TOWARDS A NEVER-ENDING LEARNING MODEL OF RECYCLABLE MEDICAL WASTE

FENG Jiaqi
20231947
jiaqifeng077@gmail.com

SHAO Qichen
20232764
shaoqichen2000@gmail.com

ABSTRACT

Medical Waste classification is a critical step in achieving efficient processing and resource utilization. However, traditional techniques face significant limitations in classification efficiency, generalization capability, and adaptability to new waste categories. To address these challenges, this study proposes an intelligent medical waste classification model based on an Image Analysis Technology and Continual Learning framework, using the *Medical Waste 4.0* dataset as the experimental foundation. To systematically explore the impact of different training strategies on model performance, four progressively designed experiments were conducted : full-batch training, multi-task learning, independent task training, and incremental learning. These experiments aim to gradually uncover the model's potential and challenges in classification performance, convergence efficiency, and task adaptability.

The experimental results show that full-batch training provides a reliable baseline for optimizing the initial performance of the model; multi-task learning significantly improves the model's learning efficiency and shared feature extraction capability; independent task training further validates the effects of task interference; and incremental learning excels in mitigating catastrophic forgetting and enhancing adaptability to new categories. Through the progressive exploration of these four experiments, this study comprehensively investigates the design and optimization strategies for intelligent medical waste classification models and provides theoretical and practical support for achieving the goal of **Never-Ending Learning**. The significance of this research lies in advancing the technological development of intelligent medical waste management, reducing treatment costs, and minimizing the threats posed by medical waste to the environment and human health.

Key Words : Medical Waste, Multi Task Learning, Continual Learning

Encadré par: Monsieur Massinissa Hamidi

CONTENTS

1 INTRODUCTION

Medical waste management is a major global challenge for the healthcare industry, particularly in the classification phase of medical waste. Although the classification standards for medical waste vary across countries, textile materials (such as gauze and bandages) and plastic products (such as infusion bags and syringes) typically make up the majority. Additionally, sharp objects (such as needles and scalpels) account for approximately 12% of medical waste and are a primary source of injury and infection for healthcare workers and waste handlers. According to WHO's 2024 statisticsOrganization (2024), about 15% of medical waste is infectious, toxic, or radioactive. If improperly managed, these hazardous wastes not only pose a direct threat to human health but also pollute soil, water, and air, causing severe ecological damage and exacerbating the spread of diseases.

However, current medical waste management systems Zhou et al. (2022) show significant shortcomings in the classification phase. As the foundational step in waste management and disposal, inefficient and inaccurate waste classification often leads to increased complexity in subsequent disposal processes and may even result in secondary pollution or infection risks. Furthermore, with the continuous advancement of medical technologies, new types of medical waste are emerging, making traditional static classification methods increasingly inadequate to adapt to these dynamic changes. These issues hinder the standardization and scientific improvement of medical waste management and limit the industry's progression toward greater efficiency and sustainability. Effectively addressing the challenges of medical waste classification could significantly improve overall management efficiency and have far-reaching implications for environmental protection and public health.

To address these challenges, this study proposes a medical waste classification model that integrates Image Analysis Techniques with Continual Learning methods, featuring a "Never-Ending Learning" capability. The model is designed to continuously accumulate knowledge, dynamically update features, and adapt to new categories while maintaining classification performance for existing categories, meeting the long-term demands of complex medical waste classification scenarios. Through image analysis techniques, the system achieves automated waste classification, significantly reducing manual intervention and improving classification efficiency and accuracy. Meanwhile, the continual learning approach enables the system to dynamically adapt to new categories, alleviating the problem of "catastrophic forgetting" and ensuring stable classification performance.

To validate the model's performance and its feasibility in achieving never-ending learning, this study designs four progressive experiments: Full-Batch Training, Independent Task Training, Multi-Task Learning, and Continual Learning. These experiments progressively explore the potential and challenges of different training strategies in terms of classification efficiency, feature extraction optimization, and task adaptability.

2 CONTINUAL LEARNING SETUP

In continual learning Chaudhry et al. (2020), a learner encounters a stream of data triplets (x_i, y_i, t_i), where x_i is an input, y_i is the corresponding target, and $t_i \in \mathcal{T} = \{1, \ldots, T\}$ is the task identifier. For each task t_i, the input-target pair $(x_i, y_i) \in \mathcal{X}_{t_i} \times \mathcal{Y}_{t_i}$ and is drawn independently and identically from an unknown distribution $P_{t_i}(\mathcal{X}, \mathcal{Y})$. This distribution characterizes the t_i-th learning task.

We assume that tasks are performed sequentially, meaning $t_i \leq t_j$ for all $i \leq j$, and the total number of tasks T is not predetermined. **//Doubtful**

Under this setup, our goal is to estimate a predictor $f_\theta = (w \circ \phi) : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$, parameterized by $\theta \in \mathbb{R}^P$, where $\mathbb{R}$ represents the set of real numbers, and P denotes the dimensionality of the model's parameter space. The predictor is composed of a feature extractor $\phi : \mathcal{X} \to \mathcal{H}$ and a classifier $w : \mathcal{H} \to \mathcal{Y}$, aiming to minimize the multi-task error.

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{(x,y)\sim P_t} \left[\ell(f(x,t), y) \right] \tag{1}$$

The objective is to minimize the average error across multiple tasks, where $\ell(f(x,t), y)$ represents the loss function measuring the discrepancy between the model's prediction and the true label. First,

the average loss is computed for each task t under its data distribution P_t using the expectation operator. Then, the overall average error across all tasks is calculated to evaluate the model's performance in a multi-task setting.

The average accuracy of a predictor is defined as

$$Accuracy_t = \frac{1}{t}\sum_{j=1}^{t} a_{t,j} \quad t \in \{1, \ldots, T\} \tag{2}$$

where $a_{i,j}$ denotes the test accuracy on task j after the model has finished experiencing task i. The average accuracy for each task is calculated by averaging the test accuracies over all tasks j such that $j \leq i$.

3 DATASETS

This dataset Bruno et al. (2023a) was created by Bruno Antonio, Massimo Martinelli, and Davide Moroni following a selection process of the most commonly used types of medical items in hospitals. The data was collected using new medical equipment to simulate medical waste. The **Medical-Waste-4.0-Dataset** Bruno et al. (2023b) was collected as part of the "Medical Waste Treating 4.0" project funded by the Tuscany Region, aimed at supporting research related to medical waste management, particularly in the fields of machine learning and computer vision.

The dataset includes the following categories : gauze, latex gloves (single and pairs), nitrile gloves (single and pairs), surgical gloves (single and pairs), medical caps, medical goggles, shoe covers (single and pairs), test tubes, and urine bags.

Classification	**Sample size**
Gauze	393
Glove pair latex	330
Glove pair nitrile	330
Glove pair surgery	300
Glove single latex	303
Glove single nitrile	333
Glove single surgery	306
Medical cap	306
Medical glasses	318
Shoe cover pair	351
Shoe cover single	312
Test tube	363
Urine bag	300
Total dataset size	**4,245**

Table 1: Dataset category and corresponding sample size

Each sample in the dataset consists of three images : a high-resolution RGB image (1920 x 1080) and a stereo pair of grayscale images, each with a resolution of 640 x 400. The naming conventions for the files are as follows : timestamp.jpg for the RGB image, timestamp_r.png for the right view, and timestamp_l.png for the left view in the stereo pair. A public dataset containing over 1400 image triplets has already been released, while a more structured dataset with over 2100 image triplets will be published in the near future.

This dataset is designed to be a valuable resource for devising and testing computer vision methods for the primary sorting of medical waste.

3.1 SEPARATION OF TASKS

To enhance the structure and organization of the dataset, we classified the categories into three distinct tasks based on their functionality, material characteristics, and usage state : gloves (Task 1), shoe covers (Task 2), and other medical items (Task 3).

Task 1 includes medical gloves, further divided by material type (latex, nitrile, and surgical-specific) and usage state (pairs or singles). This classification is based on the material properties and the form in which they are discarded, facilitating sorting and processing. This category includes the following items :

- `glove_pair_latex`
- `glove_pair_nitrile`
- `glove_pair_surgery`
- `glove_single_latex`
- `glove_single_nitrile`
- `glove_single_surgery`

By organizing gloves based on material, this task enables a more granular understanding of glove waste, such as distinguishing disposable latex gloves from the more robust surgical ones. This classification is crucial for optimizing recycling or disposal processes based on material types.

Task 2 consists of medical shoe covers, categorized by whether they are paired or single. This straightforward classification reflects the basic form differences of shoe covers during usage or disposal. This category includes the following items :

- `shoe_cover_pair`
- `shoe_cover_single`

The separation between single shoe covers and pairs mirrors their practical use in hospitals and clinics, where single covers might be used as spares, while pairs are standard for operations. Grouping shoe covers as a standalone task highlights their unique role and simplifies the identification of footwear-related waste.

Task 3 encompasses other medical items such as urine bags, gauze, medical caps, medical glasses, and test tubes. These are categorized based on their unique functionalities (e.g., liquid collection, protection, experimental purposes) and item-specific characteristics, addressing different processing needs. This category includes the following items :

- `urine_bag`
- `gauze`
- `medical_cap`
- `medical_glasses`
- `test_tube`

This grouping reflects the functional diversity of these items while ensuring their waste characteristics are considered together. These items are not as frequently disposed of as gloves or shoe covers but still constitute a significant portion of medical waste, particularly in diagnostic and procedural settings.

Overall, this separation of tasks ensures a logical organization of the dataset, improving the clarity and usability for downstream machine learning tasks. Each task aligns with specific medical waste management processes, facilitating targeted analysis and model development for different categories of medical waste.

4 EXPERIMENTS

In the process of building and optimizing a model, selecting an appropriate training method is crucial. Different training strategies significantly impact the model's performance, convergence speed,

and computational resource requirements. To achieve the final goal, we designed and conducted a series of experiments to explore the characteristics and applicability of various strategies, including ***full-batch training***, ***independent task training***, ***multi-task learning*** and ***continual learning***. The following sections will provide a detailed explanation of each method's design approach, experimental process, model architecture, and corresponding result analysis.

4.1 DATASET TRANSFORM

In convolutional neural networks, input images are required to have a fixed size, as convolutional operations and fully connected layers rely on consistent input dimensions. Mismatched image sizes can result in dimensional incompatibility errors. Furthermore, when models are trained in batches, each batch must consist of uniformly sized samples to facilitate efficient processing.

However, image datasets often originate from diverse sources with varying resolutions and aspect ratios, ranging from high-resolution images (e.g., 1920×1080) to low-resolution ones (e.g., 640×400). By applying the Resize operation during data loading, all images can be adjusted to a consistent size, thereby simplifying data preprocessing, ensuring uniformity across samples, and improving the overall efficiency of training and processing pipelines.

Additionally, while high-resolution images capture more details, they significantly increase memory usage and computational costs. For many tasks, retaining the full detail of the original resolution is unnecessary. Thus, resizing images to an appropriate size effectively reduces computational overhead and optimizes model performance.

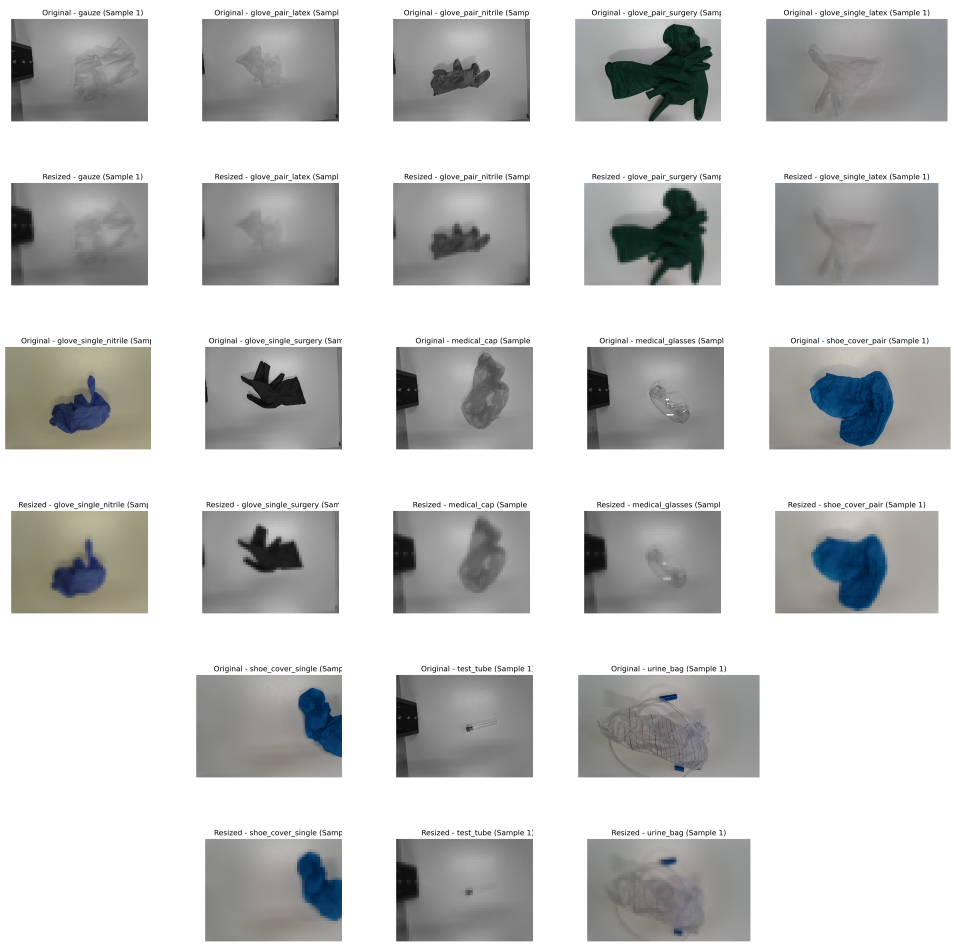


Figure 1: Original and Resize Dataset sample

The Resize operation adjusts an image to a specified target height and width, here we adjust images to (40, 64), by processing the original pixel matrix. For each target pixel, its corresponding position in the original image is calculated based on the scaling ratio. The pixel value is then determined using bilinear interpolation, a method that considers the four nearest pixels in the original image. This interpolation calculates the target pixel's value as a weighted average of these four neighbors, with weights based on their distances to the target position. This process effectively produces a resized image matrix with the desired dimensions, preserving the overall structure and quality of the original image while ensuring smooth transitions between pixels.

By selecting the first image from each category and comparing its original version with the one processed by the Resize operation, we can visually observe the overall reduction in pixel count and the change in dimensions. To facilitate comparison, the resized images are displayed at the same scale as the original images, allowing for a clearer examination of the details in both. The comparison reveals that the resized images exhibit a certain degree of blurriness due to the reduction in pixel count, resulting in a loss of details and slightly lower clarity compared to the original images. This blurriness is an unavoidable consequence of downscaling but, within the target dimensions suitable for the task, key features are typically preserved.

The input image is converted from a PIL image or NumPy array format to a PyTorch Tensor format, with pixel values normalized from the range 0-255 to 0-1. Additionally, the image data is standardized to have a mean of 0.5 and a standard deviation of 0.5, enhancing the stability and performance of model training.

4.2 Experiment 1 : Full-Batch Training - Flatten

In the full-batch training method, the entire dataset is treated as a single entity without task partitioning. During the training process, a batch of data is drawn from the entire dataset and fed into the model for training, repeating this process until the training is complete. The advantage of this approach is that it allows for a comprehensive evaluation of the model's learning ability on the entire dataset and provides an intuitive view of its performance during global training.

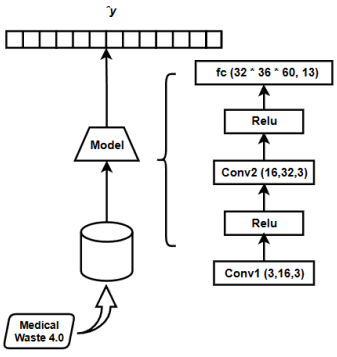


Figure 2: model_exp1

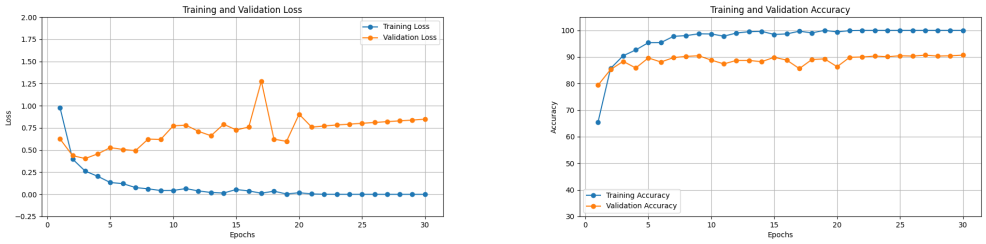
To better illustrate the training process under this strategy, we have drawn the model structure shown in the figure above. The diagram clearly depicts the data flow in full-batch training and how it is processed within the model.

Hyper-parameter	Value
Validation split	0.2
Number of layers	3
Number of Neurons in Conv1	$16 \times 38 \times 62 = 37{,}696$
Number of Neurons in Conv2	$32 \times 36 \times 60 = 69{,}120$
Number of Neurons in FC	13
Activation function	ReLU
Dropout	Not used
Learning rate	0.001
Optimizer	Adam
Batch size	16

Table 2: Hyper-parameter of model flatten

The hyper-parameter values used in the experiment are shown in the table above.

4.2.1 Analysis of Model Training Methods



(a) Loss-flatten

(b) Accuracy-flatten

Figure 3: Comparison of Training and Validation Metrics

Accuracy and Loss Rate Analysis

From these two graphs, which describe the evolution of **accuracy** and **loss** during training and validation, the following conclusions can be drawn :

Training and Validation Accuracy (Figure-Accuracy)

- Training accuracy increases rapidly and eventually reaches nearly 100%, indicating a strong ability of the model to fit the training data.
- Validation accuracy improves steadily during the first 8 epochs but starts to fluctuate between epochs 8 and 15, failing to improve further. It remains lower than training accuracy, stabilizing around 90%.

Training and Validation Loss (Figure-Loss)

- Training loss decreases rapidly at the beginning, showing that the model effectively learns the features of the training data.
- Validation loss follows a similar downward trend initially but stagnates between epochs 8 and 15, then starts to fluctuate and even increase (notably around epoch 25).

Phenomenon Analysis

Model Performance

- The model performs excellently on the training set (*training accuracy close to 100%, training loss close to 0*), demonstrating a strong ability to fit the training data.

- However, performance on the validation set is less satisfactory (*validation accuracy remains below 90% with fluctuations, and validation loss increases*), revealing limited generalization capability on unseen data.

Overfitting Phenomenon

- Starting from epoch 8, the model shows signs of overfitting Li et al. (2024): training loss continues to decrease, while validation loss ceases to decrease and begins to fluctuate or even increase.

- Training accuracy is significantly higher than validation accuracy, and validation performance stabilizes or deteriorates, confirming that the model overfits the training data in the later epochs.

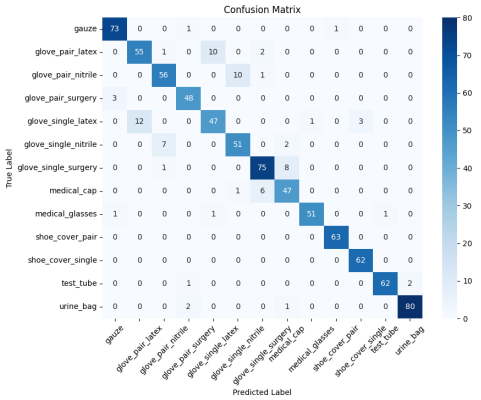


Figure 4: Confusion-matrix_flatten

Confusion Matrix Analysis

From the confusion matrix, it is evident that the model performs well in most categories, but there are some misclassification issues in certain easily confused categories. Notably, there is significant confusion between glove-related categories, such as glove_pair_latex and glove_single_latex, as well as glove_single_nitrile and glove_pair_nitrile, likely due to their similar material or physical characteristics, with noticeable confusion in distinguishing the number of gloves. Additionally, there is some misclassification between medical_cap and medical_glasses, indicating that the model struggles to differentiate the features of these categories effectively.

4.3 Experiment 2 : Independent Task Training

Consistent with the data partitioning approach used in multi-task learning, we also divided the dataset into three tasks. However, in this method, each task is trained independently. We trained a separate model for each task and evaluated the impact of task separation on model performance by recording the accuracy and loss values for each task. This experiment also serves as a baseline for comparison with other methods.

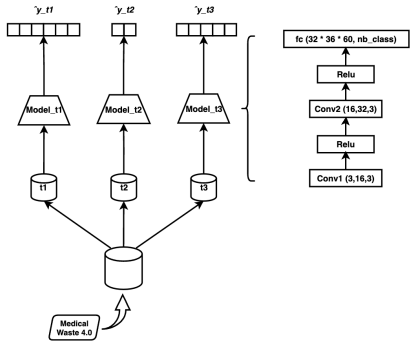


Figure 5: model_exp2

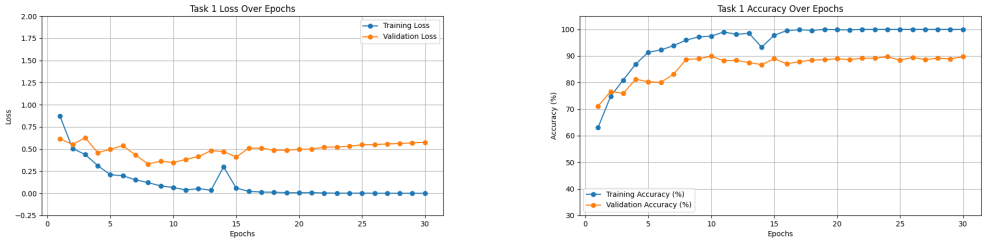
The purpose of this method is to assess the impact of task separation on model training performance and to explore the optimization potential in scenarios where there is no interference between tasks. To facilitate comparison and understanding, we also created a corresponding model diagram that visually illustrates the implementation process of independent task training.

Hyper-parameter	**Value**
Validation split	0.2
Number of layers	3
Number of Neurons in Conv1	$16 \times 38 \times 62 = 37{,}696$
Number of Neurons in Conv2	$32 \times 36 \times 60 = 69{,}120$
Number of Neurons in FC_T1	6
Number of Neurons in FC_T2	2
Number of Neurons in FC_T3	5
Activation function	ReLU
Dropout	Not used
Learning rate	0.001
Optimizer	Adam
Batch size	16

Table 3: Hyper-parameter of model specific

The hyper-parameter values used in the experiment are shown in the table above.

4.3.1 Analysis of Model Training Methods



(a) Loss-task_1 (b) Accuracy-task_1

Figure 6: Loss and Accuracy of task_1

Accuracy Rate Analysis

Training accuracy rises rapidly and plateaus near 100% after the 15th epoch, indicating excellent performance on the training dataset.

Validation accuracy stabilizes around 90% after the 10th epoch but remains lower than training accuracy, suggesting possible overfitting.

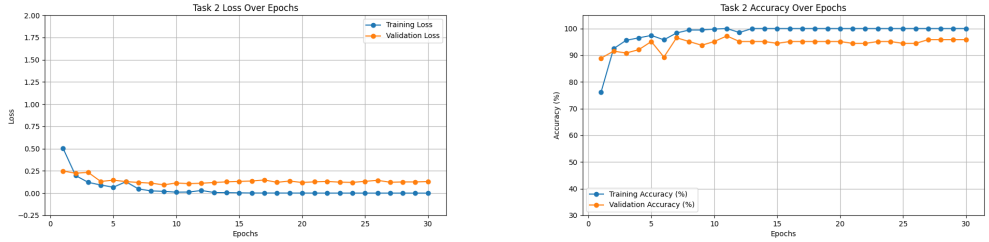
Loss Rate Analysis

Training loss decreases rapidly and approaches zero, indicating minimal error on the training dataset.

Validation loss stabilizes after an initial decline but remains significantly higher than training loss, with some fluctuations, further highlighting limited generalization ability.

Phenomenon Analysis

The model performs well on the training dataset but shows signs of overfitting on the validation dataset.



(a) Loss-task_2 (b) Accuracy-task_2

Figure 7: Loss and Accuracy of task_2

Accuracy Rate Analysis

Training accuracy rises quickly in the first few epochs and stabilizes near 100% after the 5th epoch, indicating excellent performance on the training dataset.

Validation accuracy increases rapidly in the initial epochs and stabilizes around 90%-92%, slightly lower than training accuracy, suggesting mild overfitting.

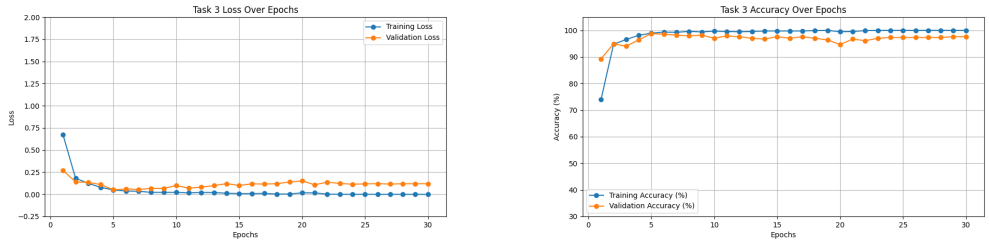
Loss Rate Analysis

Training loss decreases sharply during the first few epochs and approaches zero, showing minimal error on the training data.

Validation loss drops initially and stabilizes at a low level (around 0.1 to 0.2), with a small gap compared to training loss, indicating good generalization without significant overfitting.

Phenomenon Analysis

The model demonstrates excellent training performance and good generalization on the validation set.



(a) Loss-task_3 (b) Accuracy-task_3

Figure 8: Loss and Accuracy of task_3

Accuracy Rate Analysis

Training accuracy increases rapidly and stabilizes near 100% after the 5th epoch, demonstrating excellent learning on the training set.

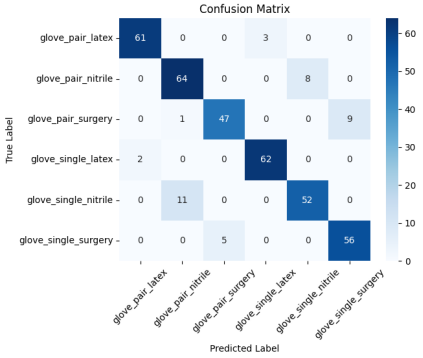
Validation accuracy stabilizes around 90%-92%, slightly lower than training accuracy, indicating mild overfitting but overall good generalization.

Loss Rate Analysis

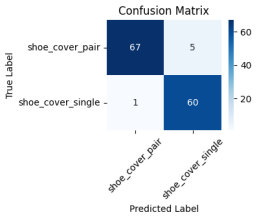
Training loss drops sharply in the early epochs and approaches zero, reflecting minimal error on the training data. Validation loss stabilizes at a low level (around 0.1 to 0.2), with a small gap compared to training loss, suggesting the model generalizes well.

Phenomenon Analysis

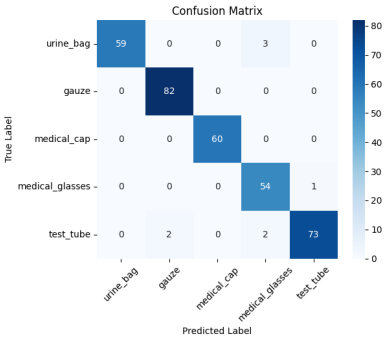
The model achieves excellent training performance and good validation accuracy.



(a) Confusion-matrix_task_1



(b) Confusion-matrix_task_2



(c) Confusion-matrix_task_3

Figure 9: Confusion-matrix

Confusion Matrix Analysis

Task 1

Main Confusion : There is significant confusion between glove_single_nitrile and glove_pair_nitrile, glove_single_latex and glove_pair_latex, as well as glove_single_surgery and glove_pair_surgery.

Reason : The confusion is primarily caused by the difficulty in distinguishing between single and paired gloves made of the same material.

Task 2

Main Confusion : There is minor confusion between shoe_cover_pair and shoe_cover_single, where 5 samples of shoe_cover_pair were misclassified as shoe_cover_single, and 1 sample of shoe_cover_single was misclassified as shoe_cover_pair.

Reason : The similar appearance of single and paired shoe covers, especially when limited image information is available, makes them difficult to distinguish.

Task 3

Main Confusion : urine_bag has 3 samples misclassified as medical_glasses. test_tube has 2 samples misclassified as gauze and medical_glasses.

Reason : These items may share similarities in certain angles or material properties, such as transparency or similar texture features.

4.4 Experiment 3 : Multi-Task Learning with Dynamic Batch size

Based on the multi-task learning framework, we divided the data into three related tasks. In each training iteration, a batch of data is dynamically sampled from each task. During training, the data first passes through the shared backbone for feature extraction and is then routed to different head models according to the task type for task-specific learning.

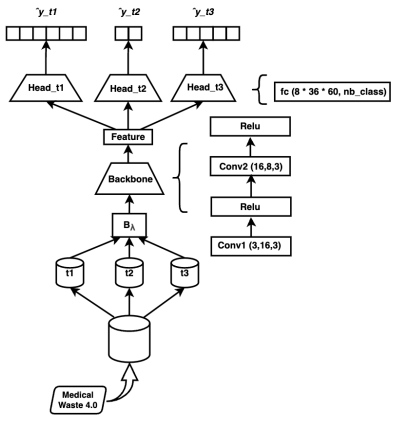


Figure 10: model_exp3

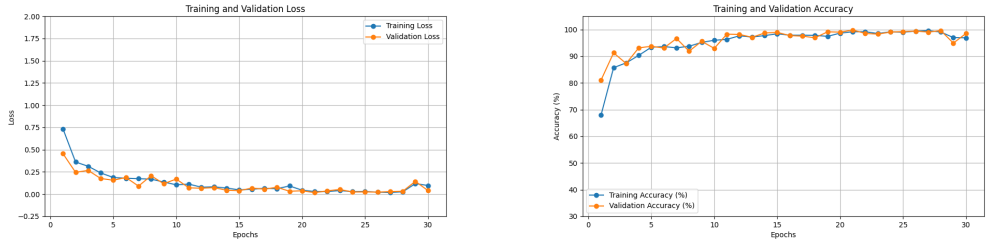
The figure above illustrates the structure of the backbone and head models, as well as the data flow between them under this strategy. The purpose of this strategy is to maximize the utilization of shared features across tasks while preserving the independent optimization direction of each task. We recorded the model's accuracy and loss variations for each task to analyze the collaborative effects among tasks and the learning efficiency of dynamic sampling.

Hyper-parameter	Value
Validation split	0.2
Number of layers	3
Number of Neurons in Conv1	$16 \times 38 \times 62 = 37{,}696$
Number of Neurons in Conv2	$8 \times 36 \times 60 = 17{,}280$
Number of Neurons in FC_T1	6
Number of Neurons in FC_T2	2
Number of Neurons in FC_T3	5
Activation function	ReLU
Dropout	Not used
Learning rate	0.001
Optimizer	Adam
Batch size of Task 1	23
Batch size of Task 2	8
Batch size of Task 3	28

Table 4: Hyper-parameter of model multitask

The hyper-parameter values used in the experiment are shown in the table above.

4.4.1 Analysis of Model Training Methods



(a) Loss-MTL

(b) Accuracy-MTL

Figure 11: Comparison of Training and Validation Metrics

Accuracy and Loss Rate Analysis

From the two graphs, which describe the evolution of **accuracy** and **loss** during training and validation, the following conclusions can be drawn:

Training and Validation Accuracy (Figure-Accuracy)

- Training accuracy increases rapidly in the first few epochs and eventually reaches 100%, demonstrating the strong ability of the model to fit the training data.
- The validation accuracy improves steadily during the first 10 epochs, reaching approximately 95%, and then stabilizes. Minor fluctuations are observed around epochs 15 and 20, but the overall trend remains stable.

Training and Validation Loss (Figure-Loss)

- Training loss decreases rapidly during the initial epochs and approaches 0 after epoch 10, indicating effective learning from the training data.
- The validation loss follows a similar decreasing trend initially but stabilizes after Epoch 10, with occasional fluctuations. Slight increases are observed around epochs 15 and 20, but these do not significantly affect the overall trend.

Phenomenon Analysis

Model Performance

- The model performs exceptionally well on the training set (*training accuracy close to 100%, training loss close to 0*), showing excellent learning capabilities for the training data.
- Validation set performance is satisfactory, with validation accuracy stabilizing around 95% and validation loss remaining at a low level, indicating good generalization ability.

Confusion Matrix Analysis

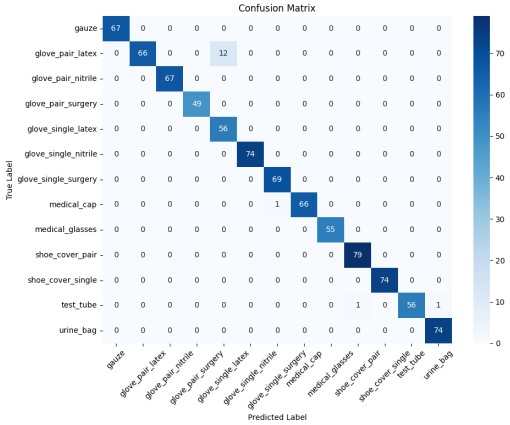


Figure 12: Confusion-matrix_MTL

In the Multi-Task Learning (MTL) experiment, the model achieved the highest classification accuracy among all experiments, with the majority of predictions concentrated along the diagonal of the confusion matrix, demonstrating exceptional classification performance. Notably, for categories such as gauze, glove_pair_nitrile, and shoe_cover_pair, the model achieved nearly 100% accuracy, highlighting the advantages of the MTL approach in feature extraction and sharing. However, some confusion was observed for similar categories, such as glove_pair_latex and glove_pair_surgery, where 12 glove_pair_latex samples were misclassified as glove_pair_surgery, likely due to their highly similar appearance or features. Additionally, misclassification in a small number of cases, such as between test_tube and urine_bag, could be attributed to imbalanced data distribution or overlapping visual features. Overall, the results of the MTL experiment validate the effectiveness of the multi-task learning framework, demonstrating its ability to significantly enhance classification performance in dynamic environments.

Dvantages

- In this approach, the dataset is split into multiple tasks, and during each training epoch, a batch of data is dynamically sampled from each task. Compared to the first method, where data from a single class is fed into the model in a concentrated manner, this dynamic sampling strategy demonstrates significant advantages in terms of training efficiency and generalization.
- Furthermore, in the head layer, task-specific outputs are separated by tasks, which further improves performance by leveraging task-specific distinctions.

4.5 Experiment 4 : Continual Learning

Sequential incremental training introduces data gradually in the order of tasks. The model is first trained on Task 1 until the desired accuracy is achieved, after which data from Task 2 is added for further training, followed by the addition of Task 3 data. At each stage, we recorded the accuracy and loss values for both the training and test sets, observing whether the performance of previous tasks was maintained or affected.

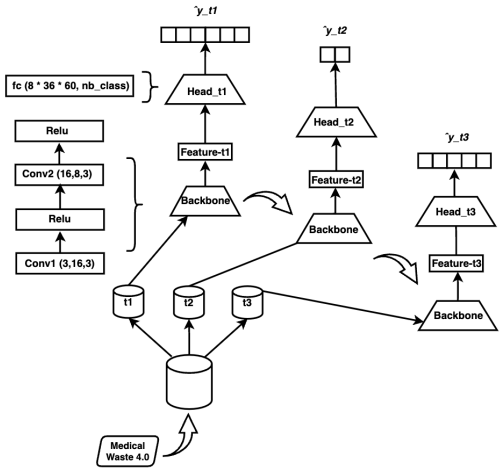


Figure 13: model_exp4

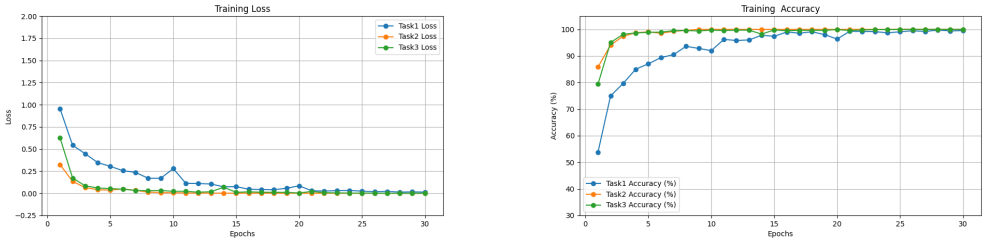
This method simulates real-world scenarios where data arrives incrementally, focusing on the model's adaptability in incremental learning and the interactions between tasks. The model diagram above illustrates the process of gradually introducing incremental data and its dynamic impact on the model's training structure.

Hyper-parameter	**Value**
Validation split	0.2
Number of layers	3
Number of Neurons in Conv1	$16 \times 38 \times 62 = 37{,}696$
Number of Neurons in Conv2	$8 \times 36 \times 60 = 17{,}280$
Number of Neurons in FC_T1	6
Number of Neurons in FC_T2	2
Number of Neurons in FC_T3	5
Activation function	ReLU
Dropout	Not used
Learning rate	0.001
Optimizer	Adam
Batch size	16

Table 5: Hyper-parameter of model continual

The hyper-parameter values used in the experiment are shown in the table above.

4.5.1 Analysis of Model Training Methods



(a) Loss-continual (b) Accuracy-continual

Figure 14: Loss and accuracy of Training

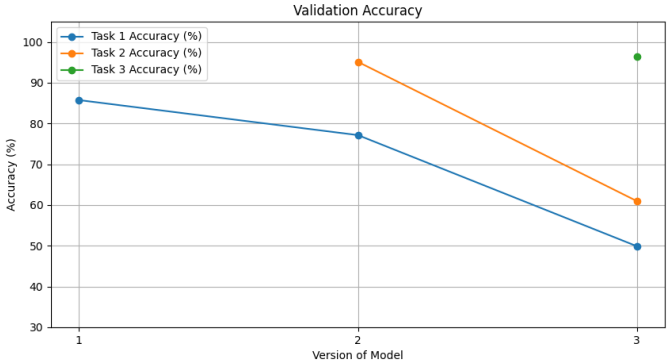


Figure 15: Accuracy of validation

Accuracy and Loss Rate Analysis

Training Loss and Accuracy (Loss-Continual & Accuracy-Continual)

- In the loss rate graph on the left, the loss values for all tasks decrease gradually over 30 epochs, indicating that the model is learning and converging.
- The initial loss for Task 1 is the highest, as it contains more data categories with higher complexity. However, by the end of training, the loss for all three tasks approaches zero.
- The loss for Task 2 and Task 3 decreases more rapidly, suggesting that these tasks are relatively simpler, making it easier for the model to capture their features.
- In the accuracy graph on the right, the training accuracy improves progressively, corresponding to the decrease in loss. The final training accuracy indicates that the model fits the training data well. However, this does not fully guarantee generalization performance on the validation set.

Validation Accuracy (Accuracy of validation)

- Task 1 : The initial model (Version 1) achieved a validation accuracy of approximately 85% on the Task 1 test set. However, after training on Task 2 and Task 3, the validation accuracy on the Task 1 test set gradually decreased. By Version 3, the validation accuracy on Task 1 dropped to approximately 50%, showing a significant reduction and demonstrating a clear case of catastrophic forgetting.
- Task 2 : In Version 2, the validation accuracy on the Task 2 test set was approximately 95%. However, in Version 3, the validation accuracy on the Task 2 test set dropped to

approximately 60%. After training on Task 3, the validation accuracy for Task 2 also decreased significantly, indicating the presence of catastrophic forgetting.

- Task 3 : After completing training on Task 3 in Version 3, the validation accuracy on the Task 3 test set was approximately 96%. Task 3, being the most recently trained task, performed the best, showing that the model has strong adaptability to the most recent task in the continual learning process.

Phenomenon Analysis

Catastrophic Forgetting

- Definition: Catastrophic forgetting refers to the phenomenon in continual learning where the model, while learning new tasks, overwrites or loses the features and knowledge from previous tasks.
- Manifestation: The validation accuracy on the Task 1 and Task 2 test sets significantly decreased as the model trained on newer tasks.

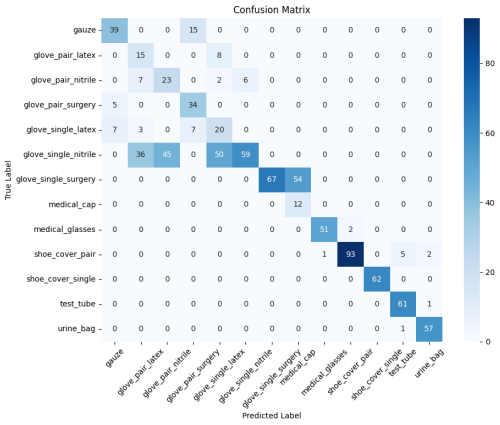


Figure 16: Confusion-matrix_continual

Confusion Matrix Analysis

From the confusion matrix, it can be observed that the model demonstrates overall excellent classification performance. Most of the classification results are concentrated along the diagonal, indicating that the model can accurately identify these categories. The classification results for shoe_cover_pair and urine_bag are particularly outstanding, with almost no apparent misclassifications. This highlights the distinctiveness of these categories' features and the model's strong ability to recognize them.

However, there is still some degree of confusion between certain categories, especially among glove-related categories. For instance, glove_single_latex and glove_single_nitrile are frequently misclassified as each other, possibly due to the high visual similarity in color and material between single gloves, making them difficult to distinguish. Additionally, some instances of glove_pair_surgery are misclassified as glove_pair_nitrile, which may reflect overlapping features between paired gloves during feature extraction.

In contrast, other categories show more stable classification performance, such as medical_glasses and test_tube, with minimal classification errors. This suggests that these categories have more distinct features, enabling the model to effectively capture their key characteristics.

Overall, in model Version 3, the validation performance on the test set for Task 1, which is related to glove categories, was the worst, followed by Task 2, while Task 3 showed the best performance. This aligns with the characteristics of Catastrophic Forgetting Li et al. (2024), where the model forgets the features and knowledge of previously trained tasks as it learns new categories.

5 CONCLUSION

This study conducted four progressive experiments : *Full-Batch Training*, *Independent Task Training*, *Multi-Task Learning*, and *Continual Learning*, to validate the feasibility of a **Never-Ending Learning Model** for medical waste classification. The experimental results demonstrated that the continual learning approach significantly enhanced the model's performance in dynamic environments, effectively alleviating the problem of catastrophic forgetting and ensuring stable classification performance. These findings provide strong support for the realization of the Never-Ending Learning Model. Furthermore, the experiments confirmed that the integration of image analysis techniques and continual learning methods not only allows dynamic adaptation to new categories but also optimizes classification efficiency, addressing the complex and evolving demands of medical waste management.

The results obtained at this stage indicate that the Never-Ending Learning Model is achievable in the context of medical waste classification. This provides a solid foundation for future research to further optimize the model and expand its application scenarios while offering reliable technical support for dynamic adaptation and long-term management in complex medical waste classification settings.

6 FUTURE WORK

By implementing multiple training sessions with result averaging, reevaluating dataset task separation, optimizing training parameters, expanding the dataset scope and mitigating catastrophic forgetting, the stability, generalization, and adaptability of the continual learning model can be significantly enhanced. These improvements offer effective solutions to address bias issues and task adaptability in continual learning.

Multiple Training and Result Averaging

Due to the potential bias in prediction results on the same test set after training the model on the same training set, multiple rounds of training can be performed to improve the stability of the results. After each training session, the prediction results are recorded, and the average of the results is taken as the final evaluation metric for model performance. Additionally, the variance of the results can be calculated to quantify the bias range of predictions, providing a more accurate assessment of model performance. This approach effectively reduces fluctuations caused by the randomness of single training sessions.

Reevaluating Dataset Task Separation

To further enhance the validation performance of the model on the test set, the reasonableness of dataset task separation needs to be reconsidered. By calculating the feature distance between categories, the similarity among categories can be evaluated, and categories with high similarity can be regrouped into the same task to reduce interference between tasks. This method helps optimize the model's feature extraction capabilities, enabling more precise classification performance on the test set and achieving better validation results.

Optimizing Model Training Parameters

During the model training process, optimization tools (such as grid search, random search, or Bayesian optimization) can be used to systematically optimize hyperparameters and training parameters (e.g., nb epoch) to enhance training efficiency and validation performance. Furthermore, different learning rates can be set for each task to adjust the training pace according to task complexity. For instance, smaller learning rates can be used for complex tasks to stabilize gradient updates, while larger learning rates can accelerate convergence for simpler tasks. Such optimization allows the model to learn task-specific features more efficiently, thereby further improving overall performance.

Expanding Dataset Scope

To enhance the inclusivity of the continual learning model, new datasets can be introduced for training. These new datasets can help the model learn more diverse features and improve its adaptability in multi-task and multi-category environments. By training on richer data, the model can better

handle more complex scenarios and simultaneously validate its generalization ability in diverse task environments. This expansion not only contributes to improving model performance but also provides stronger foundational support for continual learning.

mitigating catastrophic forgetting

To mitigate catastrophic forgetting and improve model testing performance, methods such as Elastic Weight Consolidation (EWC) can be employed. EWC identifies and protects important parameters of previous tasks using the Fisher Information Matrix, restricting their excessive updates during the training of new tasks. This approach effectively preserves the features of previous tasks, reduces forgetting, balances new and old tasks, and enhances the model's overall performance and validation results.

REFERENCES

A. Bruno, C. Caudai, G. R. Leone, M. Martinelli, D. Moroni, and F. Crotti. Medical waste sorting: a computer vision approach for assisted primary sorting, 2023a. URL https://arxiv.org/abs/2303.04720.

Antonio Bruno, Massimo Martinelli, and Davide Moroni. Medical-waste-4.0-dataset: v0.1, February 2023b. URL https://doi.org/10.5281/zenodo.7643417.

Arslan Chaudhry, Albert Gordo, Puneet K. Dokania, Philip H. S. Torr, and David Lopez-Paz. Using hindsight to anchor past knowledge in continual learning. *CoRR*, abs/2002.08165, 2020. URL https://arxiv.org/abs/2002.08165.

Yayong Li, Peyman Moghadam, Can Peng, Nan Ye, and Piotr Koniusz. Inductive graph few-shot class incremental learning, 2024. URL https://arxiv.org/abs/2411.06634.

World Health Organization. Healthcare waste. *WHO News Report*, 2024. URL https://www.who.int/news-room/fact-sheets/detail/health-care-waste.

Haiying Zhou, Xiangyu Yu, Ahmad Alhaskawi, Yanzhao Dong, Zewei Wang, Qianjun Jin, Xianliang Hu, Zongyu Liu, Vishnu Goutham Kota, Mohamed Abdulla, Sohaib Ezzi, Binjie Qi, Juan Li, Bixian Wang, Jianyong Fang, and Hui lu. A deep learning approach for medical waste classification. *Scientific Reports*, 12, 02 2022. doi: 10.1038/s41598-022-06146-2.