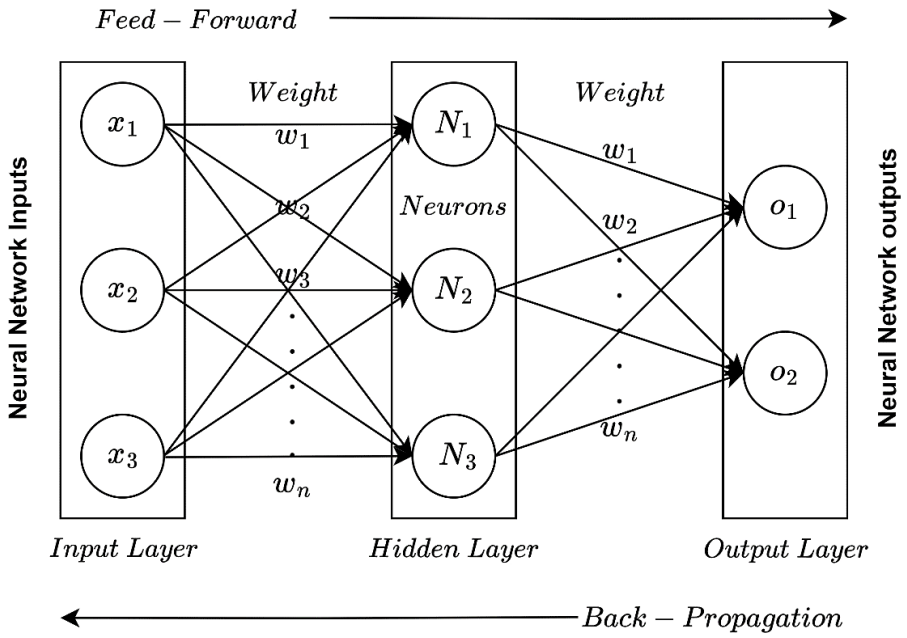

Figure 3. Feedforward Neural Network

Perform Neural network Steps:

- Feedforward Step 1: The weights of the input layer to the hidden layer nodes are calculated using the summation function and the Sigmoid function, as defined in Equations 3 and 4:

$$net_j = \sum_{i=1}^{N} w_{ji} x_i \tag{3}$$

$$O_j = f(net_j) = \frac{1}{1 + e^{-x}} \tag{4}$$

- Feedforward Step 2: Hidden layer to Output layer Nodes Calculating the weighted with sum function and Sigmoid functions as defined by Equations 5 and 6:

$$net_k = \sum_{j=1}^{M} w_{kj} O_j \tag{5}$$

$$O_k = f(net_k) = \frac{1}{1 + e^{-net_k}} \tag{6}$$

- Feedforward Step 3: Derivative of the Sigmoid function as defined by Equation 7:

$$\acute{f}(x) = \sigma(x)(1 - \sigma(x)) \tag{7}$$

- Backpropagation Step 1: Renew the weight from the output back to the hidden layer as defined by Equations 8, and (9), respectively.

$$w_{kj}^{new} = w_{kj}^{old} - \eta(\Delta w_{kj}) \tag{8}$$

$$\Delta w_{kj} = -(y_k - o_k) o_k (1 - o_k) o_j \tag{9}$$

- Backpropagation Step 2: Renew the weight from the hidden layer back to the input layer as defined by Equations 10 and 11, respectively. Where η is the learning rate, *w* is weight, *y* is the values of prediction, *o* is the actual value, and *x* is input.

$$w_{ji}^{new} = w_{ji}^{old} - \eta(\Delta w_{ji}) \tag{10}$$

$$\Delta w_{ji} = -\sum_{k=1}^{L} (y_k - o_k) o_k (1 - o_k) w_{kj} o_j (1 - o_j) x_i \tag{11}$$

- Backpropagation Step 3: Mean squared error is a loss function that is often used for Model performance, The vector of values as defined by Equation 12. Where n is all variable's number of prediction samples, y is observed values, and yˆ is the predicted value.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{12}$$

- Normalization is to normalize the input data so that it obeys a distribution with a mean of 0 and a variance of 1:

$$x = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (13)$$

- The sigmoid function is a logistic function characterized by an "S"-shaped curve (Equation 14 and Figure 4):

$$f(z) = \frac{1}{(1 + e^{-z})} \quad (14)$$

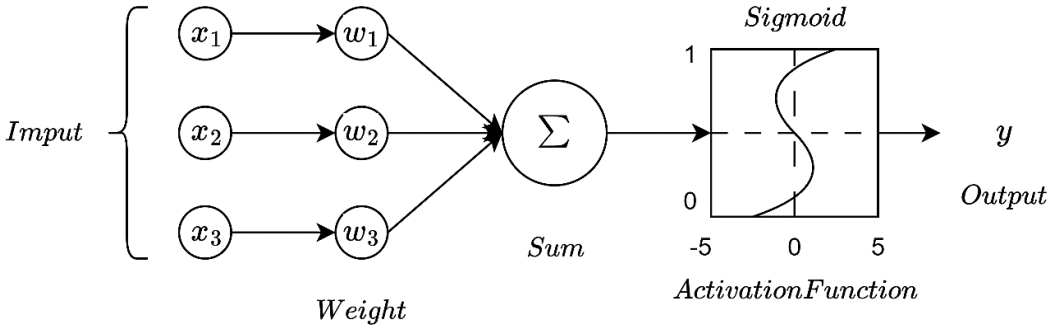

Figure 4. Active function Sigmoid

3.3. Handling of Overlapping Data

The overlapping area of the two clusters is taken as a new cluster. Set Cluster A centroid as *p* point, set Cluster B centroid as *q* point. Pythagorean theorem [34] is used to calculate the center point of the overlapping area as the centroid. The workflow for determining overlapping centers as shown in Figure 5. The formula is: $d = \frac{1}{2}\sqrt{p^2 + q^2}$.

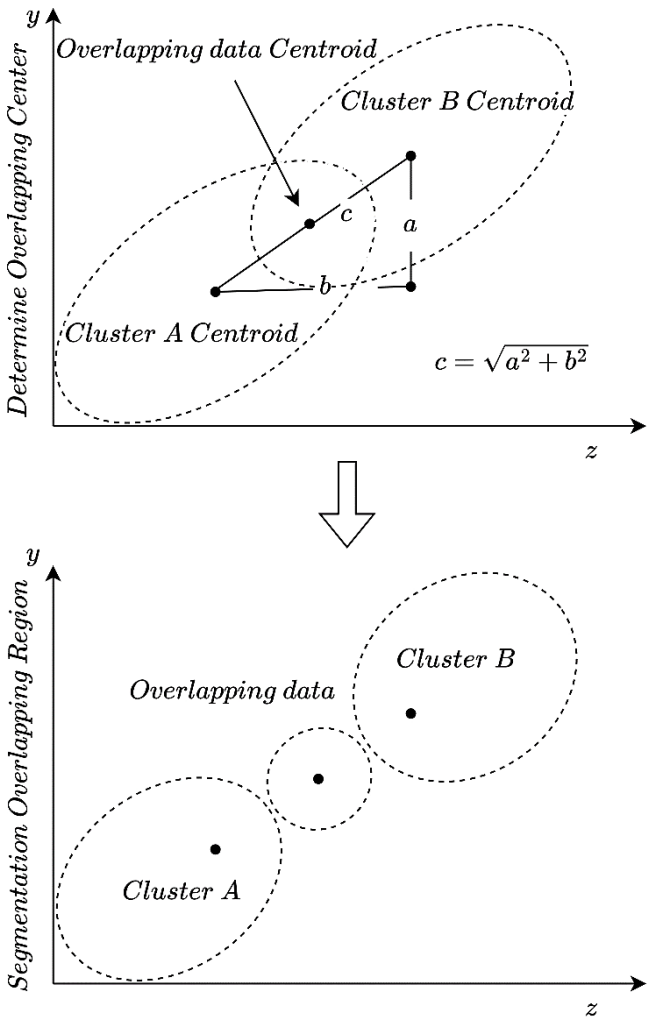

Figure 5. Handling of Overlapping Data

3.4. Parameter Setting

An overview of the hardware and software configuration used in the experiment and the parameters are shown in Table 1.

Table 1. Model Parameter Setting in Details

Model	Datasets	Dimensions	Distance function	Initialization method	Iteration	Features	Clusters	Train: Val Ratio (%)
K-means	Iris	Low	Euclidean distance	Random seeds	100	4	3	80:20
K-means	Wine	Medium	Euclidean distance	Random seeds	100	13	3	80:20
K-means	Breast Cancer	Hide	Euclidean distance	Random seeds	100	30	2	80:20

Model	Datasets	Classes	Train: Val Ratio (%)	Hidden Neurons	Learning Rate	Maximum Epochs	Activation Function	Early Stropping
F-NN	Iris	3	80:20	4	0.5	200	Sigmoid	Applied
F-NN	Wine	3	80:20	13	0.5	200	Sigmoid	Applied
F-NN	Breast Cancer	2	80:20	30	0.5	200	Sigmoid	Applied

Dataset	Operating system	Central Processing Unit	Processor	Random-Access Memory	Python	Anaconda
Iris	Windows 10 Education, version 22H2	Intel(R) Core (TM) i7-6700 CPU @ 3.40GHz	64-bit operating system, x64-based processor	Random-Access Memory (16.0 GB)	Version 3.9.12	Version 4.13.0
Wine						
Breast Cancer						

3.5. Dataset Diversity and Scalability

All experiments conducted in this study used public datasets downloaded from the UCI. Specifically, we used the Iris, Wine, and Breast Cancer datasets, where the Iris dataset contains four dimensions, the Wine dataset contains 13 dimensions, and the Breast Cancer dataset contains 30 dimensions. We experimented with these three datasets in low-dimensional, medium-dimensional, and high-dimensional datasets to obtain the effective performance of the model in different dimensions (see Figure 6). Table 2 provides more details about the experimental data.

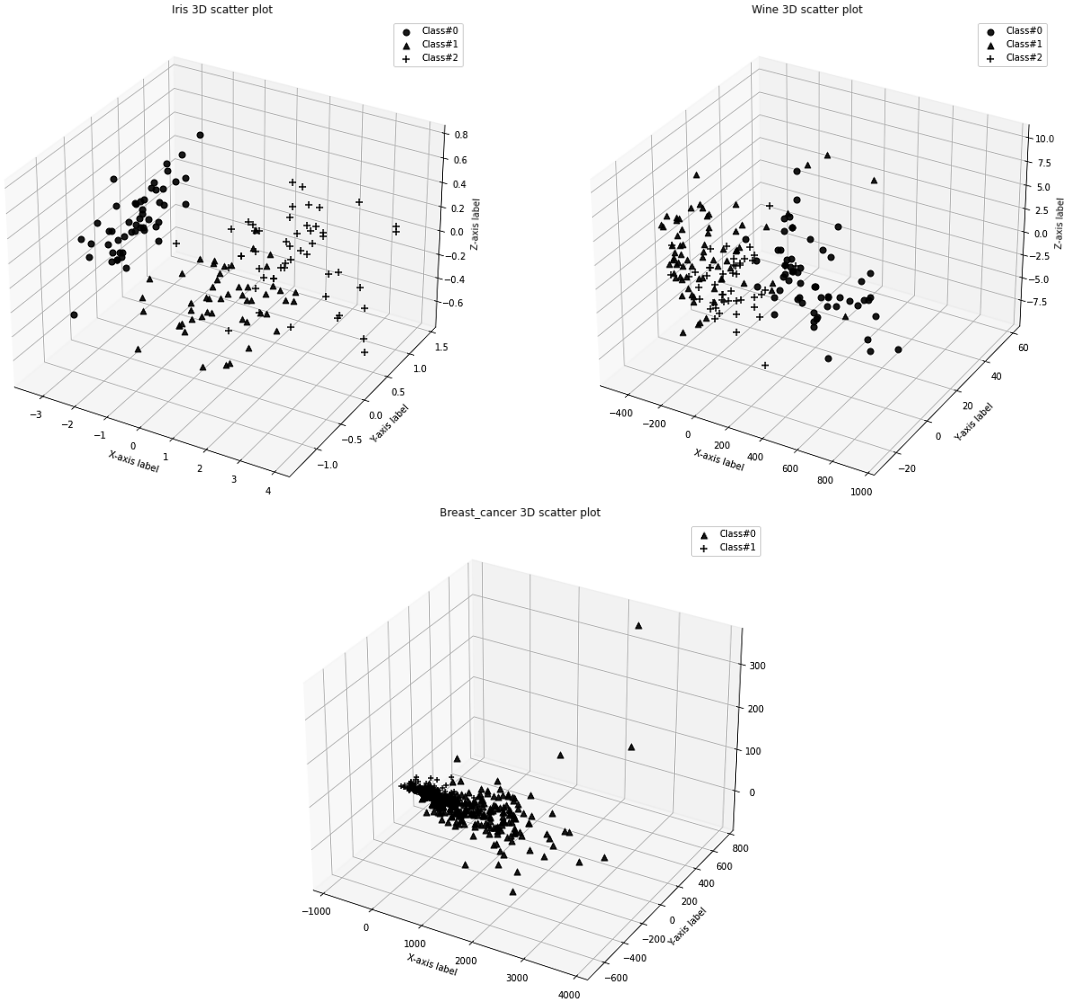

Figure 6. Dataset Diversity

Table 2. Experiment Dataset in Low, Medium, and High Dimensions Dataset

Dataset	Dimensions	Features	Class	Instances	Samples Per Class	Associated Tasks	Subject Area	Mission Values
Iris	Low	4	3	150	[50,50,50]	Classicization	Biology	No
Wine	Medium	13	3	178	[59,71,48]	Classicization	Physics and Chemistry	No
Breast Cancer	Hide	30	2	569	[212,357]	Classicization	Health and Medicine	No

3.6. Model Integration

This section delves into the specifics of the proposed segmental technique designed to address clusters overlapping. We propose a new classification model that uses a neural network algorithm to solve the overlapping areas in K-means classification. Our methodology outlines the approach for identifying overlapping clusters and introduces the segmental technique, shown in Figure 7.

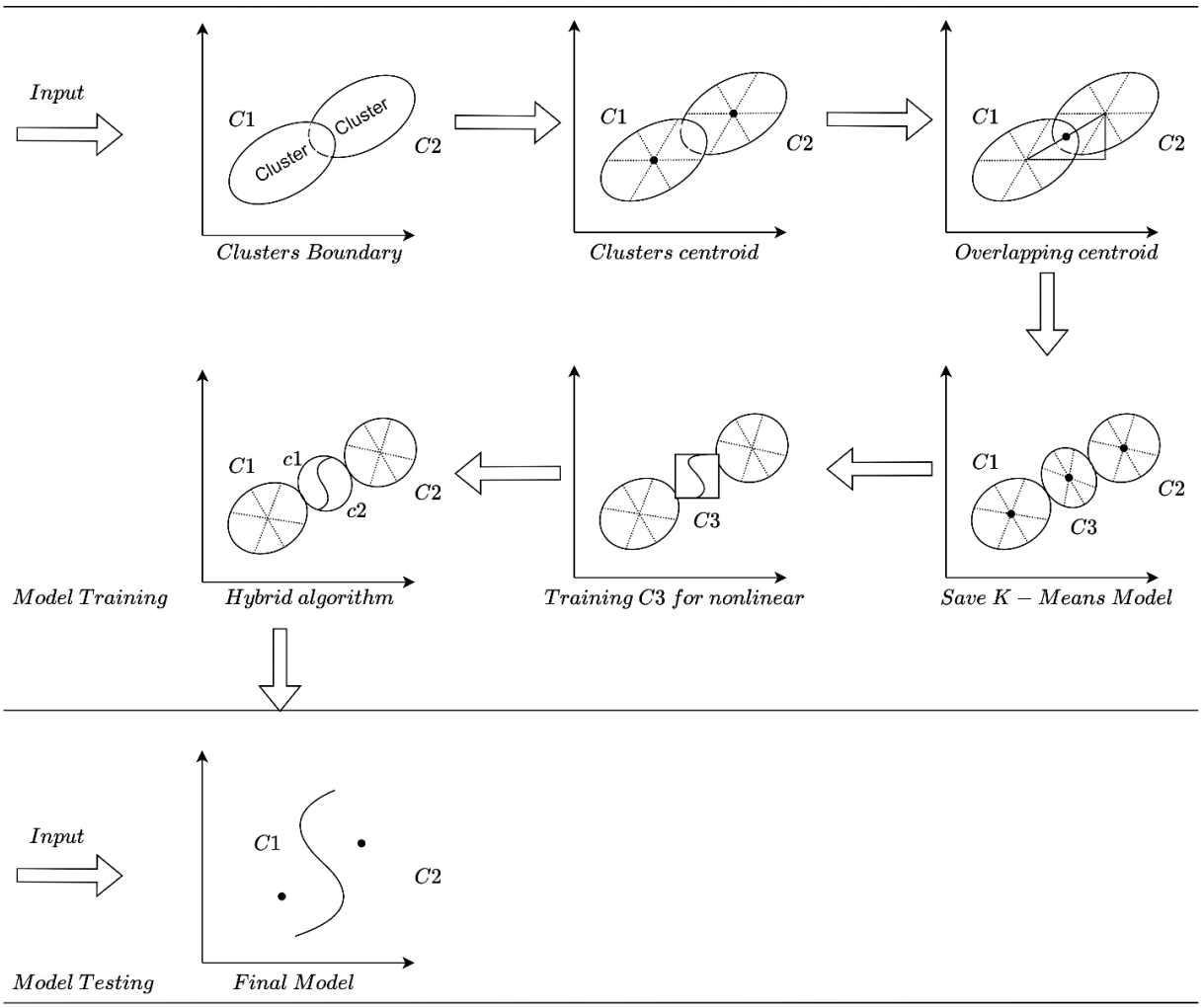

Figure 7. Process of Proposed Novel Classification Model

Our model for addressing overlapping problems involves seven key steps:

- The first step: involves inputting the cluster data afflicted by overlapping issues after applying the K-Means algorithm.
- The second step is calculating the center points of all clusters by the means.
- The third step is identifying overlapping areas.
- The fourth step is determining overlapping region data.
- The fifth step is training the Neural Networks model for overlapping region data.
- The sixth step: Integrating the features derived from the K-Means and Neural Networks models is performed to formulate the final model.
- The seventh step is evaluating the model's performance through rigorous assessment techniques.

4. Experiment and Validation

In the experiments, we used k-fold cross-validation to evaluate the robustness and generalization ability of the model design. The procedure is shown in Figure 8.

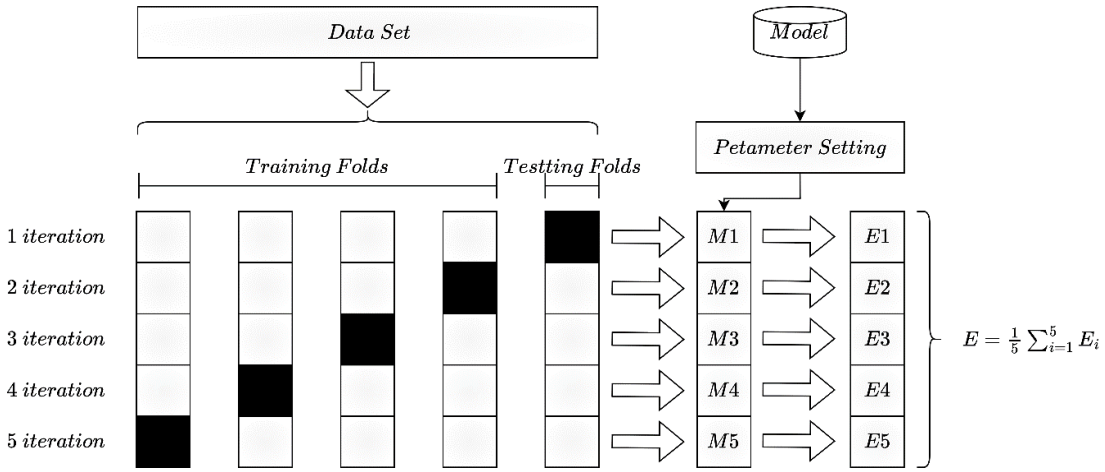

Figure 8. K-Fold Cross Validation

There are 4 mathematical methodologies for model performance measurement: Parameters: from model predicted class and actual class. There are TP: True Positive; FP: False Positive; FN: False Negative; and TN: True Negative respectively [35, 36]. The accuracy of the equation is shown in Equation 15, the Precision of the equation is shown in Equation 16, the Recall of the equation is shown in Equation 17, and the F1-score of the equation is shown in Equation 18.

$$Accuracy = \left(\frac{TN + TP}{TP + FP + TN + FN}\right) \times 100 \quad (15)$$

$$Precison = \frac{TP}{TP\ + FP} \quad (16)$$

$$Recall = \frac{TP}{TP\ +\ FN} \quad (17)$$

$$F1 - score = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right) \quad (18)$$

The standard deviation (or σ) is a measure of how dispersed the data is concerning the mean as defined by Equation 19.

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}} \quad (19)$$

Mean deviation is the distance in the cluster, it is as defined by Equation 20.

$$MD = \frac{1}{N}\sum_{i=1}^{n}|x_i - \bar{x}| \quad (20)$$

5. Experiment Results and Discussion

The Results of K-Fold cross-validation with K-Means as the control group and our model as the experimental group. To ensure the consistency of the results, the K-Means model and our model use the same folds for model training and model validation; performance results are given for the data environment of the model. The results are shown Table 3 and Figures 9 and 10.

Table 3. Experiment Results Performance in Low, Medium, and High Dimensions Dataset with Proposed Model

Dataset	Lower dimensions	Models	Accuracy	Precision	Recall	F1-Measure	Mean	Std
Iris	4	K-Means	0.8666	0.8666	0.8666	0.8666	0.8808	0.0584
		Proposed	0.9333	0.9333	0.9333	0.9333	0.8942	0.0536
Wine	13	K-Means	0.6666	0.6773	0.6626	0.6678	0.7059	0.0998
		Proposed	0.7820	0.8070	0.7820	0.7840	0.7890	0.0817
Breast Cancer	30	K-Means	0.7479	0.7263	0.6388	0.6952	0.7662	0.0406
		Proposed	0.9405	0.9218	0.9400	0.9428	0.8077	0.0394

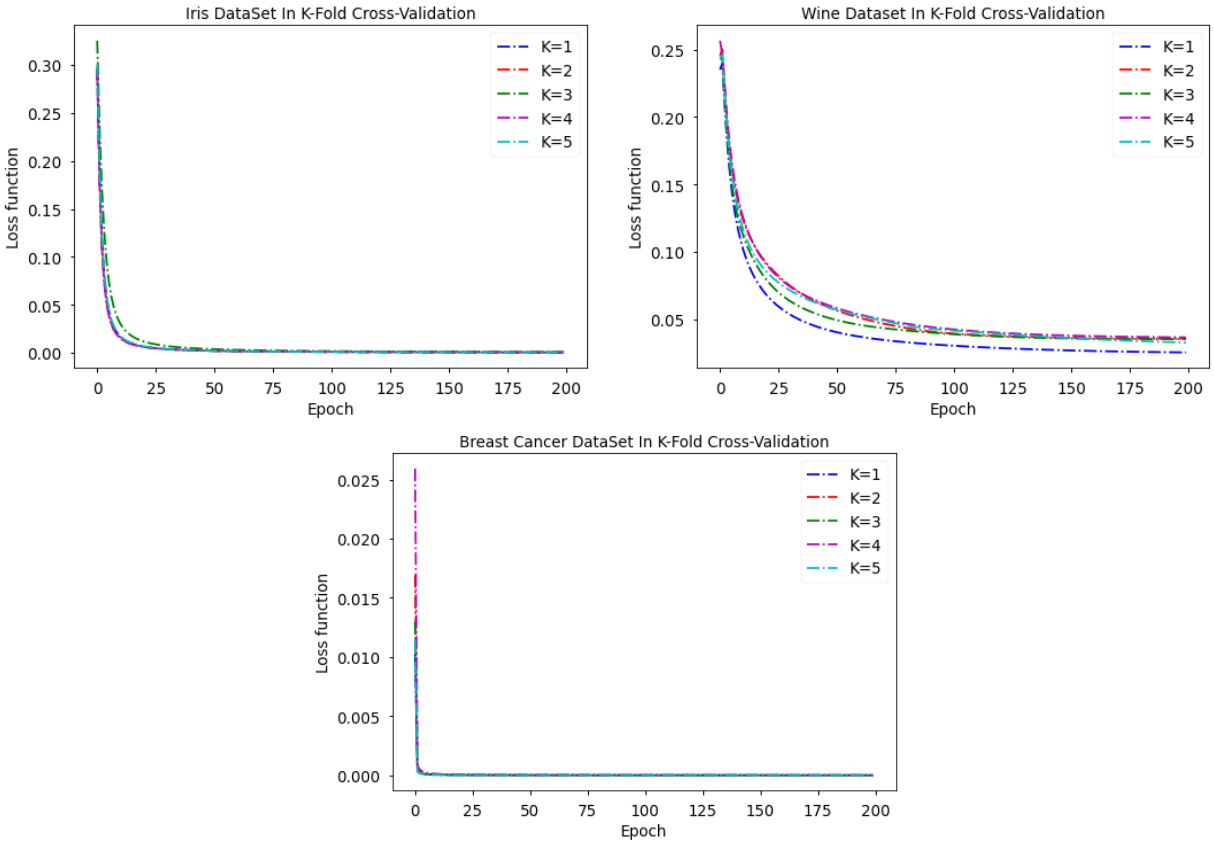

Figure 9. Loss function from Our Model. The K-Means Model and Our Model used the same folds data for training and testing; To ensure the consistency of the results, Both Model Setting are k=3 & Folds (1-5); train-validation-test splits=80% & 20%.

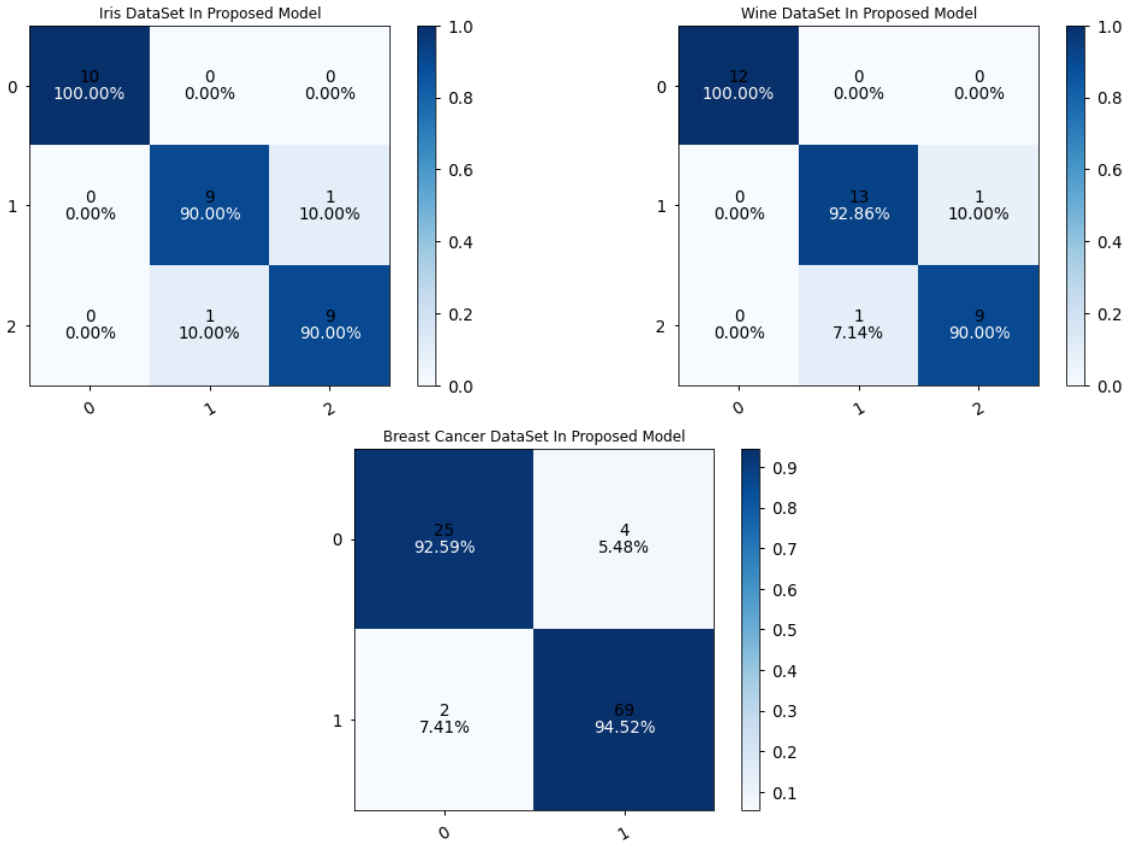

Figure 10. Confusion matrix from Our Model. K-Means Model and Our Model used the same folds data for training and testing; To ensure the consistency of the results, Both Model Setting is k=3 & Folds (1-5); train-validation-test splits=80% & 20%.

In summary, this section comprehensively evaluates the performance of our proposed model. The following is an overview of the key components used for evaluation:

Table 3: This table provides a detailed comparison of the performance metrics between the K-Means model as a control group and our proposed model. It provides a comprehensive overview of various metrics such as accuracy, precision, recall, and F1 score, allowing for a detailed evaluation of the performance of the model. To provide a more detailed picture of the performance of the model in terms of data diversity and scalability, we tested it on low, medium, and high-dimensional datasets. The dataset with four dimensions is called low dimensional, the dataset with 13 dimensions is called mid-dimensional, and the dataset with 30 dimensions is called high dimensional.

Figure 9: These plots show the performance of neural network model training with cluster overlapping data. We used 5 k-fold validation for training. So each dataset has 5 training performances. Where k represents k-fold. We use mean squared error (MSE) as the evaluation metric for the model. The iris dataset has 3 classes, and the ratio of training and evaluation is 80:20. There are 4 hidden neurons, the learning rate is 0.5, and the epochs are 200. The wine dataset has 3 classes, and the ratio of training and evaluation is 80:20. There are 13 hidden neurons, the learning rate is 0.5, and the epochs are 200. The breast dataset has 2 classes, and the ratio of training and evaluation is 80:20. There are 30 hidden neurons, the learning rate is 0.5, and the epochs are 200.

Figure 10: Confusion matrix of the model. They provide an intuitive representation of the model classification results and help understand the distribution of True Positive, True Negative, False Positive, and False Negative predictions. The prediction accuracy of the iris dataset is 100% for class 0, 90% for class 1, and 90% for class 2. The overall accuracy is 93.333%, which is higher than the original classification model. The prediction accuracy of the wine dataset is 100% for class 0, 92.86% for class 1, and 90% for class 2, with an overall accuracy of 78.20%, which is higher than the original classification model. The prediction accuracy of the breast cancer dataset is 92.59% for class 0, 94.52% for class 1, and 0.9405% for the overall accuracy, which is higher than the original classification model.

Figure 11 are used to show the classification performance between our proposed models, including the mean deviation, standard deviation, minimum score, maximum score, and median of indicators such as accuracy, precision, recall, and f1-score. They provide insights into the performance differences across various evaluation criteria. We show in detail the performance of our model and K-means model on low-latitude, mid-latitude, and high-latitude datasets respectively. We use mean deviation to detect the compactness of data within the clusters, and the smaller the number, the better the performance. Standard deviation is used to detect the size of cluster boundaries, and the larger the value, the larger the cluster. Thus, their robustness and generalization ability are comprehensively evaluated.

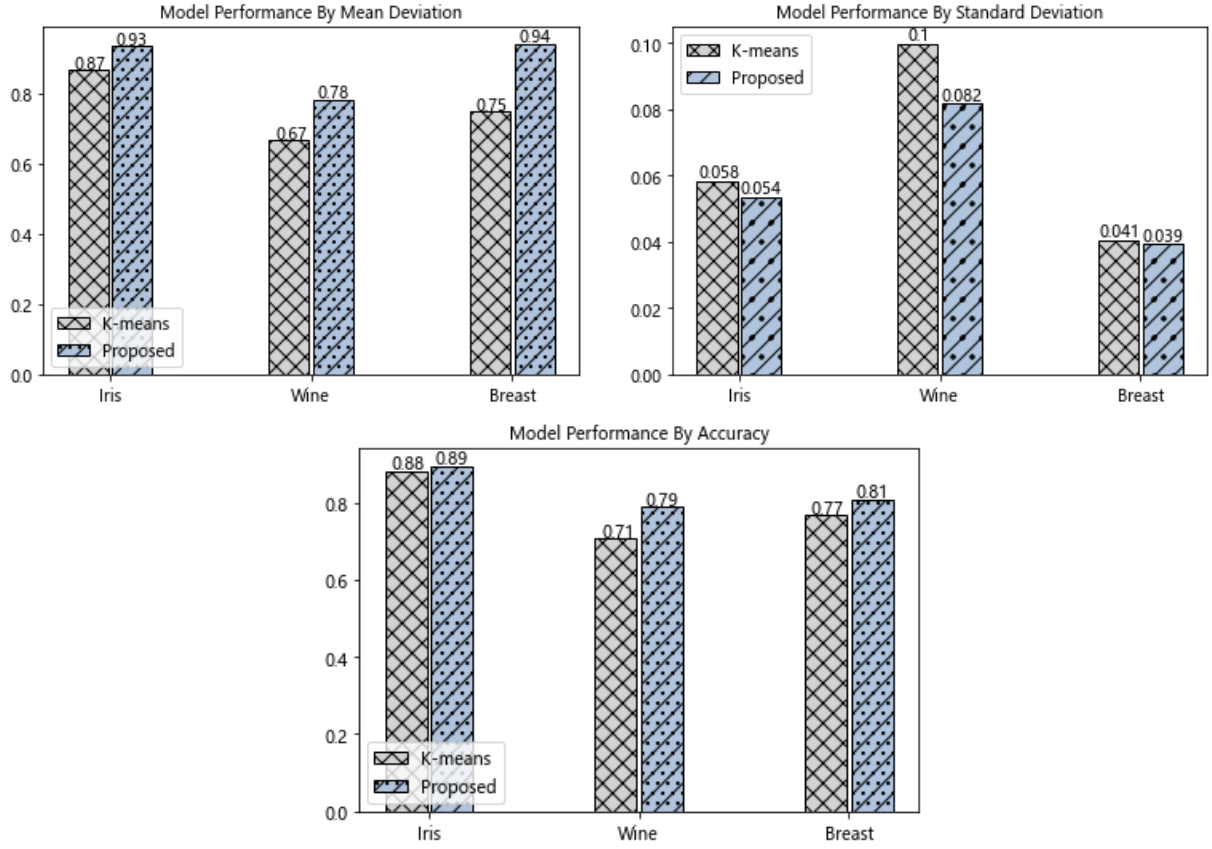

Figure 11. The bar chart comparison is from our model. To ensure the consistency of the results, the K-Means Model and Our Model used the same folds data for training and testing; Both Model Setting are k = 3 & Folds (1-5); train-validation-test splits = 80% & 20%.

In summary, we evaluated all model's performance with low, medium, and high dataset. With the experiment results the K-Means model and the proposed model have provided a comprehensive understanding of their strengths and weaknesses in various evaluation scenarios.

6. Conclusion

The study introduces a novel classification model that integrates features from both K-Means and neural networks. This algorithm is designed to identify areas of cluster overlap within classified data, effectively distinguishing categories that the K-Means algorithm alone finds difficult to discern.

A new classification model is presented to address the problem of data overlap in the clustering process. The model applies segmentation technology to divide the data in the overlapping area. Neural networks are then employed to train and identify this portion of the data, which is subsequently integrated into the classification model. This approach retains the characteristics of the K-Means algorithm while utilizing the neural network to identify data in the overlapping area. The primary focus of the model is on identifying the overlapping area and segmenting the data, as introduced in the paper.

Moreover, to evaluate our proposed model, we conducted experiments with 4, 13, and 30-dimensional datasets and compared its performance against a K-Means model. The experimental results indicate that our proposed model consistently outperforms K-Means across various cross-validation scenarios. While K-Means, as a traditional machine-learning model, classification accuracy in 3 scenarios with the public dataset, ranging from 86%, 66%, and 74%, our proposed model achieves significantly higher accuracy rates. Specifically, our model demonstrates 7%, 12%, and 20% higher accuracy compared to K-Means in the respective scenarios.

These findings provided the superiority of our proposed classification model over K-Means and validated its feasibility. Furthermore, our novel model has the potential ability for application in various classification problems. In future research, we intend to explore the applicability of our model across diverse data types and develop automatic optimization methods tailored to the unique characteristics of each dataset.

In future research, we intend to explore the applicability of our model across diverse data types and develop automatic optimization methods tailored to the unique characteristics of each dataset.

7. Declarations

7.1. Author Contributions

Corresponding, A.T.; conceptualization, C.C. and A.T.; methodology, C.C. and A.T.; software, C.C.; validation, C.C.; formal analysis, C.C. and A.T.; investigation, C.C. and A.T.; resources, C.C. and A.T.; data curation, C.C.; writing—original draft preparation, C.C.; writing—review and editing, C.C. and A.T.; visualization, C.C.; supervision, A.T.; project administration, C.C.; funding acquisition, A.T. All authors have read and agreed to the published version of the manuscript.

7.2. Data Availability Statement

The data presented in this study are available in the article.

7.3. Funding and Acknowledgements

This work was supported by King Mongkut's Institute of Technology Ladkrabang (KMITL), Thailand.

7.4. Institutional Review Board Statement

Not applicable.

7.5. Informed Consent Statement

Not applicable.

7.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

8. References

[1] Hassoun, A., Aït-Kaddour, A., Abu-Mahfouz, A. M., Rathod, N. B., Bader, F., Barba, F. J., Biancolillo, A., Cropotova, J., Galanakis, C. M., Jambrak, A. R., Lorenzo, J. M., Måge, I., Ozogul, F., & Regenstein, J. (2023). The fourth industrial revolution in the food industry—Part I: Industry 4.0 technologies. Critical Reviews in Food Science and Nutrition, 63(23), 6547–6563. doi:10.1080/10408398.2022.2034735.

[2] Meiring, G. A. M., & Myburgh, H. C. (2015). A review of intelligent driving style analysis systems and related artificial intelligence algorithms. Sensors (Switzerland), 15(12), 30653–30682. doi:10.3390/s151229822.

[3] Celebi, M. E., & Aydin, K. (2016). Unsupervised learning algorithms. Unsupervised Learning Algorithms, 8, 55-70. doi:10.1007/978-3-319-24211-8.

[4] Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. Electronics (Switzerland), 9(8), 1–12. doi:10.3390/electronics9081295.

[5] Azar, A. T., Gaber, T., Oliva, D., Ṭulbah, M. F., & Hassanien, A. E. (2020). Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020), 1153, 247–257.

[6] Chenghu, C., Jinna, H., Visavakitcharoen, A., Temdee, P., & Chaisricharoen, R. (2019). Identifying the effectiveness of arabica drip coffee on individual human brainwave. ECTI DAMT-NCON 2019 - 4th International Conference on Digital Arts, Media and Technology and 2nd ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering, 1–4. doi:10.1109/ECTI-NCON.2019.8692298.

[7] Chen, G., Liu, Y., & Ge, Z. (2019). K-means Bayes algorithm for imbalanced fault classification and big data application. Journal of Process Control, 81, 54–64. doi:10.1016/j.jprocont.2019.06.011.

[8] Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences, 622, 178–210. doi:10.1016/j.ins.2022.11.139.

[9] Wang, Z., Du, X., & Wu, L. (2022). AI-Based Secure Construction of University Information Services Platform. Security and Communication Networks, 2022(1), 1939796. doi:10.1155/2022/1939796.

[10] Khanmohammadi, S., Adibeig, N., & Shanehbandy, S. (2017). An improved overlapping k-means clustering method for medical applications. Expert Systems with Applications, 67, 12–18. doi:10.1016/j.eswa.2016.09.025.

[11] Fränti, P., & Sieranoja, S. (2018). K-means properties on six clustering benchmark datasets. Applied Intelligence, 48(12), 4743–4759. doi:10.1007/s10489-018-1238-7.

[12] Danganan, A. E., & De Los Reyes, E. (2021). Ehmcoke: An enhanced overlapping clustering algorithm for data analysis. Bulletin of Electrical Engineering and Informatics, 10(4), 2212–2222. doi:10.11591/EEI.V10I4.2547.

[13] Chen, Y. C., Chen, Y. L., & Lu, J. Y. (2021). MK-Means: Detecting evolutionary communities in dynamic networks. Expert Systems with Applications, 176, 114807. doi:10.1016/j.eswa.2021.114807.

[14] Nie, F., Li, Z., Wang, R., & Li, X. (2023). An Effective and Efficient Algorithm for K-Means Clustering with New Formulation. IEEE Transactions on Knowledge and Data Engineering, 35(4), 3433–3443. doi:10.1109/TKDE.2022.3155450.

[15] Sieranoja, S., & Fränti, P. (2022). Adapting k-means for graph clustering. Knowledge and Information Systems, 64(1), 115–142. doi:10.1007/s10115-021-01623-y.

[16] Gan, G., & Ng, M. K. P. (2017). K-Means Clustering with Outlier Removal. Pattern Recognition Letters, 90, 8–14. doi:10.1016/j.patrec.2017.03.008.

[17] Wang, P., Shi, H., Yang, X., & Mi, J. (2019). Three-way k-means: integrating k-means and three-way decision. International journal of machine learning and cybernetics, 10, 2767-2777. doi:10.1007/s13042-018-0901-y.

[18] Lu, X., Ye, X., & Cheng, Y. (2024). An overlapping minimization-based over-sampling algorithm for binary imbalanced classification. Engineering Applications of Artificial Intelligence, 133, 108107. doi:10.1016/j.engappai.2024.108107.

[19] Afridi, M. K., Azam, N., & Yao, J. T. (2020). Variance based three-way clustering approaches for handling overlapping clustering. International Journal of Approximate Reasoning, 118, 47–63. doi:10.1016/j.ijar.2019.11.011.

[20] Dai, Q., Wang, L. hui, Xu, K. long, Du, T., & Chen, L. fang. (2024). Class-overlap detection based on heterogeneous clustering ensemble for multi-class imbalance problem. Expert Systems with Applications, 255, 124558. doi:10.1016/j.eswa.2024.124558.

[21] Zhou, Q., & Sun, B. (2024). Adaptive K-means clustering based under-sampling methods to solve the class imbalance problem. Data and Information Management, 8(3), 100064. doi:10.1016/j.dim.2023.100064.

[22] Zhu, J., Jiang, Z., Evangelidis, G. D., Zhang, C., Pang, S., & Li, Z. (2019). Efficient registration of multi-view point sets by K-means clustering. Information Sciences, 488, 205–218. doi:10.1016/j.ins.2019.03.024.

[23] Ros, F., & Riad, R. (2024). Feature and Dimensionality Reduction for Clustering with Deep Learning. Springer Nature, XI, 268. doi:10.1007/978-3-031-48743-9.

[24] Lücke, J., & Forster, D. (2019). k-means as a variational EM approximation of Gaussian mixture models. Pattern Recognition Letters, 125, 349–356. doi:10.1016/j.patrec.2019.04.001.

[25] Liu, X., Fan, K., Huang, X., Ge, J., Liu, Y., & Kang, H. (2024). Recent advances in artificial intelligence boosting materials design for electrochemical energy storage. Chemical Engineering Journal, 490. doi:10.1016/j.cej.2024.151625.

[26] Vuttipittayamongkol, P., Elyan, E., Petrovski, A., & Jayne, C. (2018). Overlap-Based Undersampling for Improving Imbalanced Data Classification. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11314 LNCS, 689–697. doi:10.1007/978-3-030-03493-1_72.

[27] Huang, S., Kang, Z., Xu, Z., & Liu, Q. (2021). Robust deep k-means: An effective and simple method for data clustering. Pattern Recognition, 117, 107996. doi:10.1016/j.patcog.2021.107996.

[28] Saputra, D. M., Saputra, D., & Oswari, L. D. (2020). Effect of Distance Metrics in Determining K-Value in K-Means Clustering Using Elbow and Silhouette Method. Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019), 341–346. doi:10.2991/aisr.k.200424.051.

[29] Shi, H., Wang, P., Yang, X., & Yu, H. (2022). An Improved Mean Imputation Clustering Algorithm for Incomplete Data. Neural Processing Letters, 54(5), 3537–3550. doi:10.1007/s11063-020-10298-5.

[30] Anastassiou, G. A. (2023). Multiple general sigmoids based Banach space valued neural network multivariate approximation. Cubo, 25(3), 411–439. doi:10.56754/0719-0646.2503.411.

[31] Patel, J., Advani, H., Paul, S., & Maiti, T. K. (2022). VLSI Implementation of Neural Network Based Emergent Behavior Model for Robot Control. 2022 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics, DISCOVER 2022 - Proceedings, 197–200. doi:10.1109/DISCOVER55800.2022.9974734.

[32] Szu, H., Yeh, C., Rogers, G., Jenkins, M., Farsaie, A., & Lee, C. H. (1992). Speed up Performances on MIMD Machines. In Proceedings of the International Joint Conference on Neural Networks, 3, 742–747. doi:10.1109/IJCNN.1992.227063.

[33] Nguyen, V. A., Shafieezadeh-Abadeh, S., Kuhn, D., & Esfahani, P. M. (2023). Bridging Bayesian and Minimax Mean Square Error Estimation via Wasserstein Distributionally Robust Optimization. Mathematics of Operations Research, 48(1), 1–37. doi:10.1287/moor.2021.1176.

[34] Sefira, R., Setiawan, A., Hidayatullah, R., & Darmayanti, R. (2024). The Influence of the Snowball Throwing Learning Model on Pythagorean Theorem Material on Learning Outcomes. Journal Edutechnium Journal of Educational Technology, 2(1), 1–7.

[35] Cheng, Y., Li, Q., & Wan, F. (2021). Financial Risk Management using Machine Learning Method. Proceedings - 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence, MLBDBI 2021, 133–139. doi:10.1109/MLBDBI54094.2021.00034.

[36] Wang, L. H., Dai, Q., Wang, J. Y., Du, T., & Chen, L. (2024). Undersampling based on generalized learning vector quantization and natural nearest neighbors for imbalanced data. International Journal of Machine Learning and Cybernetics, 1–26. doi:10.1007/s13042-024-02261-w.