M³-CIL: A Multidimensional Evaluation Protocol for Class-Incremental Learning

Liedson D. O. Leite[1], **Gabriel F. Sarmento**[1], **Thyago T. M. de Oliveira**[1], **Rodolfo C. Cavalcante**[1], **Leandro L. Minku**[2], **Gustavo H. F. de M. Oliveira**[1]

[1]Sistemas de Informação – Universidade Federal de Alagoas (UFAL)
Unidade Educacional de Penedo, Penedo, AL – Brasil

[2]School of Computer Science – University of Birmingham
Edgbaston, Birmingham, B15 2TT – Reino Unido

{liedson.leite, gabriel.sarmento, rodolfo.cavalcante}@arapiraca.ufal.br

{gustavo.oliveira, thyago.oliveira}@penedo.ufal.br, l.l.minku@bham.ac.uk

Abstract. *In Class-Incremental Learning, catastrophic forgetting is a central challenge. Currently, methods which mitigate this problem are mostly evaluated by final accuracy, neglecting real operational constraints. This paper proposes the* ***M³-CIL (Multidimensional Metric Matrix)****, a protocol that investigates the practical viability of approaches across three of analysis: knowledge retention, hyperparameter sensitivity, and operational efficiency. After analyzing seven classic methods in scenarios of increasing complexity, we conclude that Replay strategies (notably ER-ACE) offer the best balance for real applications, while Regularization techniques prove ineffective in complex contexts.*

Resumo. *No aprendizado incremental de classes, o esquecimento catastrófico é um desafio central. Atualmente, métodos que mitigam este problema são avaliados sobretudo pela acurácia final, ignorando restrições operacionais. Este trabalho propõe o* ***M³-CIL (Multidimensional Metric Matrix)****, protocolo que investiga a viabilidade prática de tais abordagens em três eixos de análise: retenção de conhecimento, sensibilidade hiperparamétrica e eficiência operacional. Após analisar sete métodos clássicos em cenários de complexidade crescente, conclui-se que estratégias de Replay (especialmente o ER-ACE) oferecem o melhor equilíbrio entre estabilidade-plasticidade, enquanto técnicas de Regularização mostram-se ineficazes em contextos complexos.*

1. Introdução

Sistemas de aprendizado de máquina aplicados a problemas do mundo real, especialmente em visão computacional, operam em ambientes dinâmicos, nos quais novos dados visuais surgem continuamente [Zhou et al. 2024]. Nesses cenários, os modelos precisam ser atualizados continuamente à medida que novos dados são disponibilizados.

Nesse contexto, as redes neurais profundas, amplamente utilizadas nessas tarefas, enfrentam o problema do *esquecimento catastrófico* (*catastrophic forgetting*). Este problema está relacionado à aprendizagem de novas classes de forma dinâmica pelo modelo, onde há uma tendência de sobrescrever representações previamente adquiridas, resultando em queda de desempenho nas classes antigas [Goodfellow et al. 2015].

Na literatura, esse cenário é conhecido por Aprendizado Incremental por Classes (do inglês: *Class-Incremental Learning – CIL*) [van de Ven and Tolias 2019], no qual o modelo recebe novas classes de forma sequencial e precisa incorporá-las ao seu conhecimento ao longo do tempo. Durante a avaliação, o modelo deve classificar cada nova amostra considerando todas as classes já aprendidas, sem saber a qual etapa de treinamento a entrada pertence [van de Ven and Tolias 2019]. Esse cenário torna o problema do esquecimento catastrófico ainda mais desafiador, pois exige que o modelo mantenha desempenho consistente em um conjunto de classes que cresce continuamente.

Embora diversas abordagens tenham sido propostas para esse fim, especialmente aquelas baseadas em *replay* [Rebuffi et al. 2017] e *regularização* [Kirkpatrick et al. 2017], a maioria dos estudos avalia suas propostas predominantemente por meio da acurácia final [Chavan et al. 2023, Zhou et al. 2024]. Nesse sentido, observa-se uma lacuna metodológica relevante, uma vez que essa prática negligencia dimensões determinantes para a aplicabilidade prática no mundo real, como a capacidade de retenção de conhecimento, a robustez a variações de hiperparâmetros e o custo computacional. Como consequência, a avaliação da viabilidade dessas abordagens em cenários práticos torna-se limitada.

Para preencher essa lacuna, este trabalho propõe o **M³-CIL (Multidimensional Metric-based Audit for CIL)**, um protocolo de avaliação projetado para validar as abordagens de *CIL*. O **M³-CIL** integra três eixos de análise: (i) eficácia na retenção de conhecimento, por meio do *trade-off* entre acurácia e esquecimento; (ii) robustez, avaliada pela sensibilidade à variação de hiperparâmetros; (iii) eficiência operacional, considerando a relação entre desempenho e custo computacional bem como a escalabilidade, ao avaliar os métodos em domínios de complexidade crescente. Nesse sentido, a principal contribuição deste estudo reside na formalização da metodologia M³-CIL e na aplicação prática desse protocolo em sete métodos representativos da literatura. Além disso, o trabalho propõe uma representação espacial multidimensional das estratégias de Aprendizado Contínuo, integrando forgetting, acurácia, custo computacional e sensibilidade a hiperparâmetros em uma única visualização analítica, permitindo interpretar simultaneamente diferentes dimensões de viabilidade prática em cenários incrementais.

Para alcançar essa contribuição, o restante do artigo está organizado da seguinte forma: a Seção 2 apresenta os fundamentos teóricos; a Seção 3, os trabalhos relacionados; as Seções 4 e 5 descrevem o protocolo M³-CIL e a metodologia experimental; a Seção 6 discute os resultados; e a Seção 7 apresenta as conclusões.

2. Fundamentação Teórica

Esta seção visa apresentar de forma resumida os principais conceitos envolvidos no trabalho (Subseção 2.1) e as principais abordagens de aprendizado incremental por classes (Subseção 2.2).

2.1. Aprendizado Contínuo e Esquecimento Catastrófico

O Aprendizado Contínuo visa o desenvolvimento de sistemas capazes de aprender uma sequência de tarefas de forma incremental. Formalmente, conforme estabelecido por [Lopez-Paz and Ranzato 2017], uma sequência de tarefas é definida por $\mathcal{T} = \{T_1, T_2, \ldots, T_k\}$, onde cada tarefa T_t representa um problema específico e possui um conjunto de dados $\mathcal{D}_t = \{(x_i^t, y_i^t)\}_{i=1}^{n_t}$, em que x é a entrada e y o rótulo.

Nesse contexto, busca-se ajustar os parâmetros θ de uma rede neural $f(\cdot;\theta)$, que produz previsões a partir de x, visando minimizar o erro $\ell(f(x;\theta), y)$, que quantifica a diferença entre a previsão e o valor real. De forma geral, o aprendizado consiste em minimizar o risco empírico total ao longo de todas as tarefas, dado por $\mathcal{L}_{total}(\theta) = \sum_{t=1}^{k} \mathbb{E}_{(x,y)\sim D_t}[\ell(f(x;\theta), y)]$, em que $\mathbb{E}_{(x,y)\sim D_t}$ representa o erro médio na tarefa T_t.

Entretanto, como os dados das tarefas anteriores $D_{1:t-1}$ não estão disponíveis durante o treinamento de T_t, a atualização dos parâmetros tende a favorecer a tarefa atual em detrimento das anteriores, levando ao esquecimento catastrófico ([Goodfellow et al. 2015]) e evidenciando o dilema estabilidade-plasticidade ([Mermillod et al. 2013]).

2.2. Abordagens de Aprendizado Incremental por Classes

Na literatura de Aprendizado Contínuo, os métodos são tradicionalmente organizados em três grandes categorias: (i) baseados em regularização, (ii) baseados em replay e (iii) baseados em expansão de arquitetura [van de Ven and Tolias 2019]. Embora abordagens recentes explorem novos paradigmas, como métodos baseados em prompts ou modelos generativos, este trabalho foca exclusivamente nas famílias de regularização e replay, por permitirem uma comparação direta sob arquitetura fixa.

As abordagens de regularização assumem que o conhecimento pode ser preservado ao restringir atualizações nos parâmetros do modelo, sem a necessidade de armazenar dados passados. Para isso, modificam a função de perda da tarefa atual $\mathcal{L}_{new}$, adicionando um termo de penalização: $\mathcal{L}(\theta) = \mathcal{L}_{new}(\theta) + \lambda \cdot \mathcal{R}(\theta, \theta_{old})$. Nessa formulação, θ são os parâmetros atuais, θ_{old} representam o estado anterior do modelo, $\mathcal{R}$ mede a mudança entre esses estados, e λ controla a intensidade da restrição ([Kirkpatrick et al. 2017, Li and Hoiem 2017]).

Em contraste, as abordagens de replay preservam informações do passado por meio de um buffer de memória $\mathcal{M}$, que armazena um subconjunto representativo de amostras anteriores. Durante o treinamento, a otimização passa a considerar conjuntamente dados novos e amostras recuperadas da memória, aproximando a minimização da perda a um cenário multi-tarefa. A função objetivo pode ser escrita como: $\mathcal{L}_{replay}(\theta) = \mathbb{E}_{(x,y)\sim D_{new}}[\ell(f(x;\theta), y)] + \mathbb{E}_{(x',y')\sim\mathcal{M}}[\ell(f(x';\theta), y')]$, onde o segundo termo atua como um mecanismo de regularização, forçando o modelo a manter bom desempenho nas amostras armazenadas. Na prática, o equilíbrio entre os termos pode ser controlado implicitamente pela proporção de amostras no mini-batch usado no treinamento, o que influencia diretamente a estabilidade do modelo ao longo do tempo ([Rebuffi et al. 2017]).

3. Trabalhos Relacionados

A literatura em Aprendizado Contínuo, por meio das famílias de regularização e replay, apresentam diferentes estratégias para mitigar o esquecimento catastrófico. Entretanto, estudos recentes apontam que avaliações fundamentadas apenas na acurácia final negligenciam fatores críticos para cenários reais, como custo computacional, uso de memória e robustez [Chavan et al. 2023, Zhou et al. 2024].

No eixo da **Regularização**, destacam-se os métodos *Elastic Weight Consolidation* (EWC) [Kirkpatrick et al. 2017], *Synaptic Intelligence* (SI) e *Learning without Forgetting*

(LwF) [Li and Hoiem 2017]. Embora eficientes em memória, essas abordagens tendem a apresentar limitações de retenção em cenários mais complexos.

Por outro lado, métodos baseados em **Replay** como *iCaRL* [Rebuffi et al. 2017], *ER*, *ER-ACE*, *DER*, A-GEM e GDumb [Prabhu et al. 2020] por explorarem diferentes estratégias de armazenamento e reutilização de exemplos, geralmente têm alcançado menor esquecimento ao custo de maior demanda computacional e de memória [Chavan et al. 2023, Kumar et al. 2025].

Apesar desses avanços, a literatura ainda carece de protocolos que integrem simultaneamente a retenção de conhecimento, sensibilidade a hiperparâmetros e eficiência operacional. Nesse contexto, o **M³-CIL** se destaca por propor uma avaliação multidimensional da viabilidade prática dessas estratégias de CIL em ambientes dinâmicos e com restrições operacionais.

4. O Framework de Avaliação M³-CIL

Esta seção formaliza o **M³-CIL**, um protocolo de viabilidade prática de abordagens para o cenário de aprendizagem interativa de classes. O framework organiza a investigação das abordagens sob a ótica de três eixos complementares: (i) eficácia na retenção de conhecimento (Subseção 4.1), (ii) robustez a hiperparâmetros (Subseção 4.2) e (iii) eficiência operacional (Subseção 4.3), integrados sob uma perspectiva de análise multidimensional (Subseção 4.4).

4.1. Eixo I: Retenção de Conhecimento (Estabilidade-Plasticidade)

O primeiro eixo avalia a capacidade do modelo de equilibrar a incorporação de novas informações (*plasticidade*) com a preservação do conhecimento previamente adquirido (*estabilidade*). Conforme a metodologia proposta por [Lopez-Paz and Ranzato 2017], essas métricas são derivadas de uma matriz de desempenho $R \in \mathbb{R}^{k \times k}$, onde cada elemento $R_{i,j}$ indica a acurácia obtida na tarefa j após o modelo ter concluído o treinamento da tarefa i. Para caracterizar esse equilíbrio, utilizam-se as métricas de Acurácia Média e *Forgetting Measure*:

- **Acurácia Média (A_k)**: mede o desempenho global ao longo das tarefas, sendo definida por $A_k = \frac{1}{k} \sum_{j=1}^{k} R_{k,j}$, onde k é o número de tarefas e $R_{k,j}$ é a acurácia na tarefa j após o treinamento da tarefa k. Valores mais altos indicam maior plasticidade do modelo.
- **Forgetting Measure (FM)**: quantifica a perda de desempenho em tarefas anteriores, dada por $FM = \frac{1}{k-1} \sum_{j=1}^{k-1} \left(\max_{l \in \{1,\ldots,k-1\}} R_{l,j} - R_{k,j} \right)$, onde $\max_l R_{l,j}$ é o melhor desempenho já alcançado na tarefa j e $R_{k,j}$ o desempenho atual. Valores menores indicam maior estabilidade, ou seja, menor esquecimento.

4.2. Eixo II: Robustez a Hiperparâmetros

O segundo eixo investiga a sensibilidade das estratégias às variações de seus principais hiperparâmetros. Em cenários práticos, métodos altamente sensíveis tendem a demandar ajustes finos e custosos, reduzindo sua aplicabilidade. Para analisar esse aspecto, o comportamento dos métodos é avaliado por meio de curvas de desempenho em função dessas variações. Nesse contexto, métodos robustos tendem a apresentar desempenho estável e previsível, enquanto métodos sensíveis exibem maior variabilidade ou forte dependência de configurações específicas.

4.3. Eixo III: Eficiência Operacional

O terceiro eixo considera a viabilidade computacional com base em dois pilares: o **custo computacional**, mensurado pelo tempo de execução absoluto, e a **escalabilidade**, avaliada pela resiliência do modelo frente ao aumento da complexidade visual dos datasets. A análise conjunta desses fatores permite identificar métodos que oferecem bom desempenho sem comprometer significativamente os recursos computacionais, aspecto fundamental em aplicações reais.

4.4. M³-CIL: Análise Multidimensional das Estratégias

Para consolidar os três eixos de análise, o protocolo propõe uma representação espacial multidimensional das estratégias, permitindo interpretar simultaneamente retenção de conhecimento, plasticidade, custo computacional e estabilidade paramétrica em um único espaço analítico, apresentada posteriormente na Figura 5. Essa representação funciona como uma visualização complementar da análise M³-CIL, favorecendo a identificação de abordagens mais adequadas para cenários incrementais reais.

Além disso, o protocolo M³-CIL organiza os resultados em uma matriz de análise multidimensional capaz de integrar métricas quantitativas e critérios qualitativos de viabilidade prática através de uma tabela para os três eixos. Para cada eixo, os símbolos ($\checkmark$ e $\times$) são atribuídos com base no posicionamento relativo das abordagens em relação à média global das métricas avaliadas. Dessa forma, o M³-CIL permite comparar estratégias de Aprendizado Contínuo de maneira padronizada e replicável, considerando não apenas a acurácia final, mas também sua prontidão para ambientes dinâmicos e computacionalmente restritos.

5. Metodologia Experimental

Esta seção detalha a fundamentação do *framework* **M³-CIL** e o esboço experimental utilizado para sua validação. O fluxo de execução do protocolo proposto está sintetizado na Figura 1, que apresenta as etapas sequenciais desde o processamento dos dados até a análise final. Com base nessa estrutura, a metodologia é aprofundada nas subseções seguintes: a Seleção de *datasets* e definição dos cenários de aprendizagem (5.1); a Especificação do ambiente e arquiteturas (5.2); a Definição das abordagens e parametrização sistemática (5.3); e, por fim, os critérios para a Consolidação dos resultados via matriz de viabilidade (4.4).



Figura 1. Fluxo geral do protocolo experimental no cenário de aprendizado incremental (CIL).

5.1. Seleção dos Datasets e Cenário de Aprendizado Interativo

Os datasets adotados foram escolhidos e organizados em ordem crescente de complexidade: MNIST $\rightarrow$ Fashion-MNIST (FMNIST) $\rightarrow$ CIFAR-10. O MNIST é composto por imagens de dígitos manuscritos; o Fashion-MNIST (FMNIST) por imagens de itens de

vestuário; e o Split-CIFAR10, uma versão particionada do CIFAR-10, contém imagens coloridas de objetos do mundo real. Essa escolha permite avaliar o comportamento dos métodos em diferentes níveis de complexidade visual. O MNIST e o FMNIST consistem em imagens de 28×28 pixels em escala de cinza (60.000 amostras de treino e 10.000 de teste) e são utilizados para analisar o desempenho inicial em cenários mais simples. Em contraste, o Split-CIFAR-10 apresenta imagens coloridas (RGB) de 32×32 pixels, com maior variabilidade e complexidade textural, impondo um desafio mais próximo de aplicações reais.

Definidos os datasets, adota-se o cenário de *Class-Incremental Learning* (CIL), no qual o modelo é exposto a uma sequência de tarefas ao longo do tempo. Neste trabalho, o fluxo é organizado em cinco tarefas consecutivas, cada uma introduzindo duas novas classes que devem ser incorporadas ao conhecimento já existente. Como consequência, deve realizar a predição considerando todas as classes já aprendidas, o que torna o problema significativamente mais desafiador. Como a ordem de apresentação das classes influencia diretamente os resultados nesse cenário, cada experimento é replicado com cinco sementes aleatórias fixas (123, 456, 789, 101 e 202), que definem diferentes sequências de classes. Os resultados são reportados em termos de média e desvio padrão das sementes.

5.2. Definição dos Modelos Base, Arquitetura e Parâmetros de Treinamento

As arquiteturas das redes neurais experimentadas foram fixadas para enfatizar apenas o efeito das estratégias de aprendizado contínuo, evitando que limitações arquiteturais interfiram na análise. Para garantir a capacidade de representação adequada foi utilizado um Perceptron Multicamadas (SimpleMLP) com duas camadas ocultas de 400 unidades e ativação ReLU para MNIST e FMNIST e uma Rede Neural Convolucional (SimpleCNN) composta por quatro camadas convolucionais e uma camada densa final para o Split-CIFAR10. Todos os modelos de redes neurais foram otimizados via SGD com momento de 0.9, taxa de aprendizado fixa em 0,01 e *batch size* entre 32 e 64, dependendo do dataset, alinhados às configurações de referência da biblioteca Avalanche. Os experimentos foram executados em *PyTorch* utilizando GPU RTX 4070 e 64GB RAM

5.3. Definição das Abordagens Avaliadas e Hiperparametrização

Para este estudo, foram selecionadas as duas famílias de abordagens mais relevantes no cenário de Aprendizado Contínuo: regularização e replay. A avaliação foi conduzida considerando métodos representativos de cada família e com foco na análise de sensibilidade a hiperparâmetros (Eixo II).

No grupo de Regularização, foram avaliados: o EWC, com coeficiente de penalidade $\lambda \in \{1, 100, 1000, 10000\}$; o LwF, variando o peso de destilação $\alpha \in \{0.1, 1, 10, 100\}$ com temperatura fixa de 2.0; e o Synaptic Intelligence (SI), com $\lambda \in \{0.1, 1, 10, 100\}$. Esses métodos permitem analisar o impacto da restrição sobre parâmetros importantes na retenção de conhecimento, sem o uso de memória explícita.

No grupo de Replay e Gerenciamento de Memória, foram avaliados o ER-ACE, o DER e o GDumb, todos com variação no tamanho do buffer de memória ($\mathcal{M}$) entre 100, 200 e 500 amostras, permitindo observar a influência direta da memória na escalabilidade e desempenho. O DER foi configurado com $\alpha = 0.1$ e $\beta = 0.0$, com foco na preservação de *logits*. Adicionalmente, avaliou-se o A-GEM, variando o número de padrões por experiência entre 50 e 200, com tamanho de amostra fixo de 64.

A análise de sensibilidade foi conduzida variando apenas o hiperparâmetro responsável por controlar o dilema estabilidade-plasticidade em cada método (e.g., a força de regularização λ ou o tamanho do buffer $\mathcal{M}$), mantendo os demais fixos conforme as implementações de referência da biblioteca Avalanche. Essa estratégia evita efeitos de confusão oriundos de uma otimização combinatória multivariável (*grid search*), garantindo que a comparabilidade reflita a robustez inerente à arquitetura de cada abordagem, e não um ajuste exaustivo.

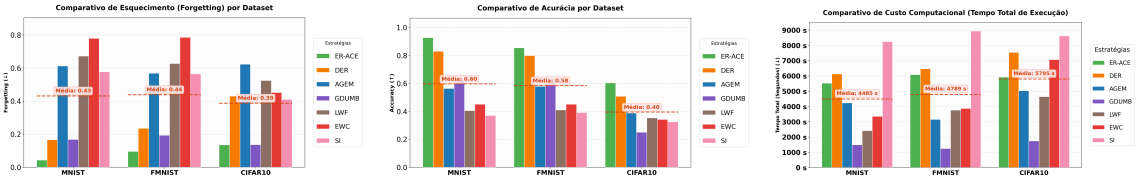
6. Resultados e Discussões

Nesta seção, os resultados experimentais são analisados sob a ótica do **M³-CIL**. Inicialmente, na Seção 6.1 avalia-se o equilíbrio entre retenção de conhecimento e adaptação (Eixo I). Em seguida, na Seção 6.2 investiga-se a sensibilidade a hiperparâmetros das abordagens (Eixo II). Posteriormente, na Seção 6.3 analisam-se a eficiência operacional e a escalabilidade sob aumento de complexidade visual (Eixo III). Por fim, na Seção 6.4 os resultados são integrados na Figura e Matriz de Resultado do protocolo.

6.1. Eixo I - Eficácia de Retenção e Equilíbrio Entre Métricas

Esta análise investiga o equilíbrio entre estabilidade e plasticidade por meio da relação entre *forgetting* e acurácia média. A Figura 2 apresenta esse comportamento sob duas perspectivas complementares: forgetting (Figura 2a) e acurácia (Figura 2b). Em ambos os gráficos, as barras representam a média global das abordagens ao longo dos três datasets avaliados (MNIST, FMNIST e CIFAR-10), permitindo observar simultaneamente o comportamento geral dos métodos e sua degradação conforme aumenta a complexidade visual do cenário incremental.

Os resultados mostram um padrão consistente favorável às estratégias baseadas em Replay, especialmente ER-ACE e DER, que apresentam simultaneamente maiores níveis de acurácia e menores taxas de forgetting. O ER-ACE destaca-se por preservar melhor o desempenho em tarefas anteriores mesmo com a introdução contínua de novas classes. Além disso, observa-se que, embora todas as abordagens sofram degradação com o aumento da complexidade visual, os métodos de Replay mantêm comportamento mais estável ao longo do gradiente MNIST $\rightarrow$ CIFAR-10. Por outro lado, os métodos de Regularização (EWC, SI e LwF) apresentam degradação mais acentuada, concentrando-se em regiões associadas a maior forgetting e menor acurácia final, especialmente nos cenários mais complexos.

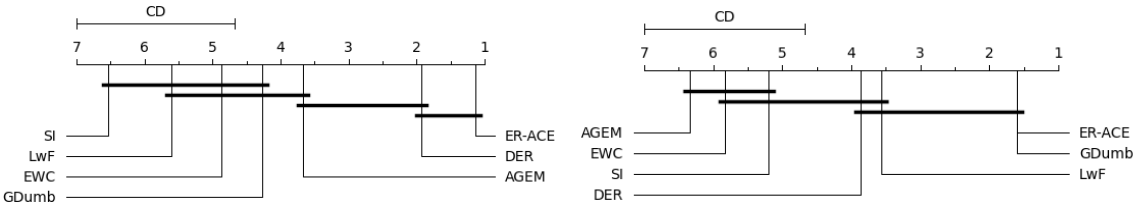


(a) Forgetting **(b) Acurácia** **(c) Tempo (s)**

Figura 2. Métricas para as estratégias de Aprendizado Contínuo.

Para complementar a análise, a Figura 3 apresenta o ranking estatístico global utilizando o teste *AutoRank*, baseado no teste de Friedman com pós-teste de Nemenyi. Nessas figuras, os métodos são organizados da esquerda para a direita, do pior para o

melhor desempenho médio. As conexões horizontais indicam ausência de diferença estatisticamente significativa entre os métodos conectados, considerando a distância crítica (*Critical Distance*) do teste de Nemenyi. Assim, métodos não conectados apresentam diferença estatisticamente relevante em desempenho. Dado isso, Os resultados reforçam experimentalmente a superioridade do ER-ACE em acurácia e retenção, mantendo equivalência estatística apenas com o DER.



(a) Acurácia **(b) Forgetting**

Figura 3. Ranking de Friedman com pós-teste de Nemenyi.

6.2. Eixo II - Análise de sensibilidade a hiperparâmetros e Eficiência Operacional Prática

O segundo eixo investiga a sensibilidade das abordagens à variação de seus principais hiperparâmetros. A Figura 4 apresenta as curvas de sensibilidade ao longo dos datasets: o eixo horizontal representa a variação normalizada do principal hiperparâmetro de cada abordagem (Subseção 5.3), e o eixo vertical, o desempenho médio correspondente. As linhas indicam a tendência média, e as regiões sombreadas, a variabilidade entre execuções. Sombras estreitas indicam comportamento consistente entre configurações e sementes; sombras amplas evidenciam maior instabilidade e dependência paramétrica.

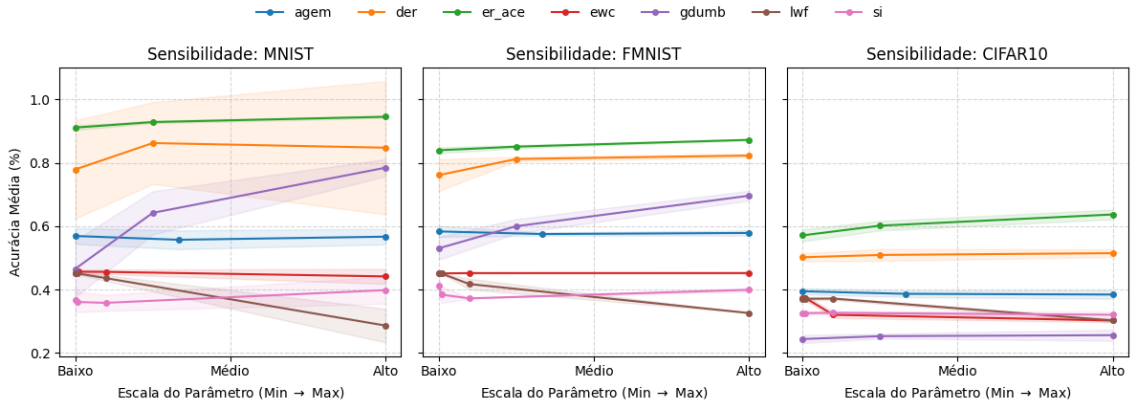


Figura 4. Variação dos hiperparâmetros das abordagens ao longo dos datasets.

Os resultados mostram que ER-ACE e DER apresentam comportamento mais estável mesmo diante de mudanças significativas nos hiperparâmetros, mantendo curvas relativamente suaves e baixa dispersão ao longo dos datasets. Esse comportamento indica maior previsibilidade operacional e menor sensibilidade às configurações experimentais.

Em contraste, métodos de Regularização exibem desempenho estagnado ou forte dependência de regiões específicas de configuração, frequentemente acompanhados de maior variabilidade. Esse comportamento sugere que o aumento da intensidade da

regularização não é suficiente para resolver o esquecimento catastrófico em cenários mais complexos, além de tornar o desempenho mais dependente de ajustes específicos.

6.3. Eixo III - Eficiência Operacional: Escalabilidade e Custo Computacional

O terceiro eixo investiga a eficiência operacional sob duas perspectivas: custo computacional e escalabilidade frente ao aumento da complexidade visual. A Figura 2 sintetiza essa análise com a média global de cada abordagem por métrica, relacionando desempenho, retenção e custo operacional. A Figura 2c apresenta o tempo total de execução acumulado, com barras que representam a média global do custo computacional nos cenários MNIST, FMNIST e CIFAR-10, evidenciando como o processamento escala com a complexidade visual.

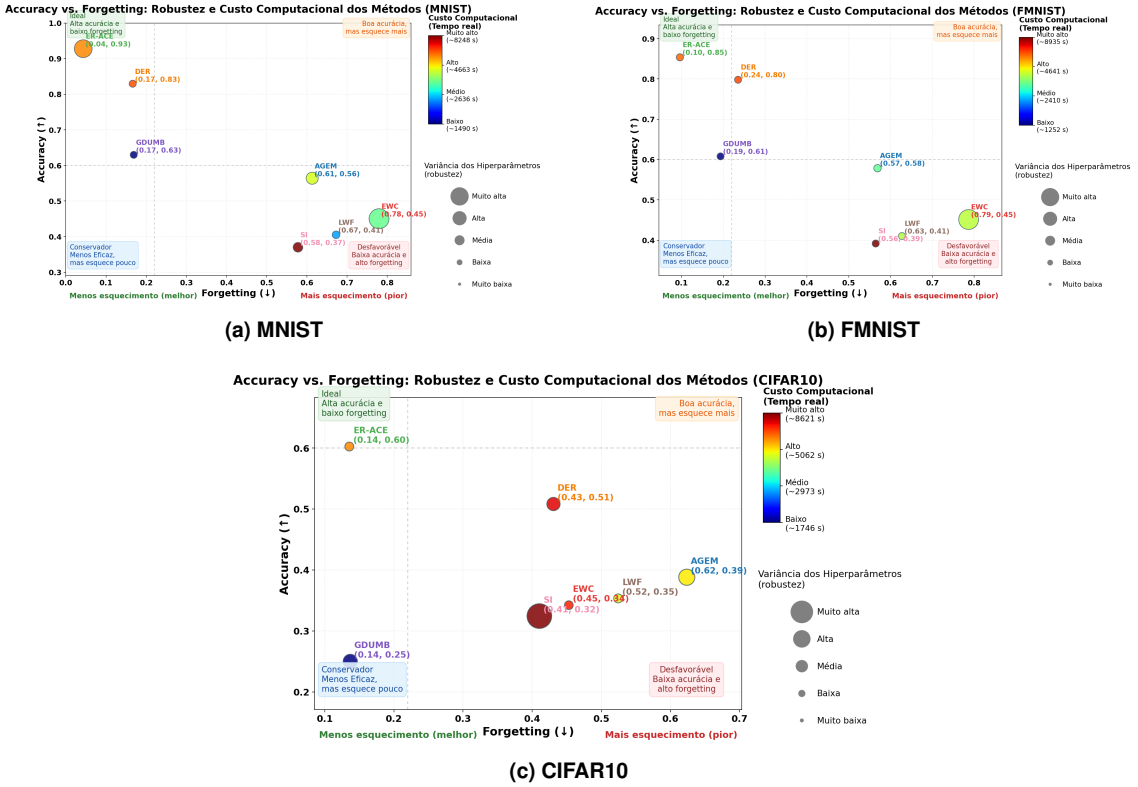
Os resultados mostram que métodos baseados em Replay tendem a demandar maior custo operacional devido ao gerenciamento contínuo do buffer de memória. Ainda assim, o ER-ACE mantém um equilíbrio favorável entre desempenho e custo computacional, combinando elevada acurácia e baixo forgetting sem apresentar crescimento desproporcional do custo computacional. Em contrapartida, o SI apresenta uma das maiores demandas de processamento, mas sem ganhos equivalentes em retenção ou acurácia. Esse comportamento sugere baixa eficiência operacional, especialmente em cenários com restrições computacionais. Já abordagens como GDumb e A-GEM apresentam menor custo absoluto, porém acompanhado de desempenho significativamente inferior.

A análise conjunta da Figura 2 também evidencia o comportamento das abordagens ao longo do gradiente de complexidade visual (MNIST $\rightarrow$ FMNIST $\rightarrow$ CIFAR-10). Observa-se que estratégias de Regularização apresentam degradação progressiva em acurácia e aumento do forgetting, enquanto métodos de Replay mantêm comportamento relativamente mais estável. Embora abordagens de Replay também sofram degradação associada à limitação do buffer frente ao aumento da diversidade visual, o ER-ACE preserva melhor equilíbrio entre retenção, plasticidade e custo computacional, demonstrando maior escalabilidade prática.

6.4. M³-CIL: Análise Multidimensional das Estratégias

A Figura 5 representa a principal materialização visual da análise multidimensional proposta pelo protocolo M³-CIL, integrando em uma única representação os diferentes eixos de viabilidade prática discutidos ao longo deste trabalho: forgetting, acurácia, custo computacional e estabilidade paramétrica. Em cada gráfico, o eixo horizontal representa o *forgetting*, associado à capacidade de retenção de conhecimento, enquanto o eixo vertical representa a acurácia média, relacionada à plasticidade do modelo. A escala de cores indica o custo computacional das abordagens, enquanto o tamanho das bolhas representa a variabilidade dos resultados entre diferentes execuções e configurações. Dessa forma, métodos posicionados próximos da região superior esquerda, acompanhados de bolhas menores e colorações associadas a menor custo, representam estratégias mais adequadas para aplicações práticas.

As Figuras 5a, 5b e 5c permitem observar como essa relação multidimensional evolui conforme aumenta a complexidade visual do problema. No cenário MNIST, praticamente todas as abordagens apresentam desempenho competitivo devido à simplicidade do domínio visual. Entretanto, mesmo nesse cenário inicial, estratégias de Replay já demonstram vantagem em retenção de conhecimento.



(a) MNIST **(b) FMNIST**

(c) CIFAR10

Figura 5. Espacialização multidimensional do protocolo M³-CIL integrando retenção de conhecimento, acurácia, custo computacional e robustez hiperparamétrica.

No FashionMNIST, a separação entre as famílias de métodos torna-se mais evidente. Estratégias de Replay permanecem concentradas em regiões associadas a maior acurácia e menor forgetting, enquanto métodos de Regularização passam a apresentar maior dispersão e degradação de desempenho. Essa diferença torna-se ainda mais explícita no CIFAR-10, cenário mais complexo e mais próximo de aplicações reais. Nesse caso, métodos de Regularização deslocam-se para regiões associadas a maior forgetting e menor acurácia, evidenciando limitação estrutural na preservação do conhecimento. Em contraste, o ER-ACE mantém-se consistentemente próximo da região de maior viabilidade prática, preservando elevado desempenho mesmo sob aumento da complexidade visual.

Além disso, o tamanho reduzido das bolhas associadas ao ER-ACE indica menor variabilidade entre execuções, sugerindo maior sensibilidade a hiperparâmetros e previsibilidade operacional. Esse comportamento reforça os achados discutidos nos Eixos II e III, evidenciando que o método não apenas apresenta elevado desempenho médio, mas também maior estabilidade frente a diferentes condições experimentais.

A Tabela 1 consolida os resultados obtidos nos três eixos do protocolo M³-CIL, sintetizando a análise multidimensional proposta neste trabalho. Diferentemente de avaliações tradicionais fundamentadas apenas em acurácia final, a matriz permite interpretar simultaneamente retenção de conhecimento, sensibilidade a hiperparâmetros e eficiência operacional, funcionando como um *checklist* de viabilidade prática para estratégias de Aprendizado Contínuo.

Tabela 1. Matriz de Auditoria Qualitativa das Estratégias (Checklist M³-CIL).

Família	Método	Eixo I	Eixo II	Eixo III	Veredito
Regularização	EWC	×	✓	✓	Limitado
	LwF	×	✓	✓	Limitado
	SI	×	×	×	Reprovado
Replay	**ER-ACE**	✓	✓	✓	**Aprovado**
	DER	✓	✓	×	Ineficiente
	GDumb	×	✓	✓	Inconsistente
	A-GEM	×	×	✓	Reprovado

Sob a ótica da Tabela 1, observa-se que estratégias de Replay concentram o maior número de conformidades simultâneas entre os três eixos avaliados, indicando maior equilíbrio entre retenção, robustez e eficiência operacional. Nesse contexto, abordagens classificadas como *aprovadas* representam métodos capazes de satisfazer consistentemente os critérios mínimos definidos pelo protocolo M³-CIL. Em contrapartida, classificações como *limitado* ou *inconsistente* refletem estratégias comprometidas por degradação progressiva de desempenho, elevada sensibilidade paramétrica ou alta variabilidade entre execuções. Já métodos considerados *ineficientes* apresentam custo computacional desproporcional aos ganhos obtidos, enquanto abordagens *reprovadas* falham simultaneamente em múltiplos critérios fundamentais do protocolo.

Em particular, o ER-ACE destaca-se por satisfazer consistentemente os requisitos definidos pelo M³-CIL, demonstrando maior prontidão para cenários incrementais práticos. Em contraste, métodos de Regularização apresentam apenas conformidade parcial. Embora algumas dessas abordagens exibam baixo custo computacional, a degradação progressiva em retenção de conhecimento e a maior sensibilidade paramétrica comprometem sua viabilidade em ambientes mais complexos.

Dessa forma, a análise consolidada da Tabela 1 valida a principal hipótese do M³-CIL: avaliações fundamentadas exclusivamente em acurácia final são insuficientes para caracterizar a viabilidade prática de estratégias de Aprendizado Contínuo. Consequentemente, os resultados demonstram que métodos de Replay, especialmente o ER-ACE, oferecem o melhor compromisso entre retenção de conhecimento, sensibilidade a hiperparâmetros e eficiência operacional.

7. Conclusões e Trabalhos Futuros

Este trabalho formalizou e aplicou o protocolo **M³-CIL (Multidimensional Metric-based Audit for CIL)** para investigar a viabilidade de estratégias de Aprendizado Contínuo no cenário de **Class-Incremental Learning (CIL)**. A principal contribuição é a proposição do protocolo, estruturado em três eixos operacionais e representado espacialmente (Figura 5), permitindo uma análise multidimensional dos resultados. A visualização evidencia que a acurácia final, isoladamente, é insuficiente para avaliar a viabilidade prática em ambientes CIL.

Os resultados demonstraram que estratégias de **Replay** apresentam o melhor equilíbrio operacional, com destaque para o **ER-ACE**, que obteve maior estabilidade, robustez e eficiência computacional nos três eixos. Em contraste, métodos de regularização apresentaram limitações na retenção de conhecimento em cenários mais complexos.

Apesar dos resultados obtidos, o estudo concentrou-se em estratégias clássicas

de Replay e Regularização sob arquiteturas convencionais, não abrangendo paradigmas recentes baseados em transformers, prompts ou modelos fundacionais. Como trabalhos futuros, propõe-se estender o protocolo a cenários de *Domain-Incremental Learning*, conjuntos de dados maiores e mecanismos de compressão de memória voltados à eficiência operacional.

Referências

Chavan, V., Koch, P., Schlüter, M., and Briese, C. (2023). Towards realistic evaluation of industrial continual learning scenarios with an emphasis on energy consumption and computational footprint. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11472–11484.

Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., and Bengio, Y. (2015). An empirical investigation of catastrophic forgetting in gradient-based neural networks.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Kumar, S., Marklund, H., Rao, A., Zhu, Y., Jeon, H. J., Yueyang, L., and Van Roy, B. (2025). Continual learning as computationally constrained reinforcement learning. *Foundations and Trends® in Machine Learning*, 18(4):913–1053.

Li, Z. and Hoiem, D. (2017). Learning without forgetting.

Lopez-Paz, D. and Ranzato, M. (2017). Gradient episodic memory for continual learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6470–6479.

Mermillod, M., Bugaiska, A., and BONIN, P. (2013). The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in Psychology*, Volume 4 - 2013.

Prabhu, A., Torr, P. H. S., and Dokania, P. K. (2020). Gdumb: A simple approach that questions our progress in continual learning. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II*, page 524–540, Berlin, Heidelberg. Springer-Verlag.

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., and Lampert, C. H. (2017). icarl: Incremental classifier and representation learning.

van de Ven, G. M. and Tolias, A. S. (2019). Three scenarios for continual learning.

Zhou, D.-W., Wang, Q.-W., Qi, Z.-H., Ye, H.-J., Zhan, D.-C., and Liu, Z. (2024). Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9851–9873.