arXiv:2602.00957v1 [cs.LG] 1 Feb 2026

从漂移到适应，再到失败的机器学习模型：工业机器学习运维中的迁移学习

Waqar Muhammad Ashraf伦敦大学学院 图灵研究所 伦敦，英国

Talha Ansar工程与技术大学 巴基斯坦拉合尔

Fahad Ahmed萨希瓦尔燃煤电厂 巴基斯坦萨希瓦尔

Jawad Hussain萨希瓦尔燃煤电厂 巴基斯坦萨希瓦尔

MuhammadMujtaba Abbas工程与技术大学 巴基斯坦拉合尔

Vivek Dua *伦敦大学学院，英国伦敦

摘要

模型适应生产环境对于可靠的机器学习运维（MLOps）至关重要，但当机器学习模型在数据漂移下失效时，如何制定系统化的模型更新框架却较少受到关注。本文比较了采用迁移学习的模型更新策略，包括集成迁移学习（ETL）、全层迁移学习（ALTL）和末层迁移学习（LLTL），用于更新失效的前馈人工神经网络（ANN）模型。本文以某660兆瓦热电厂空气预热器装置的烟气差压为案例进行分析，因为电厂的负荷循环使其呈现出间歇过程的特征。采用三种迁移学习技术更新失效的ANN模型后发现，对于批次大小为5天的情况，ETL的预测准确率相对高于LLTL和ALTL。然而，对于以较大批次（8天）训练的模型，ALTL更适合进行有效更新。不同批次大小下，各种模型更新技术的计算需求（超参数调优和模型训练）呈现出不同趋势。通过这一基于间歇过程的工业案例研究获得的基础性实证见解，可帮助机器学习运维从业者使失效模型适应数据漂移，从而准确监控工业过程。

关键词 模型失效 · 数据漂移 · 以人为中心的监控 · 工业过程

1 引言

工业运行具有非线性特征，这意味着过程条件随时间变化以及设备更换，可能会导致过程变量之间的相关模式发生变化。在“固定数据分布”上训练的“静态”数据驱动模型无法处理工业过程的非平稳数据流，因此不适合部署用于长期监控任务。这是由数据漂移（过程输入变量之间的数据分布发生变化导致协方差偏移）和/或概念漂移（过程输入变量与输出变量之间的潜在关系发生变化）造成的，进而导致模型失效。为使模型在新的数据分布上保持预测准确率，需使用新的数据流重新训练数据驱动模型。此方法可能带来的问题是“灾难性遗忘”，即模型参数经过调整以适应新数据集后，可能无法准确预测模型曾经训练过的历史数据。最初，Geoffrey E. Hinton 假设，对模型参数进行微小调整以适应新数据时，这些调整的累积效应可能会相互抵消，从而不影响历史数据 [1]。然而，McCloskey 和 Cohen [2] 以及 Ratcliff [3] 表明，即使模型仅使用少量新数据进行更新，在互不重叠的数据集上依次训练简单神经网络也会造成显著遗忘。为解决潜在的灾难性遗忘问题，持续

*通讯作者：v.dua@ucl.ac.uk

学习方法用于更新模型参数，并在模型更新机制中通过经验回放和正则化技术，平衡可塑性（学习新的动力学特性）与稳定性（保留历史信息）之间的权衡。另一方面，在线学习会在模型投入生产环境后，持续使用单条数据或一批数据重新训练模型。在线学习使模型能够高度适应传入的数据流或概念漂移，但也会遭遇灾难性遗忘。即时学习是另一种模型更新范式：它针对传入的数据流训练局部模型，并从历史数据中查询相似数据。即时学习能够高精度地近似系统的非线性动力学。然而，它的计算需求和内存需求较高；当系统进入缺少历史数据的新运行模式时，也可能无法进行外推。为在提高生产环境中模型预测准确率的同时降低计算开销，也会采用迁移学习方法：将基于历史数据维护的已训练模型（源域）的参数迁移到“目标”域（发生漂移的数据），只需少量数据即可完成适配。文献研究指出，更新模型对于使其适应不断变化的工业运行至关重要。文献 [4], 中，研究人员在检测到漂移后调整神经网络所有层的参数，并将更新后的模型部署于数字孪生应用。另一项研究 [5], 中，研究人员为完整更新模型参数设定了轮数预算和学习率范围。类似地，迁移学习也被用于使模型适应源域 [6]。然而，尽管研究者大多关注模型参数更新，但尚未系统研究批量大小对基于迁移学习的模型更新技术更新模型参数所产生的影响。更重要的是，尚不清楚批量大小与用于神经网络的迁移学习模型更新技术——包括集成迁移学习（ETL）、全层迁移学习（ALTL）和末层迁移学习（LLTL）——结合后会如何影响模型的预测准确率。围绕模型更新技术的选择所涉及的这些根本问题，尤其是如何权衡可塑性与稳定性以及计算开销，尚未得到系统研究，因而缺乏对解决这些问题的机理性认识。鉴于当前研究的不足，本文研究批量大小及基于迁移学习的模型更新技术（ETL、ALTL、LLTL）对更新前馈人工神经网络以应对概念漂移的影响。本文以一座 660兆瓦热电厂所安装的空气预热器的烟气差压（压差）为案例进行研究。烟气压差与负荷循环（升负荷和降负荷）相关，并被视为间歇过程，这一点可从数据可视化中观察到。因此，可以将本案例的研究结果推广至各类化学工程应用中的间歇过程动力学。本文系统比较三种技术更新模型后的预测准确率、更新模型在权重空间中的变化以及其计算需求。此外，本文还利用 SHapley 加性解释（SHAP）研究特征重要性的演变，以揭示模型更新后在工业监测任务中的稳定性。通过这一工业案例研究获得的基础性实证见解，可帮助机器学习运维从业者调整失效模型以应对数据漂移，从而准确监测工业过程。

2 方法

图1展示了本研究开发的基于迁移学习的模型更新框架，用于监测空气预热器中的烟气压差。分别收集一月和四月各两个月的数据，用于开发模型，批次大小分别为5天和8天。首先，使用一组批次（5天和8天各一个）在超参数优化（学习率、隐藏层神经元数量）的条件下训练人工神经网络模型。随后，使用测试数据（数据集的20%）评估模型的决定系数（R_2）和均方根误差（RMSE）。之后，将训练好的模型部署为机器学习运维系统，用于预测其余批次数据中的烟气压差。当日平均误差（RMSE和平均绝对误差）连续三次高于测试数据误差（模型训练期间观测到的误差）的两倍时，即触发模型更新机制。采用模型失效所在批次的数据以及此前连续批次的数据（用于模型再训练的缓冲数据），通过迁移学习驱动的方法（ETL、LLTL和ALTL）更新模型。随后，再次将这三种方法训练的模型部署到机器学习运维系统中，以预测其余收集到的数据。由此可以将模型的预测准确率与收集数据中已有的真实值进行比较。我们借助SHAP框架，研究不同批次和模型更新方法下训练后及更新后模型的特征重要性变化，以评估模型稳定性。特征变化受过程动态特性的影响，但也可能因模型更新无效及模型中形成的映射而受到干扰。此外，还记录模型训练和更新所耗费的计算时间，以评估基于迁移学习的模型更新方法在机器学习运维中的适用性。

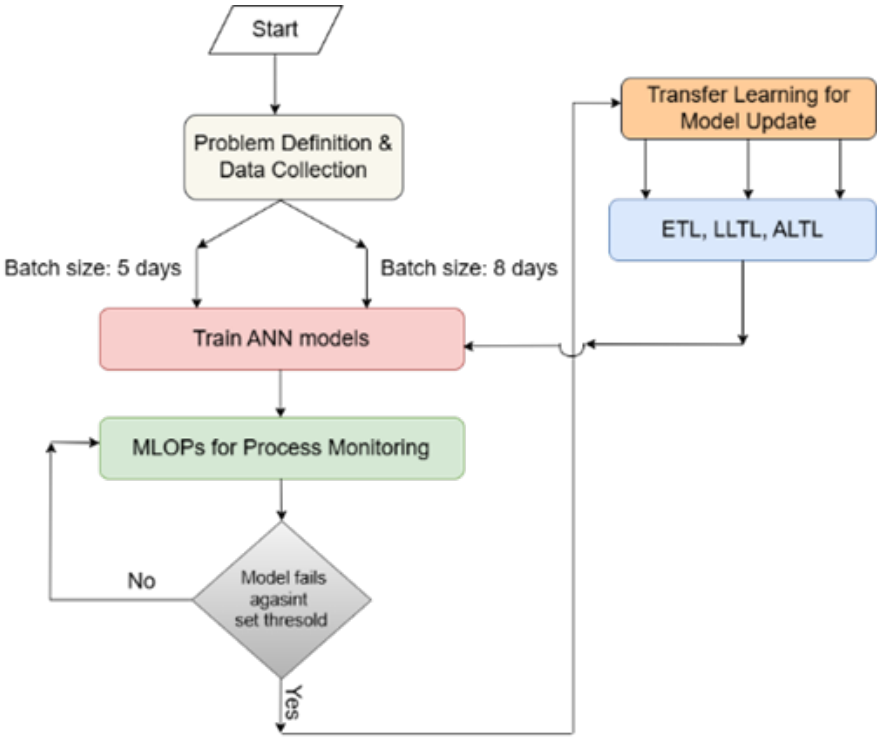

图 1：基于迁移学习的过程监测模型更新框架示意图。

3 结果

3.1 数据特征

用于对空气预热器中的烟气压差（Pa）建模的运行变量如下：烟气入口温度（℃）、烟气出口温度（℃）、二次风入口温度（℃）、二次风出口温度（℃）、一次风入口温度（℃）、一次风出口温度（℃）、氧气入口（%）、氧气出口（%）。为探究模型处理监测任务中不同数据采集采样频率的能力和效能，分别收集一月和四月的变量历史运行数据，平均间隔分别为10分钟和8分钟。这使我们能够模拟实际情形：数据采集采样频率会受到数据存储和维护成本的影响，进而考察模型如何应对这种变化，同时确保预测准确率。图2(a)展示了一月和四月若干运行变量的数据分布曲线。通过目视可以发现变量的数据分布不对称且存在数据漂移，并进一步使用总体稳定性指数（PSI）和克拉默-冯·米塞斯（CvM）进行量化。PSI范围为1.89至23.15，CvM统计量范围为72.4至658.9，且p值小于0.001，这些结果证实了以多时间尺度采样获取的两个数据集中存在漂移。一次风入口温度在两项指标上均表现出最强漂移（PSI=23.15，CvM=658.9），而烟气压差（PSI=1.89）和氧气出口（CvM=72.4）的漂移相对较小，但仍足以引起注意，可能导致模型失效。分别计算两个月中变量对之间的相关性（见图2(b)），结果表明两个数据集中变量对之间的相关性发生了变化（数据中存在概念漂移）。这说明空气预热器存在不同运行模式，有助于评估模型部署到生产环境后的监测能力。

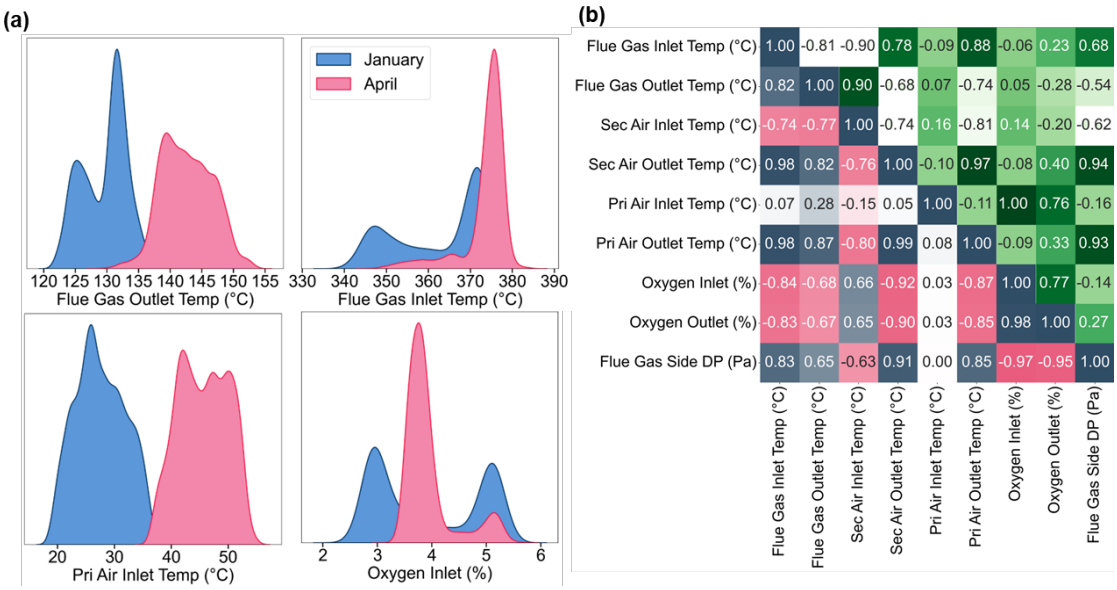

图2：空气预热器热交换数据的特征。(a) 一月和四月收集的数据中可观察到分布偏移。(b) 两个月份的运行变量之间的相关结构也发生了变化（下三角：一月，上三角：四月）。

3.2 在两种不同批次大小下训练和更新 ANN 模型

前馈人工神经网络模型分别使用一组来自一月数据的两个批次进行训练，批次时长分别为5天和8天。训练期间，对人工神经网络模型的超参数（隐藏层神经元数量、学习率、隐藏层数量）进行优化，并使用测试数据（数据集的20%）评估模型。使用5天批次训练的模型测试R_2为0.9824（均方根误差0.0381），而使用8天批次训练的模型测试R_2为0.9807（均方根误差0.0385）。训练阶段结束后，将使用5天和8天批次大小训练的模型部署到生产环境中，以监测空气预热器中的烟气压差。图3(a,b)分别展示了这两个模型的监测准确度。值得注意的是，与使用8天批次大小训练的人工神经网络模型相比，使用5天批次大小训练的人工神经网络模型在预测一月份烟气压差方面表现相对更佳。4月7日至10日期间，两个模型均越过了人工设定的误差阈值，即连续三次超过测试误差上限的两倍。图3(a,b)底部展示了实际响应与基于模型的响应（参考值）之间的对应关系。采用不同数据采样频率训练的模型失效，表明当数据采样频率与模型训练时所用频率不同时，训练好的模型无法很好地扩展。图!2所示的过程动态特性和数据漂移也可能导致模型失效。模型失效会触发基于三种技术（ETL、ALTL和LLTL）构建的模型更新流程，我们分别针对5天和8天的批次大小执行了该流程（使用模型失效所在批次的数据以及紧接其前的批次数据）。需要指出的是，在模型更新期间，仅在较窄范围内迭代优化学习率，不改变模型架构。随后，将采用三种模型更新技术训练的模型部署到机器学习运维系统中，以监测空气预热器中的烟气压差，覆盖四月剩余的数据。图3(c,d)分别展示了采用ETL、ALTL和LLTL训练的模型所得实际响应与基于模型的响应之间的对应关系，分别对应5天和8天的批次大小。同样，图3(e,f)也绘制了一致性图，以直接比较实际响应与基于模型的响应。ETL的均方根误差和平均绝对误差分别为：0.0564、0.0414（5天），0.0652、0.0541（8天）；ALTL：0.0674、0.0563（5天），0.0384、0.0315（8天）；LLTL：0.0400、0.0321（5天），0.0537、0.0438（8天）。将实际响应与基于模型的响应可视化，并结合采用三种更新技术训练的模型所计算出的误差，有助于我们检查模型部署到工业环境后的局部预测性能。对于使用5天批次大小训练的模型，ETL方法监测烟气压差的准确度似乎相较ALTL和LLTL更高（图3(a)）。另一方面，对于使用8天批次大小训练的模型，ALTL的预测性能相对更佳（图3(b)）。对于较小的批次大小，集成模型可能会对预测响应取平均，从而提供稳定的性能。然而，随着批次大小增加，ALTL会更新所有层中的模型参数，使模型适应最新的数据流。这会增加“遗忘”运行历史的风险，并使模型倾向于适应近期的运行动态特性。

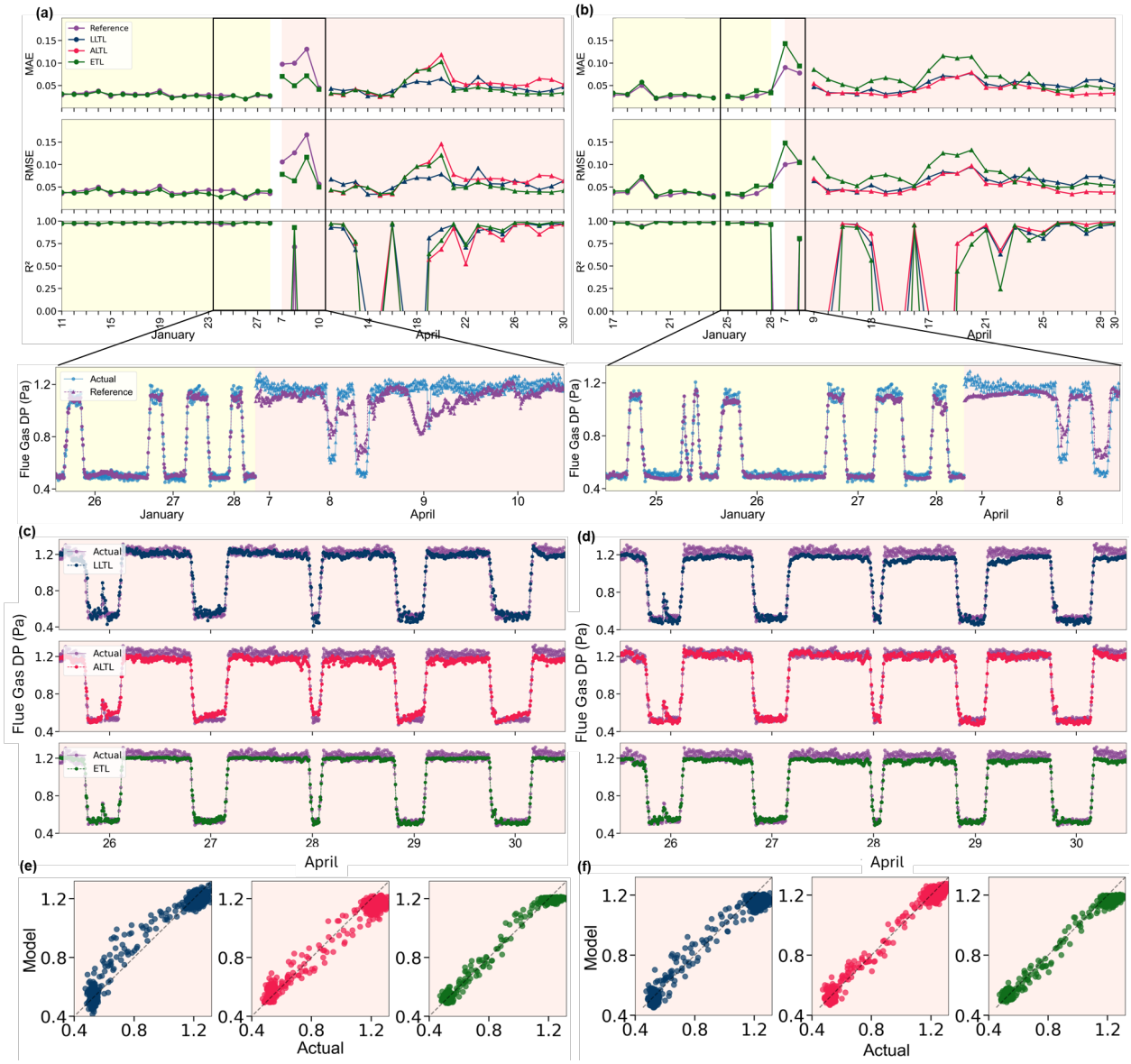

图3：采用（a）5天和（b）8天批次大小训练的人工神经网络模型的机器学习运维，用于监测空气预热器各处的烟气压差。训练后的人工神经网络模型在一月份表现出良好的监测性能。然而，它们在4月7日至10日期间失效，因此实施了模型更新技术。图中还映射了故障区域对应的实际响应和基于模型的响应。（c、d）展示了采用 LLTL、ALTL 和 ETL 技术更新的模型在四月剩余数据上的表现。（e、f）为一致性图，用于比较三种模型更新技术在四月剩余数据上的实际响应和基于模型的响应。

3.3 模型更新后逐层权重空间分布的见解

获取采用三种技术更新的模型逐层权重空间，并将其可视化，以深入了解对应5天和8天批次大小的模型更新阶段之后，模型参数如何调整或变化。将人工神经网络各层的权重空间分布以小提琴图形式可视化，分别对应5天和8天的批次大小，如图4(a,b)所示。总体而言，使用5天和8天批次大小训练的模型，其人工神经网络最后一层的权重空间分布均有所扩大。这可能是由于误差信号从人工神经网络的最后一层开始反向传播，从而调整模型参数。对于ALTL，隐藏层的权重空间似乎比参考分布（预训练

模型空间）可以解释模型在适应四月收集的数据分布中观察到的漂移时，泛化能力为何有所提升。然而，集成方法在拓宽或压缩权重空间方面似乎呈现出不一的结果，这可以归因于多个集成模型根据批次大小消化数据。

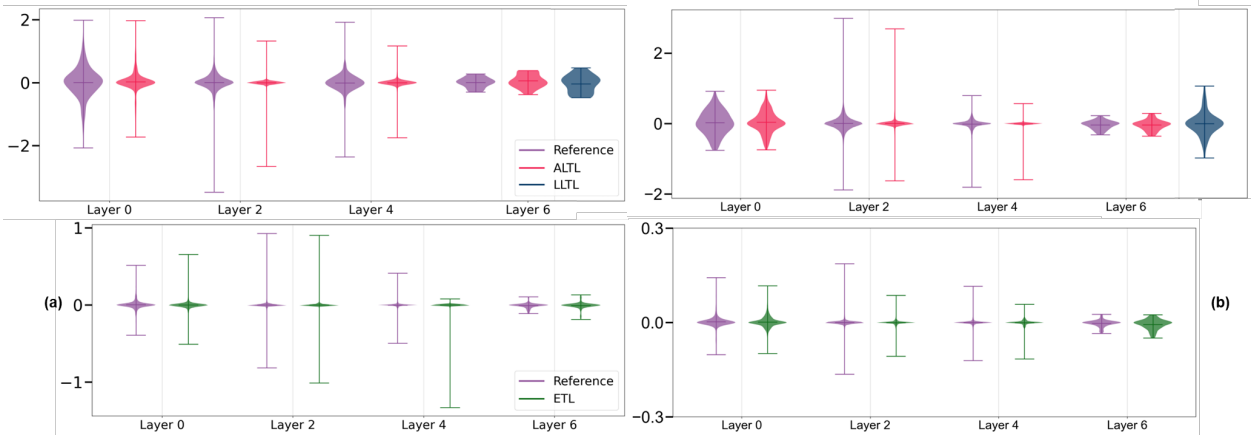

图4：权重空间分　布随模型更新而变化，批次大小分别为（a）5天和（b）8天。I　　n
共同的最后一层权　重空间会拓宽，以适应数据漂移。

3.4 特征重要性的演变

SHAP用于确定预测烟气压差的特征重要性，并据此研究模型更新后特征重要性的变化。参考列表示模型更新前的特征重要性，用作基线，以比较模型更新后的特征重要性排序。如图5(a、b)所示，在批次大小分别为5天和8天的模型更新中，二次风出口温度被发现是影响烟气压差最显著的特征。模型更新后，采用LLTL、ALTL和ETL方法更新的模型仍将二次风出口温度列为最显著的特征。多个ETL列明确跟踪连续批次间隔中的依次更新，每列对应更新后生成的不同版本及其特征重要性排序。三种模型更新技术得出的特征重要性排序一致，尤其是最显著操作变量的排序，这体现了特征重要性分析的准确性，也表明训练后的模型能够稳定地捕捉过程动态特性。此外，这还表明，即使数据中出现概念漂移和数据漂移，特征重要性也可能保持不变（参见图2）。

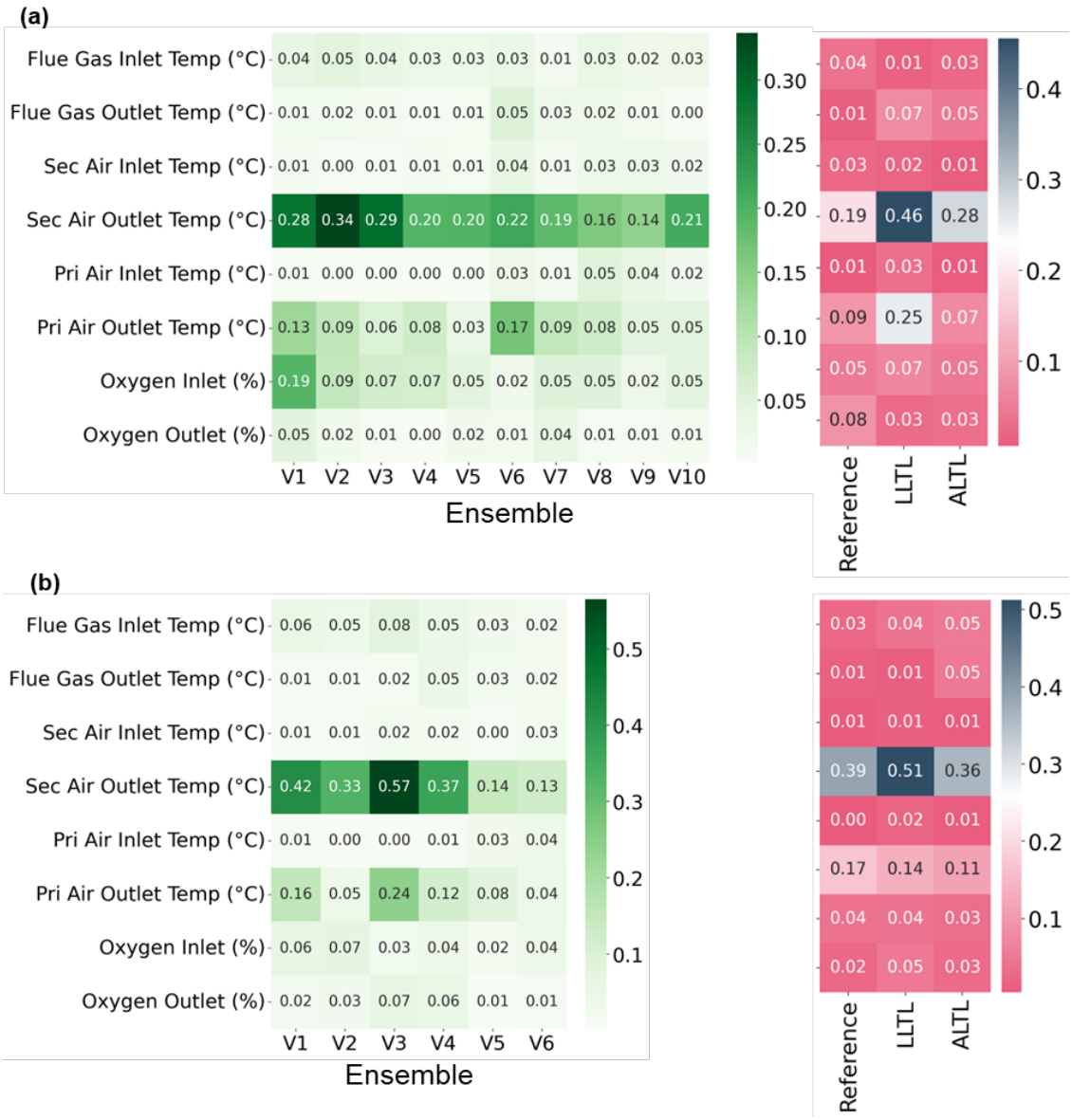

图5：对于 ETL、LLTL 和 ALTL，批次大小为 (a) 5 天和 (b) 8 天时，模型更新过程中基于 SHAP 的特征重要性变化。尽管二次风出口温度 (°C) 始终是最主要的特征，但在信息匮乏期间，一次风出口温度 (°C) 的重要性会暂时上升，以弥补新工况的信息不足（(a) 中的 V6 和 (b) 中的 V3）。积累了足够的数据后，模型趋于稳定，二次风出口温度 (°C) 重新成为最重要的特征。

尽管不同模型更新方法下特征重要性的幅度有所不同，但特征重要性的排序几乎保持不变，显著特征尤为如此。这可能表明模型参数的调优具有稳定性，不受批次大小影响；图3(a、b)所示的稳定误差剖面也印证了这一点。

3.5 模型更新技术的计算性能

图6(a,b)记录并展示了批次大小为5天和8个八小时段时，训练参考模型以及采用三种技术更新该模型所需的计算时间。对于参考模型，采用8天批次大小训练的模型所需的计算时间相对较长（超参数调优时间：1113 s，训练时间：28.4 s），而采用5天批次大小训练的模型所需时间较短（超参数调优时间：868 s，训练时间：14.4 s）。这可能归因于批次数较少时，模型以非光滑方式收敛至误差最小值。不过，采用批次

大小为5天的批次训练的模型，总体而言，使用 LLTL、ALTL 和 ETL 方法进行模型更新所耗费的计算时间，比采用8个八小时段批次大小训练的模型更长。这一现象可能反映出，基线模型训练中的随机误差收敛所产生的参数值，相较于平滑收敛的模型，能够更快地适应新的数据流。比较相同批次大小下模型更新所需的计算时间时，呈现出混合趋势。然而，无论批次大小如何，超参数调优时间仍然长于训练时间。

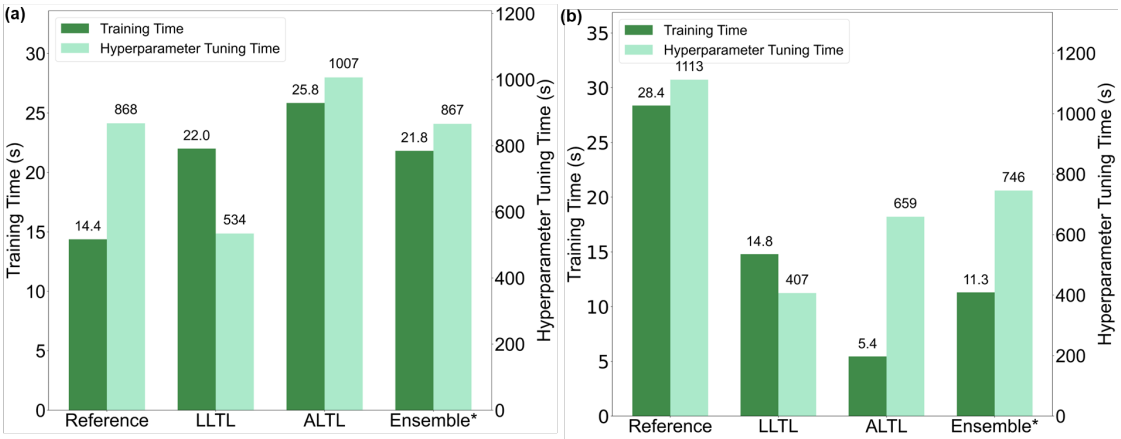

图6：比较批次大小为(a)5天和(b)8天时，超参数调优和模型训练所消耗的计算时间。

4 结论

本文探讨了当前馈人工神经网络模型因漂移（数据漂移和概念漂移）而在机器学习运维中失效时，如何选择模型更新技术。我们研究了批次大小（5天或8天）对前馈人工神经网络模型预测准确率的影响，该模型用于监测安装在一座660兆瓦热电厂中的空气预热器的烟气压差（将其视为间歇过程）。研究发现，模型失效由数据漂移引起；针对5天批次大小（小批次大小）实施的集成方法，其监测性能优于LLTL和ALTL技术。这可归因于集成模型中人工神经网络模型各层之间的权重空间扩大，而使用LLTL和ALTL技术时，权重空间则会受到压缩。对模型更新技术所需计算时间（超参数调优和模型训练）的比较表明，对于5天的批次大小，ETL更新模型耗时更长。不过，对于较大批次大小而言，监测性能更优的ALTL在超参数优化后，更新模型所需时间最短。这一点对机器学习运维实践具有更广泛的意义，因为具有随机误差收敛曲线的模型，可能比初始误差收敛曲线平滑的模型更快适应漂移（数据漂移和概念漂移）。针对工业间歇过程进行模型更新所获得的这些基本见解，凸显了数据批次大小以及模型更新技术选择在工业监测应用中的作用。这些发现还有助于排查机器学习运维流水线的问题，并处理失效的机器学习模型，使其适应多变的工业生产环境。

参考文献

[1] G. E. Hinton. 分布式表征。可访问：http://shelf2.library.cmu.edu/Tech/19334156. pdf，1984年。[2] M. McCloskey 和 N. J. Cohen。联结主义网络中的灾难性干扰：序列学习问题。载于学习与动机心理学，第109–165页。爱思唯尔，1989年。[3] R. Ratcliff。再认记忆的联结主义模型：学习与遗忘函数所施加的约束。心理学评论，97(2):285，1990年。[4] C. M. Rebello 和 I. B. R. Nogueira。化学工程中的数字孪生：用于识别、实施、在线学习和不确定性评估的集成框架。《计算机与化学工程》，200:109178，2025年。

[5] B. Metcalfe 等。面向工业规模补料分批发酵实时预测的机器学习运维（MLOps）软测量。《计算机与化学工程》，194:108991，2025。[6] M. F. Theisen、G. M. Meesters 和 A. M. Schweidtmann。利用过程拓扑迁移图神经网络以进行软测量建模。arXiv 预印本 arXiv:2502.06826，2025。