云环境中以机器学习为核心的工作负载预测模型性能分析

Deepika Saxena、Jiten dra Kumar、Ashutosh Kumar Singh，*IEEE*高级会员和 Stefan Schmid

摘要——由于异构服务类型和动态工作负载具有高度的可变性和多维性，精准估计资源使用量是一项复杂且充满挑战的任务。过去几年，资源使用量和流量预测受到了研究界的广泛关注。研究人员利用机器学习的计算能力和学习能力，开发了许多基于机器学习的工作负载预测模型。本文首次对多种机器学习驱动的云工作负载预测模型开展系统性综述，并基于性能分析进行比较研究。本文首先阐述预测性资源管理的重要性，随后以示意图介绍这些工作负载预测模型的运行设计、研究动机及相关挑战。本文聚焦于现有最先进工作负载预测方法的理论概念和数学原理，将不同预测方法划分为五个类别，并介绍其分类体系。本文全面综述并比较了属于不同机器学习模型类别的主要预测方法。研究将所有五类基于机器学习的工作负载预测模型部署在同一平台上，通过实验分析，利用三种不同的基准云工作负载跟踪数据进行系统调查与比较。本文评估了最先进方法的关键性能指标，以供比较；最后讨论了其中的权衡与重要备注。

索引词—云计算、深度学习、量子神经网络、集成学习、混合学习、进化神经网络、预测。

arXiv:2302.02452v1 [cs.DC] 5 Feb 2023

1 引言

THE 云计算（CC）范式具备快速弹性、资源池化、外包服务管理、广泛的网络访问和按使用量付费模式等特性，能够为企业、学术界、研究机构及所有利益相关者提供可扩展的计算途径，同时将前期资本投入降至最低 [1, 2]。在全球 IT 基础设施面临种种不确定性之际，CC 通过保持发展势头并应对其中的不一致性，成为推动业务进步的催化剂 [3]。根据一份最新调查报告 [4],，预计到 2030 年，全球云计算市场规模将达到 1,554.94 亿美元，复合年增长率（CAGR）为 15.7%。此外，物联网（IoT）、雾计算和边缘计算、信息物理系统等新兴技术都高度依赖 CC，因为它们的存储和计算能力不足 [5]。

1.1 研究动机

云服务提供商（CSP）通过对数据中心的物理资源进行虚拟化 [6–10] ，在满足计算实例需求的同时实现收益最大化，并享有快速扩展能力 [11–13]。因此，云计算基础设施的全面管理完全取决于对资源进行细粒度配置，包括存储、处理和网络等 [14–18]。资源需求随时间呈现高度波动，导致物理机利用率过高或过低，并引发服务级别协议（SLA）违规问题 [19]。在负载到达峰值时，虚拟机资源的总需求会超过服务器的可用资源容量，导致服务器过载和性能下降，例如，部分虚拟机可能崩溃、资源不可用时间延长以及响应时间增加等。而资源需求不足则会造成计算资源浪费。为管理资源容量的动态和随机需求，或应对过载／负载不足，需要实时将虚拟机从过载／负载不足的服务器迁移到具有充足资源容量的其他服务器，这会导致执行延迟。在这种情况下，通过预先估算有效处理到达的工作负载至关重要。准确的负载预测可减少资源浪费和功耗，并通过仅启用所需数量的物理机来减少活动服务器数量。精确的工作负载信息有助于预先预留资源，从而有效执行和管理即将到来的工作负载，减少响应时间、SLA 违规、过度配置和配置不足问题，并提高资源利用率、可靠性和服务可用性。 [20–23]。

1.2 工作负载预测视角

图 1 通过“云基础设施”与“工作负载预测”模块之间的交互式信息流，展示了工作负载预测在物理资源管理中的视角与作用

• *D. Saxena* 任职于德国法兰克福歌德大学计算机科学系。电子邮箱：*13deepikasaxena@gmail.com, d.saxena@em.uni-frankfurt.de J.Kumar* 任职于印度泰米尔纳德邦蒂鲁吉拉伯利的库鲁克谢特拉国家理工学院计算机应用系。电子邮箱：*jitendra@nitt.edu A. K. Singh* 任职于库鲁克谢特拉国家理工学院计算机应用系。电子邮箱：*ashutosh@nitkkr.ac.in S. Schmid* 任职于德国柏林工业大学、奥地利维也纳大学和德国弗劳恩霍夫安全信息技术研究所。电子邮箱：*stefan.schmid@tu-berlin.de*。

本文已被 IEEE《并行与分布式系统汇刊》接受。© 2023 IEEE。允许个人使用本材料。对于任何其他用途，无论是在当前还是未来的媒体中，包括为广告或促销目的重印/再版本材料、创作新的合集作品、转售或再分发至服务器或列表，或在其他作品中重复使用本作品的任何受版权保护部分，均须获得 IEEE 的许可。本作品可免费用于综述和引用。

以及“工作负载预测”模块。融入虚拟化技术的服务器集群 $\{ServerP_1, ServerP_2,\ldots, ServerP_n\}$ 能够创建大量虚拟机形式的虚拟实例，以满足云用户的服务需求。借助调度器和虚拟机监控程序，虚拟化技术使各种应用（如 AWS、Docker、MS Azure 和 GCP 等）能够共享物理机器。资源使用信息会受到监控

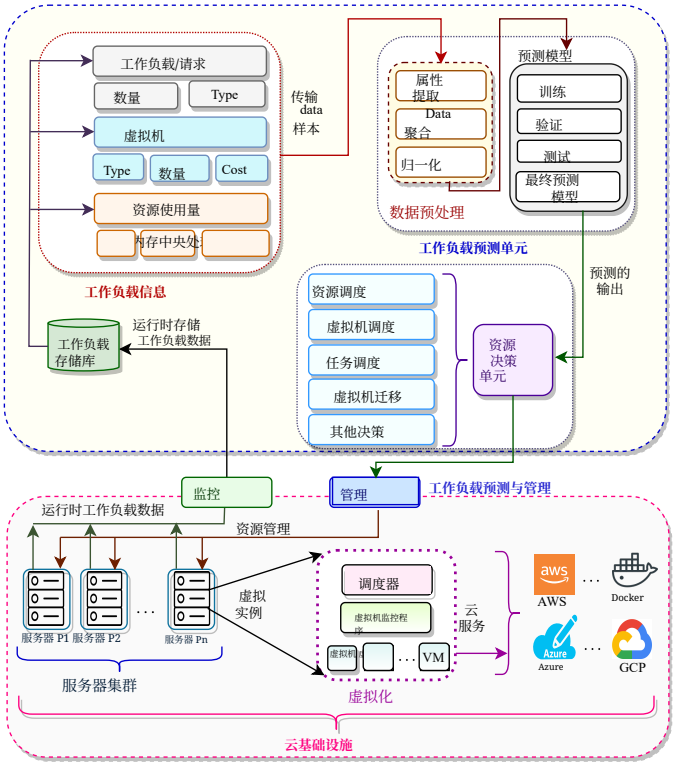


图1：工作负载预测的示意图及其应用

并记录在工作负载预测与管理单元中的工作负载存储库内。工作负载存储库中的原始信息包括请求的数量和类型、虚拟机的数量、类型及成本，以及资源（即中央处理器、内存、带宽）的使用情况；这些信息会被检索出来并传输至工作负载预测单元。在数据预处理过程中，会从原始数据样本中提取重要属性，并进行汇总和归一化。随后采用工作负载预测模型，该模型通过训练、验证和测试等一系列阶段生成并不断演进，以实现实时工作负载预测。最终预测模型会分析并估算有关资源使用量、请求数量和类型等信息，以制定有效的资源管理决策。预测的资源信息有助于主动安排所需的物理资源，避免运行时资源配置延迟，同时满足服务质量（QoS）约束。

1.3 研究挑战

显然，云工作负载预测在主动自动扩缩和资源动态管理中发挥着至关重要的作用，从而提升系统的可扩展性、吞吐量、可持续性和容错性通过主动预测系统故障来实现。然而，云工作负载预测仍面临一些重大挑战，具体讨论如下：

- 异构工作负载：云用户提交不同类型的应用请求，这些请求需要异构的资源容量，并具有不同的优先级和与各自服务级别协议相关的定价策略。
- 不确定的资源需求：资源需求会随时间变化——在小时、天、周、月乃至年尺度上，取决于工作负载类型以及用户提交的执行期限。有时流量会出现突发 [24]，这使得预测即将到来的资源需求并确定资源分配变得困难。
- 动态适应：由于云环境高度多变且动态，它会受到意外波动的影响，这给工作负载预测的适应能力带来了关键挑战，即需要进行调整或重新生成，以适应不断变化的工作负载并保持高效运行。
- 数据粒度和预测窗口大小：确定数据样本或数据粒度的合适大小，以及预测窗口的长度（即较短或较长的时间间隔），是另一项关键挑战，它会直接影响相关模式的学习以及从提取的模式中建立相关性。

1.4 论文结构与贡献

本文全面研究了基于机器学习的云工作负载预测模型。研究首先在第1节讨论云计算（CC），阐述云计算环境中工作负载预测的重要作用及其研究动机。随后，本文通过示意图及说明，展示负载预测与管理在云数据中心（CDCs）中的应用。之后，本文讨论研究挑战，指出准确的工作负载预测对于资源管理的必要性及其显著影响，并探讨相关关键问题。第2节介绍工作负载预测的基本操作流程。第3节阐述本文采用的研究方法。本研究旨在全面综述针对广泛云工作负载预测而提出的最具代表性和开创性的机器学习模型。因此，第4节讨论基于进化神经网络的预测模型，第5节和第6节分别综述基于深度学习和混合学习的预测模型，第7节和第8节则分别讨论基于集成学习和量子学习的预测模型。此外，第9节在统一平台上实现上述五类预测模型，并根据各项关键绩效指标（KPI）评估和比较其性能。最后，第10节总结本研究，讨论不同类别预测模型之间的权衡，并阐述云工作负载预测领域新兴的研究挑战及其可能的解决途径。据作者所知，这是首篇

旨在针对云计算（CC）资源管理背景下的基于机器学习的工作负载预测模型开展全面实验研究的论文。本文的主要贡献如下：

- 根据概念和运行特征，将基于机器学习算法的优秀且最新的云工作负载预测模型划分为不同类别，并构建分类体系（图3）。
- 为阐释各类别工作负载预测方法对应的通用概念设计和运行设计，本文提出了五种具体的机器学习模型架构及其运行策略。
- 本文对最新研究成果的所有关键细节进行了深入讨论和比较，并分析其特征，以确定未来的研究范围，从而解决相应类别预测模型的局限性。
- 本文在同一平台上实现了五类预测模型所对应的方法，并开展了深入的实验分析和比较，通过关键绩效指标衡量其性能，随后讨论了各种权衡及值得注意的备注。

表 1 列出了全文中使用的符号、符号说明和缩写的解释。

表 1：符号及其解释

符号	定义	符号	定义
$\mathcal{W}^{Ac}$	实际工作负载	$\mathcal{X}$	单元信息
$\mathcal{W}^{Pr}$	预测工作负载	$\mathcal{D}$	输入数据
$\mathcal{G}_1$	LSTM 第一层	$\mathcal{B}$	bias
$\mathcal{CF}_{RU}$	先前资源使用信息	$\mathcal{WT}$	权重矩阵
$\mathcal{G}_2$	LSTM 的 sigmoid 层	w	神经权重
n	输入层节点数	Θ	量子比特
p	隐藏层中的节点数	$\uplus$	激活函数
z	基预测器（BP）的数量	y^{In}	量子比特输入向量
$\mathcal{MSE}$	均方误差	$\mathcal{MAE}$	平均绝对误差
m	数据样本数量	ENN	进化神经网络
CSP	云服务提供商	CC	云计算
$LSTM$	长短期记忆	CDC	云数据中心
RNN	递归神经网络	DBN	深度置信网络
	双向长短期记忆网络	DNN	深度神经网络
SGD	随机梯度下降	GRU	门控循环单元
PLR	分段线性表示	LR	逻辑回归
TSA	顶部稀疏自编码器	CP	正则多线性
OED	正交实验设计		萨维茨基-戈莱滤波器 Savitzky-Golay
PSO	P粒子群优化		反向传播 反向传播
ENN	进化神经网络	QoS	服务质量
$RMSE$	均方根误差	MAE	平均绝对误差
$MAPE$	平均绝对百分比误差	MSE	均方误差
	均方分段根误差	VM	虚拟机
$MRPE$	平均相对预测误差	DL	深度学习
CDF	累积分布频率		对数均方根误差
$SaDE$	自适应差分进化	AEF	绝对误差频率
$BaDE$	双阶段自适应差分进化	$EQNN$	进化量子神经网络
$TaDE$	三重自适应差分进化	$c\text{-}NOT$	受控非门
$ARIMA$	自回归积分滑动平均	HL	混合学习
QNN	量子神经网络	EL	集成学习

2 工作负载预测操作流程

工作负载预测所需的基本步骤通过图2中所示的操作设计示例进行了概述。考虑输入数据 $\{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_n\} \in \mathcal{D}$ 从工作负载存储库中保存的原始数据提取而来。数据提取 会从原始数据中筛选相关属性，以改进模式学习，并在提取的模式之间建立更直观的相关性。数据聚合

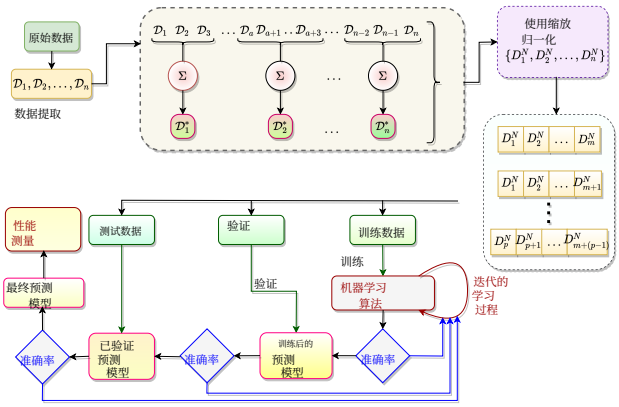


图2：负载预测操作流程

操作会按照选定的预测窗口大小（例如，五分钟）对提取的数据进行组装，例如，将 $\{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3\}$ 组合起来生成一个聚合数据样本 $\mathcal{D}_1^*$。同样，$\{\mathcal{D}_a, \mathcal{D}_{a+1}, \mathcal{D}_{a+2}\}$和 $\{\mathcal{D}_{n-2}, \mathcal{D}_{n-1}, \mathcal{D}_n\}$ 经过聚合后，分别生成数据样本 $\mathcal{D}_a^*$ 和 $\mathcal{D}_n^*$。聚合数据样本会缩放到特定范围 $[x_a, x_b]$，并使用公式（1）转换为归一化数据样本 $\{\mathcal{D}_1^N, \mathcal{D}_2^N, ..., \mathcal{D}_n^N\}$，其中 ϖ_{min}^{In} 和 ϖ_{max}^{In} 分别是输入数据集的最小值和最大值，而归一化数据向量记为 ϖ^{In}，它是所有归一化输入数据取值的集合。实验中，x_a 和 x_b的取值分别设为 0.0001 和 0.999。这些归一化取值按照公式（2）整理为二维输入和输出矩阵，分别记为 ϖ^{In} 和 ϖ^{Out}：

$$\hat{\varpi^{In}} = x_a + \frac{d_i - \varpi_{min}^{In}}{\varpi_{max}^{In} - \varpi_{min}^{In}} \times (x_b) \tag{1}$$

$$\varpi^{In} = \begin{bmatrix} \varpi_1 & \varpi_2 & .... & \varpi_z \\ \varpi_2 & \varpi_3 & .... & \varpi_{z+1} \\ . & . & .... & . \\ \varpi_m & \varpi_{m+1} & .... & \varpi_{z+m-1} \end{bmatrix} \varpi^{Out} = \begin{bmatrix} \varpi_{z+1} \\ \varpi_{z+2} \\ . \\ \varpi_{z+m} \end{bmatrix} \tag{2}$$

因此，按照图2中滑动窗口模块所示，构建了预测滑动窗口。这些归一化数据样本分为三类，即训练数据、验证数据和测试数据。指定一个机器学习算法作为预测模型，并将训练数据输入其中，使其在迭代学习或优化过程中学习特定模式。随后，使用 RMSE 或 MSE 等误差评估函数评估该预测模型，并对其准确率进行测试。如果达到期望的准确率，则得到训练后的预测模型，并再次将验证数据输入该模型，以检查预测准确率。如果达到最佳准确率，则建立验证后的预测模型 。最后，将测试数据/未见数据输入验证后的预测模型，并进行准确率评估。如果连续达到期望的准确率，则部署最终预测模型；否则，预测模型的相应阶段将退回迭代学习过程。通过获得用于云资源管理的预测输出，衡量其性能。

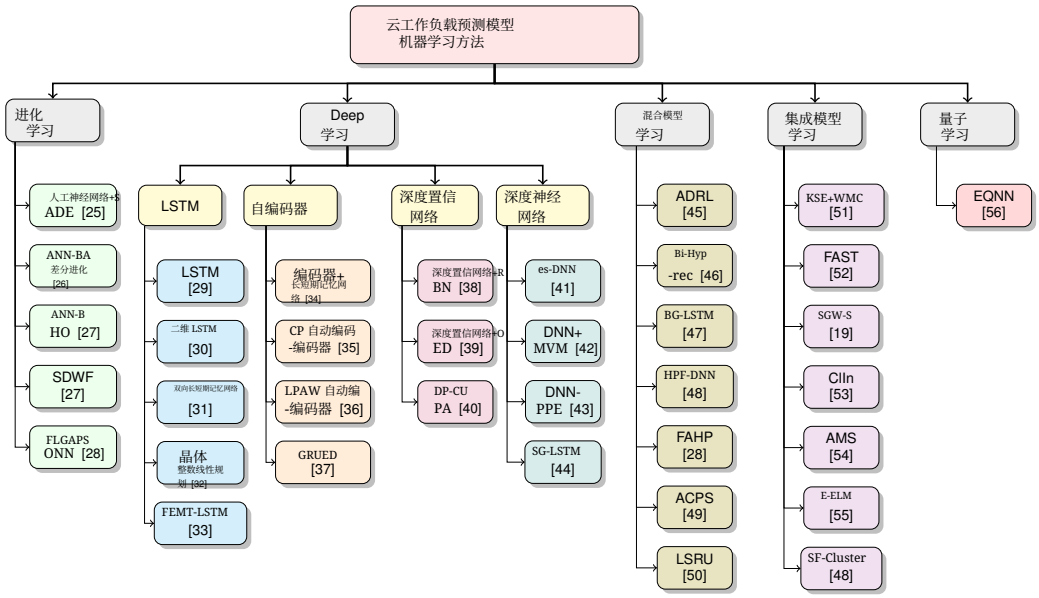


Fig. 3：基于机器学习的工作负载预测模型的分类与分类体系

3 研究方法

本节详细阐述综述方法。综述过程包括收集主要的云工作负载预测论文，这些论文提出的方法基于机器学习算法和概念。我们检索、研读并分析了发表于 IEEE、ACM 数字图书馆、Elsevier、Springer 和 Wiley 等顶级期刊及会议数据库中的开创性质量预测模型，以进行比较研究。此外，我们根据所提出方法的基础，通过识别主要研究对收集的论文进行筛选；随后采用特定的纳入标准，将方法相似或特征重叠的论文归入同一类别。为避免研究过程中的任何偏置，后续各节的综述由一位作者撰写，并经与另一位合著者讨论及反复审阅后定稿。不同作者对现有预测模型进行了全面探讨，以确保本研究的完整性并比较模型性能。在为各类别选取相关工作时，我们剔除并避免纳入平均及低于平均水平的研究，以便对现有最佳工作负载预测方法进行清晰、简洁的综述。因此，图3展示了分类及分类体系表示，根据概念和运行特征，将现有预测模型划分为监督式机器学习方法的特定类别和子类别。

相应地，五种互不重叠的工作负载预测类别分别为进化学习、深度学习、混合学习、集成学习和量子学习。在进化学习类别中，候选方法 ANN+SaDE [25], ANN-BADE [26], ANN-BHO[27], SDWF [27], 和 FLGAPSONN [28] 采用了进化优化算法进行学习过程。

或神经网络层的权重更新过程。深度学习领域的相关研究进一步分为四个子类，包括长短期记忆（LSTM）单元、自编码器、深度置信网络和深度神经网络。LSTM 子类包括 LSTM[29], 2D LSTM [30], 双向长短期记忆网络 [31], Crystal ILP [32], 和 FEMT-LSTM [33] ，它们主要基于 LSTM 模型的功能。自编码器子类包括 Encoder+LSTM [34], CP 自动编码器 [35],LPAW 自动编码器 [36], 和 GRUED [37] ，这些方法是在传统自编码器的基础上进行一些有益改进而衍生出来的。同样，DBN+RBN [38], DBN+OED [39], DP-CUPA [40]以及 es-DNN [41], DNN+MVM [42], DNN-PPE [43], SG-LSTM [44] 分别属于深度置信网络和深度神经网络子类。混合学习类别代表了多种机器学习算法和方法的集成，其中包括 ADRL [45], Bi-Hyprec [46], BG-LSTM [47], HPF-DNN [48],FAHP [28], ACPS [49], 和 LSRU [50]。同样，集成学习通过基学习器和决策的概念来估计最终结果。KSE+WMC [51], FAST[52], SGW-S [19], CIIn [53], AMS [54], E-ELM [55], 和 SF-Cluster [48] 等研究采用了集成学习。最后，量子学习类别包括 EQNN 模型 [56] ，该模型基于量子计算和神经网络学习原理开发。

4 基于进化神经网络的预测模型

借助进化优化算法实现学习过程的神经网络称为进化神经网络（ENN）。图 4 展示了 ENN 的示意图和运行视图。考虑一个前馈神经网络 n-p_1-p_2-q，其输入层和输出层分别包含 n、q 个节点，三个连续隐藏层分别包含 p_1和 p_2 个节点。将一个训练输入数据向量：$\{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_n\} \in \mathcal{D}$ 输入到

该神经网络的输入层中，其中一层的每个节点都通过突触/神经权重 $\{w_1, w_2,\ldots, w_z\}$与下一层的所有节点相连，$\in \mathcal{W}$、z 表示任意两个连续神经层之间权重连接的总数。训练输入向量（$\mathcal{D}$）通过加权连接，利用方程（3）、（4）和（5）进行前向传播；$\uplus$ 是对每个神经元计算的线性函数；$\mathcal{B}$ 是偏置向量。权重连接的取值 $\{w_1, w_2, \ldots, w_z\}$ 决定输入向量对神经元输出向量的影响，并决定神经元之间突触互连的强度。

$$\uplus = (\mathcal{D}_1 \times w_1) + (\mathcal{D}_2 \times w_2) + \ldots + (\mathcal{D}_n \times w_n) \quad (3)$$

$$\mathcal{D}^{\dagger} \cdot \mathcal{W}^{\dagger} = (\mathcal{D}_1 \times w_1) + (\mathcal{D}_2 \times w_2) + \ldots + (\mathcal{D}_n \times w_n) \quad (4)$$

$$\uplus = \mathcal{D}^{\dagger} \cdot \mathcal{W}^{\dagger} + \mathcal{B} \quad (5)$$

ENN 预测模型通过调整 ENN 的偏置和连接权重取值 $\{w_1, w_2, \ldots, w_n\} \in \mathcal{W}^{\dagger}$，以最小化预测误差。该学习过程通过应用不同的进化优化算法来实现，这些算法从由 $\mathcal{Z}$ 个网络 $\{w_1^{\dagger}, w_2^{\dagger}, \ldots, w_{\mathcal{Z}}^{\dagger}\}$ 组成的随机种群中选出最优网络。算法通过更新网络种群，广泛探索和利用多样化的网络种群，反复优化神经连接权重的取值。在连续的训练轮次中，通过评估适应度函数来选择最佳候选网络，例如使用均方根误差（RMSE）、平均绝对百分比误差（MAPE）等估算预测误差。为实现 ENNs 的在线优化，学习过程会定期重复进行，使用工作负载的先前样本和最新样本，以离线模式进行训练和再训练，同时实时分析实时工作负载。目前已有多种利用 ENN 进行云工作负载预测的方法，本节旨在进一步全面讨论和分析这些方法。

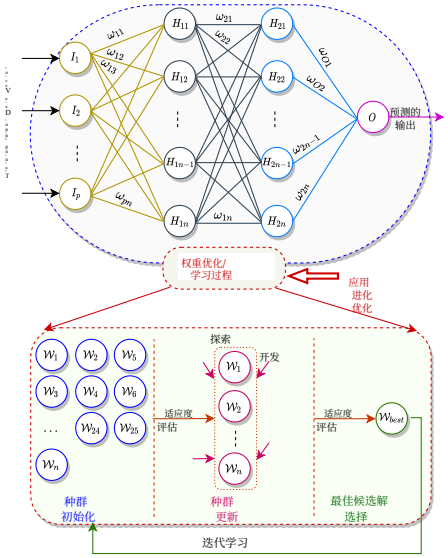


图 4：基于 ENN 的负载预测

本文提出了一种基于 ENN 的云工作负载预测方法 [25]，其中采用自适应差分进化（SaDE）训练三层前馈神经网络。在学习过程中，通过选择性地采用三种变异策略，并随后进行均匀交叉，利用探索和开发操作更新网络群体。与采用反向传播训练的神经网络（BPNN）相比 [57]，该方法取得了更高的准确率，因为前者进行多维学习，而后者仅优化单个解。Kumar 等人 [26] 提出了一种基于双相自适应差分进化（BaDE）学习的神经网络，该方法采用双重自适应机制：在开发过程中对交叉进行自适应，在探索阶段对变异进行自适应，从而提高神经网络的学习效率。因此，该方法在预测准确率方面优于 SaDE [25]。文献 [58] 提出了一种自适应神经网络，使用三重自适应差分进化（TaDE）算法调整网络连接权重。在此方法中，自适应机制作用于交叉、变异和控制参数生成，从而增强预测模型的学习能力。Kumar 等人 [59] 采用黑洞优化（BHO）算法优化神经网络权重，并构建用于动态资源扩展的工作负载预测模型。该进化优化算法会更新恒星（即随机初始化的网络向量）的移动，并跟踪其位置，以判断其是否到达事件视界。此外，文献 [27] 通过在前馈神经网络的迭代学习过程中引入局部和全局黑洞（即最优解）的概念，对现有 BHO 算法进行改进，提出了增强型 BHO（即 E-BHO）。该方法还计算近期预测结果的偏差，并将其用于提高后续预测的准确率。Malik 等人 [28] 提出了一种基于 ENN 的多资源利用率预测方法。在这项工作中，采用结合遗传算法（GA）和粒子群优化（PSO）的混合进化算法优化函数链接神经网络（FLNN）的权重。该方法与 FLNN、结合 GA 的 FLNN（FL-GANN）以及结合 PSO 的 FLNN（FLPSONN）进行了比较，并通过这些方法验证了其性能。

表2总结了现有基于 ENN 的工作负载预测模型，重点列出了其主要特征、实现细节、性能指标、参数调整情况，以及学习过程中的预期计算复杂度。 [25, 26, 58] 中讨论的方法基于差分进化，需要调整多个控制参数，包括交叉率、变异率，跟踪动态或固定的学习周期，并记录每个训练轮次中失败和成功候选项的数量。而[27, 59] 中采用 BHO 的预测方法只需跟踪事件视界半径，并根据各候选项的适应度值与视界半径的比较结果来更新下一代种群。因此可以分析得出，与基于 BHO 的预测方法相比，基于 DE 的预测方法消耗的训练时间更多、涉及的计算复杂度更高，所需的训练轮数也更多。基于上述因素， [27]

著名贡献者（时间线）	模型/方法/框架	工作流程/策略	数据集	实现/仿真tool	预测的 parameters	错误指标	结果或备注
Kumar 等人 [25] (2018年)	ANN-SaDE	前馈神经网络采用自适应差分进化算法进行训练	NASA and 萨斯喀彻温省	MATLAB	数量 of 请求 per 单位时间	MSE	错误最多降低至 0.001 且准确率提高了是反向传播的 168 倍
Kumar 等人 [26] (2020年)	ANN-BaDE	三种变异策略和三种交叉策略基于差分进化算法中的自适应机制进行优化人工神经网络	美国国家航空航天局，萨喀彻温省，谷歌 集群 跟踪记录	Python	数量 of 请求，中央处理器、内存使用率	MSE	准确率最高分别提升至 91% 和 97%，优于 SaDE 和反向传播
Saxena等人 [58] (2020年)	ANN-TaDE	三种变异策略和三种交叉策略，采用基于控制参数调优的自适应方法，在差分进化算法中优化人工神经网络	NASA and 萨斯喀彻温省	Python	数量 of 请求 per 单位时间	RMSE	准确率提高了 97.4% 以及 94.8% 优于反向传播 and SaDE，分别
Kumar 等人 [59] (2016年)	神经网络 with 黑洞优化 tion	黑洞算法用于神经网络的学习过程，该网络使用预处理后的训练数据	来自美国国家航空航天局、卡尔加里和萨斯喀彻温省 Web 服务器	MATLAB	数量 of 请求 per 单位时间	MSE	错误减少了多达 134 倍 优于反向传播
Kumar 等人 [27] (2021年)	Self 定向工作负载预测方法 （SDWF）	通过计算捕捉预测误差趋势近期预测结果的偏差，并应用于提高后续预测的准确率	NASA and 萨斯喀彻温省 HTTP 跟踪记录，Google 集群	MATLAB	数量 of 请求，中央处理器、内存使用率	MSE	与对比方法相比，错误率最高可达 99.99%
Malik 等人 [28] (2022年)	FLGAPSONN: FLNN +遗传算法+ 粒子群优化	将预处理训练数据输入 FLNN 其通过遗传算法的混合算法进行优化和粒子群优化	谷歌 集群 跟踪记录	Python	CPU，内存，磁盘使用率	MAE	提高了 准确率 by 21.87%、13.75% 和 30.55% 相较于 FLPSONN、FLGANN，以及 FLNN，分别

是所有讨论的方法中最适宜的一种。

5 基于深度学习的预测模型

深度学习模型是一类预测模型，对云计算领域产生了巨大影响。深度

学习策略如图5所示，其中，预处理后的训练输入数据样本：$\{\mathcal{D}_1, \mathcal{D}_2,...、\ \mathcal{D}_n\}$被输入长短期记忆网络（LSTM）、门控循环单元（GRU）、深度信念网络（DBN）、深度前馈神经网络（DNN）、自编码器、递归神经网络（RNN）等深度学习算法，以进行学习。定期调整和重新调整各深度学习算法的基本超参数 $\{Hyp1, Hyp2, ...、\ HypN\}$，以创建并更新基于深度学习的预测模型。随后，使用验证数据对训练好的模型进行评估，以估计其性能和准确率。由此，可获得使用最优或最适宜超参数优化的深度学习模型。对于

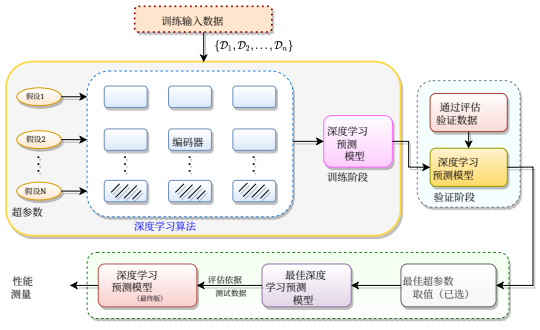


图5：基于深度学习的预测运行视图

实例采用基于 LSTM 的深度学习算法进行预测。实际负载的资源使用量（$\mathcal{Z}^{Ac}_{\mathcal{RU}}$）作为历史工作负载存储，并作为输入馈入神经网络输入层，以预测未来资源使用量。基于 LSTM-RNN 的预测模型由包含四个神经网络层的单元构成，其中前一个单元（$\mathcal{X}^{t-1}$）信息会传递到当前单元（$\mathcal{X}^t$）。第一层（$\mathcal{G}_1$）应用式（6），以确定传递到下一状态的先前资源使用信息（$\mathcal{CF}^t_{\mathcal{RU}_i}$）的量；其中，$\mathcal{WT}$ 是权重矩阵，$\mathcal{B}$ 是偏置值，$\mathcal{Z}^{Pr}_{t-1}$ 和 $\mathcal{Z}^{Pr}_t$分别是先前输出和当前输入。单元状态通过两个网络层更新，即 $sigmoid$层（$\mathcal{G}_2$）使用式（7）确定要更新的取值（$\mathcal{I}^t$），以及 $tanh$ 层使用式（8）生成新的候选值向量（$\hat{}\mathcal{X}^t$）。最后，式（9）结合两个输出以更新单元状态。

$$\mathcal{CF}^t_{\mathcal{RU}_i} = \mathcal{G}_1(\mathcal{WT}_{\mathcal{CF}} \cdot [\mathcal{Z}^{Pr}_{t-1}, \mathcal{Z}^{Ac}_t] + \mathcal{B}_{\mathcal{RU}}) \quad (6)$$

$$\mathcal{I}^t = \mathcal{G}_2(\mathcal{WT}_{\mathcal{I}} \cdot [\mathcal{Z}^{Pr}_{t-1}, \mathcal{Z}^{Ac}_t] + \mathcal{B}_{\mathcal{I}}) \quad (7)$$

$$\hat{\mathcal{X}}^t = tanh(\mathcal{WT}_{\mathcal{X}} \cdot [\mathcal{Z}^{Pr}_{t-1}, \mathcal{Z}^{Ac}_t] + \mathcal{B}_{\mathcal{X}}) \quad (8)$$

$$\mathcal{X}^t = \mathcal{CF}^t_{\mathcal{RU}_i} \times \mathcal{X}^t + \hat{\mathcal{X}}^{t-1} \times \mathcal{I}^t \quad (9)$$

对服务器上不同虚拟机的预测流量（$\mathcal{Z}^{Pr}_{t+1}$）的资源利用率（RU）进行汇总，以主动识别任何过载，并将预测资源利用率最高的虚拟机迁移到高效服务器，从而缓解过载。后续章节将全面综述现有工作负载预测模型，这些模型属于深度学习方法的四个不同类别。

5.1 LSTM-RNN

文献 [29] 提出了一种使用基于长短期记忆的循环神经网络（LSTM-RNN）的细粒度云工作负载预测模型，该模型能够学习长期依赖关系，并实现高准确率的主机负载预测。Tang [30]通过利用隐藏层的基于周的依赖关系和权重并行化算法，提出了一种二维LSTM神经网络单元结构。该研究通过给出并行LSTM算法的数学描述，并采用误差反向传播方法对其进行优化，改进了LSTM算法。其性能通过上海超级计算中心的真实工作负载进行了验证。Tuli等人 [34] 提出了一种针对云环境的自动落后任务（处理速度较慢的任务）预测与缓解方法

该方法使用编码器LSTM网络，能够处理异构主机和任务特征的波动性。编码器分析资源使用量和负载信息，并将这些信息传递给LSTM。此外，为防止LSTM模型发散，还将输入矩阵的指数移动平均纳入考虑。文献 [32]介绍了一种名为CrystalLP的、基于LSTM神经网络的存储工作负载预测方法。该方法构建了存储工作负载时间序列模型，用于收集预期的工作负载模式，从而实现精准的自适应调度和负载均衡。随后，实现了基于LSTM的工作负载预测器，并使用由随机梯度下降（SGD）与Adam优化器相结合的算法对其进行训练或优化。Gao 等人 [31] 提出了一种基于多层双向长短期记忆网络（Bi-LSTM）的任务失败预测算法。该算法由一个输入层、两个Bi-LSTM层、一个输出层和逻辑回归（LR）层组成，用于预测任务是失败还是完成。不同于仅使用前向状态的传统LSTM，Bi-LSTM同时处理前向状态和后向状态，从而能够更准确地估计较近和较远输入特征的权重。Ruan 等人 [33] 结合云工作负载特征，建立了用于云服务器工作负载预测的转折点预测模型。此后，构建了一种结合基于规则过滤的分段线性表示（PLR）算法的云特征增强深度学习模型，用于工作负载转折点预测。该模型的性能评估表明，与现有最先进的方法相比，该模型在F1 分数方面有所提升，体现出其预测准确率的有效性。

5.2 自动编码器

本文提出了一种基于高效正则多线性（CP）分解的深度学习模型，用于 [35]预测云中的工业工作负载。该模型通过双射将基本自编码器转换到张量空间，从而构建CP自编码器。在该模型中，CP分解格式的基本自动编码器之后接入了CP分解格式的堆叠自编码器模型。堆叠自编码器用于学习工作负载信息中的相关特征，而CP分解则用于大幅压缩这些特征，以提高训练效率。Chen等人 [36] 建立了一种基于深度学习的云工作负载预测算法（L-PAW），其中包含一种顶部稀疏自编码器（TSA），用于有效提取工作负载的基本表示。该方法结合了门控循环单元（GRU）和循环神经网络（RNN），以消除长期记忆依赖，从而更准确地预测未来的云工作负载。

5.3 深度置信网络

Qiu 等人 [38] 提出了一种深度置信网络（DBN），由多个层级的受限玻尔兹曼机（RBM）和一个用于预测云工作负载的回归层组成。在该模型中，DBN 从所有虚拟机中提取高级特征，并使用回归层预测虚拟机即将到来的负载。它利用先验知识，以无监督的方式高效地学习重要模式。Zhang 等人 [39] 开发了一种基于 DBN 方法的负载预测模型，该模型由堆叠的 RBM 构成，并使用反向传播算法来最小化其损失函数。该模型将方差分析和正交试验设计（OED）技术融入 DBN 的参数学习中，其预测准确率高于 ARIMA。此外， [60] 还提出了一种类似的基于 DBN 的方法，无需手工设计特定特征，即可捕捉云指标数据中的高方差，用于短期资源需求和长期负载预测。Wen 等人 [40]提出了一种基于 DBN 和粒子群优化（PSO）的 CPU 使用率预测算法，名为 DP-CUPA。该算法包括三个主要步骤：预处理训练数据样本；采用自回归和灰色模型作为基础预测模型；以及训练 DBN。在学习过程中，该算法利用 PSO 估算 DBN 参数。

5.4 深度神经网络

Xu 等人提出了一种高效的监督学习型深度神经网络（esDNN）算法 [41]，用于提取和学习历史数据的特征，并准确预测未来工作负载。该多变量数据被转换为监督学习时间序列，并采用经过改进的门控循环单元（GRU），使其能够适应工作负载的变化，从而实现准确预测，并克服梯度消失和梯度爆炸的局限性。文献 [42]提出了一种基于深度神经网络的工作负载预测方法（称为 DNN-MVM），用于处理来自多个虚拟机的工作负载预测。该方法采用预处理和特征选择引擎，直接处理来自这些虚拟机的数据。该模型根据历史负载对数据进行分类，为云服务提供商提供更多信息或知识，以便进行资源管理和优化。它有助于预测未来的资源峰值需求。该模型使用网格工作负载档案（GWA）数据集进行了验证。Bi 等人 [44] 提出了另一种基于深度神经网络的工作负载预测模型，该模型在任务平滑之前执行对数运算，以尽量减小标准差。随后，采用 Savitzky-Golay（S-G）滤波器消除原始数据序列中的极值点和噪声干扰。该模型采用基于深度神经网络的长短期记忆网络（SG-LSTM）来提取任务时间序列的复杂特征。该模型采用梯度裁剪方法实现时间反向传播（BPTT）算法，以消除梯度爆炸问题，并使用 Adam优化器完成模型优化。该模型通过 Google 集群数据集进行评估，以验证其相较于现有方法的有效性。文献 [43]提出了一种基于深度神经网络学习的功率预测引擎（DNN-PPE），其中包含数据采集、数据预处理和预测模块。该方法利用递归自编码器进行短时细粒度预测，以跟踪数据中心内功耗的快速变化。

6 混合模型预测模型

基于混合机器学习的预测模型结合了不同的机器学习算 N 法（例如，

著名贡献者（时间线）	模型/方法/框架	工作流程/策略	数据集	实现/仿真工具	预测的 pa-参数	错误指标	结果或备注
Kumar 等人 [29] (2018年)	长短期记忆网络-RNN	循环中的网络组合保留学习信息，执行特定操作以生成-循环中下一个网络的输出	Web 服务器 HTTP 跟踪记录	MATLAB	No. of 请求每单位时间	MSE	将均方根误差降低至 0.00317
Tang [30](2019年)	二维长短期记忆网络	基于周依赖性对输入数据进行分析并使用权重并行化算法以改进长短期记忆网络的优化	1 工作负载 of the 上海超级计算机	未提及	工作负载数天内	MSE	大规模实时工作负载的准确率有所提高——云服务
Tuli等人 [34] (2021年)	编码器 LSTM	编码器用于准备输入矩阵，并且长短期记忆网络预测负载信息	PlanetLab 跟踪记录	CloudSim and Python	CPU，内存，带宽	MSE，MAPE	缩短了执行时间，减少了源争用和能耗并将 SLA 违规减少了 13%，11%，16% 和 19%
Gao 等人 [31] (2020年)	双向长短期记忆网络	训练数据经由输入层、两个双向长短期记忆网络-长短期记忆网络层、一个输出层和逻辑回归层传播，during 学习过程	55,55,55 任务跟踪记录	Tensorflow in Python	task 故障 rate	F1 分数	93% 的准确率和 87% 的任务故障预测准确
Ruan 等人 [32] (2021年)	CrystalLP	时间序列模型收集负载模式和长短期记忆网络使用随机梯度下降+Adam优化器训练的负载预测模型	网络搜索档案 SPC 跟踪记录	Keras 库，Python	请求大小	平均绝对百分比误差，均方根误差，MAE	实现了 1.10% 改进——MAPE方面的改进，以及更好的MAE性能，优于现有方法
Ruan 等人 [33] (2022年)	FEMT-LSTM	一种考虑云工作负载特征，随后是特征增强深度学习模型	Google 集群、阿里巴巴、HPC Grid 工作负载	Keras 库，Python	CPU 使用率	二进制交叉-熵，F1，精确率，召回率	F1 分数提升了 6.6% 优于现有方法
Zhang 等人 [35] (2018年)	CP 自动-编码器	典范多项式分解压缩了特征，堆叠式自动编码器学习模式-用于预测的模式	PlanetLab 跟踪记录	MATLAB	CPU 利用率	平均绝对百分比误差，RMSE	实现了更高的训练效率-效率和预测工业工作负载的准确性负载
Chen等人 [36] (2019年)	L-PAW 自动-编码器	TSA 提取工作负载模式，门控循环单元+循环神经网络预测即将到来的负载	TensorFlow 1.4.0，Python	DUX 集群，阿里巴巴，and Google 集群	CPU，内存，disk，I/O 使用率	MSE，CDF	准确率优于 LSTM、循环神经网络、门控循环单元的准确率
Qiu 等人 [38] (2016年)	深度置信网络+受限玻尔兹曼机	深度置信网络在学习的同时提取显著模式并使用回归进行预测	PlanetLab 跟踪记录	CloudSim	CPU 利用率	MAPE	提升了性能 up 至 1.3% 优于现有方法
Zhang 等人 [39] (2017年)	深度置信网络+正交实验设计	预处理后的数据被输入深度置信网络+正交实验设计通过反向传播算法调优的模型算法	谷歌集群跟踪记录	Python	CPU，RAM 使用率	MSE	均方误差达到以下范围 $[10^{-4}, 10^{-3}]$
Wen 等人 [40] (2020年)	DP-CUPA	深度置信网络预测负载信息，粒子群优化估计——其调优参数的适应度值	谷歌集群跟踪记录	未提及	CPU 使用率	MSE，平均绝对百分比误差，MAE	优于 自回归模型、深度置信网络、灰色模型
Xu 等人 [41] (2022年)	es-DNN	监督学习将多变量数据转换为时间序列，并应用改进的门控循环单元	阿里巴巴和谷歌集群跟踪数据	TensorFlow Python 2.2.0	CPU 使用率每个时间单位间隔	平均绝对百分比误差，MSE，RMSE	减少的活跃数量高效地优化主机成本
Bhagtya等人 [42] (2021年)	DNN-MVM	对来自多个虚拟机的选定数据进行分类、预处理，并将其传递给DNN-MVM进行学习进程	Grid 工作负载档案 (GWA) 跟踪记录	谷歌 Colab 使用 Keras	CPU，内存，and Disk 利用率	MSE	实现了超过 85% 的预测准确率，针对每个再来源
Bi 等人 [44] (2019年)	SG-LSTM	萨维茨基-戈莱滤波器为长短期记忆网络平滑数据，以便进行更准确的预测	谷歌跟踪记录	Python	中央处理器、内存	对数均方根误差，R^2	优于 BPNN、LSTM、SG-LSTM 和 SG-BPNN
Li 等人 [43] (2016年)	DNN-PPE	预处理后的数据被传递给递归自编码器	世界 cup 98 (WC98) 和 Clark 网络跟踪数据	Python	功率、请求数请求	-	79% 错误减少量规范预测

法），并将一种算法的结果输入另一种算法（单向），以构建高效的机器学习模型，实现精确、准确的预测。这些混合模型通过多种方式组合构建，例如“分类 +分类”；“分类 + 聚类”；“聚类 + 聚类”；“聚类 + 分类”；“分类 + 回归”；“聚类 + 回归”等。这些组合会根据目标问题的需求和挑战进行选择，以尽量减少特征噪声，并

偏置、降低方差并提高预测准确率。混合预测模型的示意图如图 6 所示。在该模型中，原始数据和

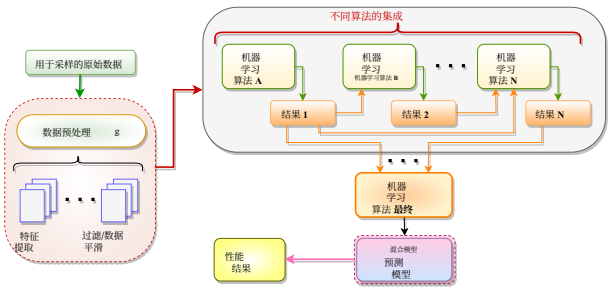


图 6：混合预测模型

复杂数据样本通过提取重要特征、聚合以及缩放或归一化，对数据样本进行过滤和平滑预处理。预处理后的数据样本依次输入一系列 N机器学习算法，并将这些算法的输出作为其他机器学习算法的输入，从而生成用于性能测量和按预期用途部署的混合预测模型。以下讨论基于混合方法的工作负载预测模型的主要重要贡献。

Kardani 等人 [45] 开发了一种基于混合异常感知深度强化学习的资源扩展方法（ADRL），用于云环境中的动态资源扩展。该方法提出了一种基于深度强化Q学习的决策方案所采用的异常检测方法，可识别系统中的异常状态并相应地触发操作。这项工作设置了全局和局部两个层级的决策者，以管理必要的扩展操作。ADRL仅执行必要操作，便改善了服务质量（QoS），并提高了系统稳定性。 [46] 提出了一种名为“BHyPreC”的混合循环神经网络（RNN）预测模型，该模型在堆叠式LSTM和GRU之上加入双向长短期记忆网络（Bi-LSTM），用于预测虚拟机资源

使用量。该模型分别提升了Bi-LSTM、LSTM和GRU模型的非线性数据分析能力，并证实其准确率优于其他统计模型。它采用结合历史窗口大小的网格搜索技术来优化模型，并确定最佳窗口大小组合。Bi 等人 [47] 提出了一种名为“BG-LSTM”的集成深度学习方法，该方法结合双向长短期记忆网络和网格长短期记忆网络，以实现对工作负载和资源时间序列的高质量预测。在预处理过程中，该方法使用 Savitzky-Golay（SG）滤波器，在平滑工作负载之前降低标准差。它能够有效提取相对较长时间序列中的复杂非线性特征，并实现较高的预测准确率。 [61]提出了一种分层毕达哥拉斯模糊深度神经网络（HPFDNN），用于预测所需云资源的数量。原始采样数据的神经表示作为一种补充方法，用于清晰地解释真实结果，这超出了模糊逻辑的应用范围。用户可以利用深度神经网络的预测结果确定预期的云服务数量，从而有助于降低成本。

本文提出了一种基于 MAPE-k 控制回路的混合式自主资源配置模型，用于多层应用 [28]。该模型结合了模糊层次分析法（FAHP）。实验结果表明，与其他方法相比，所提出的方案在分配的虚拟机数量、响应时间和成本方面表现更优。文献 [49]提出了一种自适应分类预测方案（ACPS），首先根据工作负载特征将工作负载划分为不同类别，并自动为各类别分配不同的预测模型。此外，通过建立混合 0–1 整数规划模型，将工作负载分类问题转化为任务分配问题，并利用改进的分支定界算法快速求解。Peng等人[37]采用一种名为“GRUED”的基于GRU的编码器-解码器网络，其中包含两个门控循环神经网络（GRNNs），以解决这些问题。研究通过实验评估了该网络在云计算中对主机多步超前工作负载的预测效果。Shuvo等人 [50]提出了一种新颖的混合方法，名为“LSRU”，用于

提高预测准确率。LSRU 将 LSTM 与 GRU 结合起来，既用于短期预测，也用于工作负载突发情况下的长期预测。

7 集成预测模型

集成方法使用多个“基础预测”（BP）模型或“专家”来预测事件预期的未来结果。集成模型的最终结果是通过投票引擎组合各位专家的预测结果计算得出的。基于集成的预测方法的概念架构如图 7 所示。历史数据和实时数据样本经过预处理后，构建一个“滑动窗口”，由此生成的输入数据向量将输入所有基础预测器 $\{BP_1, BP_2,...、 BP_z\}$，其中 z 是用于集成学习的基础预测器数量。每个基础预测器的估计结果都会被赋予一个权重值，以表示其在最终预测结果中的重要性。

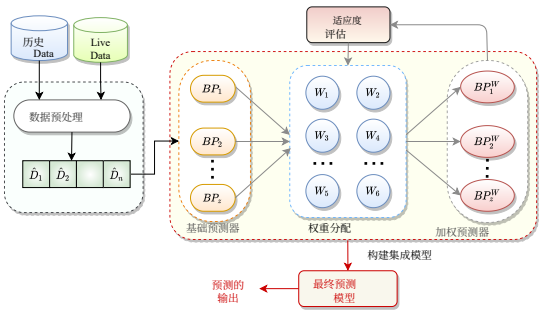


图 7：集成预测模型

权重分配及其更新需要通过学习过程实现，并采用合适的优化方法，例如多分类回归、基于优先级的方法或进化学习算法。与存在高方差、准确率低、特征噪声和偏置等问题的单独预测方法相比，集成学习方法更为有效。基本上，在集成方法中，各种机器学习模型彼此独立地进行预测，随后由基于各基础预测器对应权重值的投票系统（硬投票或软投票）确定最终预测结果。以下将讨论基于集成学习的主要模型。

Singh 等人 [51] 通过扩展和改进两种在线集成学习方法——加权多数（WM）和可模拟专家（SE），解决了广泛工作负载的预测问题。他们将经典的 SE 从二元结果空间扩展到 k 元结果空间，使其适用于解决任意 k 类问题（称为“kSE”）。加权多数集成模型的参数以增量方式重新生成，使这些算法的计算效率更高，更适合处理海量在线数据流（称为“WMC”）。研究使用包含 1570 台服务器的大型数据集对这些模型进行了评估，并验证了采用扩展版算法后，约 91% 的服务器能够被正确预测。Feng 等人 [52] 提出了一种用于工作负载预测的集成模型，该模型结合了自适应滑动窗口和时间局部性整合，命名为 FAST。该模型开发了一种自适应滑动窗口算法，综合考虑趋势与时间的相关性以及未来工作负载的随机波动，以较低的开销最大限度地提高预测准确率。此外，该模型还引入了时间局部性概念，用于基于误差的集成策略中的局部预测器行为。通过开发一种多类回归加权算法，构建了完整的集成模型。研究使用 Google集群跟踪数据集验证了该模型的性能。文献 [19]提出了一种用于工作负载时间预测的集成模型，该模型将 Savitzky–Golay（SG）滤波器和小波分解与随机配置网络相结合，以预测未来工作负载。在该模型中，任务时间序列先经 SG 滤波器平滑，再通过小波分解方法分解为多个分量。该模型名为“SGW-S”，能够刻画趋势分量和细节分量的统计特征，并通过更快的学习实现了性能提升。

一种名为“CloudIn-

著名 贡献者 （时间线）	模型/ 方法/ 框架	工作流程/策略	数据集	实现/ 仿真工具	预测的 pa- 参数	误差指标	结果或备注
Kardani 等人 [45] (2021年)	ADRL: Deep 强化学习+Q-学习	基于深度Q学习的强化学习模型，以应对中央处理器和内存瓶颈问题，有助于重新进行资源缩放和决策	基于Web的 Rice 大学竞价系统（RUBiS）	Python with 基于 Java CloudSim	CPU，内存，响应-time	MSE	提升服务质量和稳-定性
Karim 等人 [46] (2021年)	BiHyPrec: 双向长短期记忆网络 + 长短期记忆网络 + GRU	将预处理后的数据输入协同模型 由双向长短期记忆网络、长短期记忆网络和门控循环单元构成，能够实现一种基于深度学习的方法，以有效应对的复杂性和非线性 时间序列数据	Bitbrains 跟踪数据	Python and 谷歌 Colaboratory	中央处理器使用率	MSE，平均绝对百分比误差，平均绝对误差、均方根误差	表现 优于 ARIMA、长短期记忆网络、门控循环单元、双向长短期 记忆网络
Bi 等人 [47] (2021年)	BG-LSTM: 双向长短期记忆网络 + 网格长短期记忆网络	使用 Savitzky-Golay（SG）对数据进行平滑处理，被传递给双向长短期记忆网络和 GridL-STM 混合模型进行预测	Python	Google 集群跟踪记录	CPU and 随机存取存储器使用率	MSE，均方根对数误差，R^2	更高的准确率 over SG-LSTM、SG-Bi-LSTM，SG-GridLSTM
Chen等人 [61]（2021年）	HPFDNN: 层级的毕达哥拉斯 + 模糊 DNN	毕达哥拉斯模糊逻辑、神经表示，并集成深度神经网络，并采用网络训练方法，结合自适应学习率 以最大限度降低云服务的成本，用户	卡内基 梅隆大学数据集	未提及	请求数求	准确率 cost 和总成本	节省了 202.48 美元 实时请求 现有方法
Khorsand 等人 [28] (2018年)	FAHP	针对工作负载开发了 FAHP 和支持向量回归算法-负载预测、资源配置以及 适当的自动扩缩容决策	Clark Net、美国国家航空航天局、合成工作负载	CloudSim and Open 来源 RUBIS	请求到达，响应 time, cost	MSE，归一化均方误差，均方根误差，均方分段根误差	降低租赁 云服务提供商的资源 质量保证
Liu 等人 [49]（2017年）	ACPS	不同的预测模型会自动进行 根据工作负载特征进行签名	谷歌 集群跟踪记录	Python	响应 时间和成本	Mean 错误，mean 相对预测误差（平均相对预测误差）	预测误差为 re-降低了 40.86% 线性回归
Peng等人 [37]（2018年）	GRUED	GRU 编码器将可变长度的工作负载序列 将序列转换为定长向量，而 GRU 解码器将向量表示映射回 可变长度的未来工作负载序列	Google 集群和 Dinda 工作负载	Python, UNIX 系统	中央处理器、作业自回归模型	MAE，平均绝对百分比误差，均方根误差，root 平均分段平方 错误（均方分段根误差）	降低预测误差 在长短期记忆网络上
Shuvo等人 [50]（2020年）	LSRU: LSTM + GRU	一些统计方法，如自回归、指数平滑、自回归移动平均，以及 ARIMA 用于预测，并输入长短期记忆网络 并与门控循环单元组合，以提高预测准确率	Bitbrains	Kaggle	CPU，磁盘，内存，带宽	MAE，平均绝对百分比误差，均方根误差、均方误差	降低预测误差 优于长短期记忆网络和门控循环单元

sight”（ClIn）的云资源预测模型，基于集成预测方法，在 [53] 中提出。该模型通过训练基于统计特征的分类器，精准估算作业到达情况，并仔细找出最适用的机器学习预测方法。它采用多个局部预测器或专家，并动态确定各局部预测器的重要权重（或贡献），据此构建集成预测模型。借助支持向量机分类器进行多类回归，定期优化自适应权重分数，从而在相应的预测区间内选择准确率最高的最合适预测模型。该模型使用三类不同的工作负载进行测试，包括 Web、集群和高性能计算工作负载，以证明其相较于现有预测模型的有效性。Baig 等人 [54] 提出了一种名为“自适应模型选择器”（AMS）的集成式自适应云资源估算模型。该模型通过训练基于统计特征的分类器，精准估算资源利用率，并确定最适用的机器学习预测方法。它使用随机决策森林（RDF）构建分类器，通过定期训练和重新训练，预测给定滑动窗口数据的最佳模型。选定的特征和识别出的预测方法会被记录为训练数据，并传递至滑动窗口以进行学习。该模型的性能使用 Google 集群、阿里巴巴和 Bitbrains 数据集进行评估。Kumar 等人 [55]提出了一种名为“E-ELM”的基于集成学习的工作负载预测模型，该模型采用极限学习机，并通过投票引擎为其相关预测结果赋予权重。在这项工作中，研究者应用一种受黑洞理论启发的元启发式算法，以确定最优权重。该模型的准确率通过 Google 集群跟踪数据中的中央处理器和内存需求请求进行评估，并针对 PlanetLab 虚拟机跟踪数据中的 CPU 利用率，Chen等人[48] 提出了一种自适应预测方法，该方法采用集成模型和基于减法模糊聚类的模糊神经网络（ESFCFNN）。该方法对用户偏好和需求进行刻画，并使用若干基础预测器构建集成预测模型。

8 基于量子神经网络的预测模型

基于量子神经网络（QNN）的预测模型是一种智能预测模型，它兼具神经网络的机器学习能力和量子力学的计算能力，从而实现高精度的预测准确率。它本质上是一种神经网络，其中使用量子比特神经元和量子比特权重，而非实数值，训练数据的信息也以量子比特的形式传播。量子比特可表示为 1、0，或这些状态的任意量子叠加。数学上，可表示为：$|\Psi> = \alpha|0> + \beta|1>$，其中 α 和 β 是复数，分别表示状态 $|0>$ 和 $|1>$ 的概率振幅。基于量子神经网络的预测模型示意图如图 8 所示，该模型由一个输入层、多个隐藏层和一个输出层组成，分别包含 n、p 和 q 个量子比特节点，对应于 n-p-q 量子比特网络架构。不同层的量子比特神经元之间的连接权重也采用量子比特形式，并在学习过程中进行调整。量子比特神经元的状态转换由各种量子门实现，例如旋转门、受控非门等。训练数据样本的取值经过提取和聚合，形成特定时间间隔（例如 5 分钟）的数据，并使用归一化函数将其缩放至特定范围。归一化后的数据值

著名贡献者（时间线）	模型/方法/框架	工作流程/策略	数据集	实现/模拟tool	预测的 pa-参数	误差指标	结果或备注
Singh 等人 [51] (2014年)	kSE+WMC	SE 被扩展到 k元结果空间，WM 则得到改进，以实现增量式且计算高效的学习过程	数据集 G(1570)	未提及	服务器工作负load	MAPE	89% 准确的预测与……相比基准算法的 13–24% 算法
Feng 等人 [52]（2022年）	FAST	自适应滑动窗口考虑所有类型的具有时间局部性概念的工作负载趋势，用于基于错误的集成，以提升预测准确率	谷歌 集群跟踪记录	Python	中央处理器、内存	绝对误差，均方根误差，R^2	改进 预测准确率 by 14.99% 达到 27.55，RMSE 为达到 22.57% 至76.86%，针对 R^2
Bi 等人 [19]（2019年）	SGW-S	SG 滤波器与小波分解相结合与随机配置网络结合以预测高准确率的工作负载	谷歌 集群跟踪记录	未提及	task 到达 rate	均方误差，R^2	使用 Savitzky–Golay 滤波器小波分解有助于提高预测准确率
Kim 等人 [53]（2020年）	ClIn: 集成模型 with 多类回归	不同的局部预测器构成一个集成模型预测模型，通过动态确定并更新每个预测器的重要权重	Web、集群、HPC 网格工作负载	Python	单位时间内的作业到达单位时间	归一化均方根误差，绝对误差	Up to 15%–20% less 不足/过度配置，兼具高成本效益和低 SLA 违规
Baig 等人 [54]（2019年）	AMS	使用……预测即将到来的资源需求。RDF 分类器选出的最佳预测模型预测即将到来的资源需求，该分类器会随时间定期更新	阿里巴巴，Bitbrains，Google 集群	Python	CPU	均方根误差，平均绝对误差	改进的预测准确准确率 from 6%-27% 优于当前方法——论
Kumar 等人 [55] (2020年)	E-ELM	基于极限学习机的局部预测器经过训练和选择采用一种权重更新方法，并借助元启发式算法	MATLAB	Google 集群和 PlanetLab 跟踪记录	中央处理器、内存	均方根误差、平均绝对误差	均方根误差最多降低 99.20% 优于现有方法
Chen等人 [48]（2015年）	减法式模糊聚类	用户偏好和需求得到表征到集成预测模型减法基于模糊聚类的模糊神经网络	数据流统计跟踪记录	CloudSim	网络资源跟踪记录	均方误差、平均绝对误差	有效的资源去-求预测

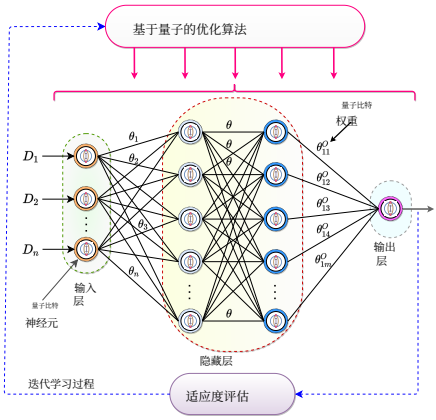


图 8：量子神经网络

通过应用公式 (10) 和 (11) 中量子比特旋转的作用，转换为量子态值或量子比特；

$$y^{In}{}_i = f(\Theta_i{}^{In}) \tag{10}$$

$$\Theta_i = \frac{\pi}{2} \times \mathcal{D}_i \tag{11}$$

其中，$\mathcal{D}_i$ 是 i^{th} 输入数据点，Θ_i 是网络的 i^{th} 量子输入点。QNN 模型从实际工作负载中提取相关且重要的模式，并分析 n先前的工作负载值，以估算数据中心下一（$n+1$）th 个时刻的工作负载信息。QNN 的学习过程由一种基于量子比特的优化算法实现，该算法能够操控、探索和利用量子比特来重新生成种群，并防止出现停滞问题。使用 QNN 的工作负载预测方法仍处于起步阶段。

迄今为止，唯一一项采用基于 QNN 的模型进行工作负载预测的研究见 [56]。该模型将工作负载信息编码到量子比特中，并通过网络传递这些信息，以估算工作负载或资源需求，从而主动提高预测准确率；它利用了量子计算的计算效率。

C-NOT 门的旋转和反向旋转效应在隐藏层和输出层中充当激活函数，以优化量子比特权重。研究人员开发了自平衡自适应差分进化（SB-ADE）算法，用于优化量子比特网络权重。该模型使用三类不同的工作负载进行评估，其预测准确率显著优于现有方法。 [62]介绍了一种使用复数的工作负载预测模型，该模型将强大的学习能力和更高的准确率应用于采用多值神经元（MLMVN）的多层神经网络预测模型，从而缩短了用时。

9 性能评估

9.1 实验设置

仿真实验在一台服务器上进行，该服务器配备两颗 Intel® Xeon® Silver 4114 中央处理器，拥有 40 个处理器核心，时钟频率为 2.20 GHz。计算机安装了 64 位 Ubuntu 16.04 LTS，配备 128 GB 随机存取存储器。所有选定的、基于进化神经网络（ENN）、集成学习（EL）、混合学习（HL）、深度学习（DL）和进化量子学习的最佳预测方法，均使用 Python 3.7 实现；各项预设参数及其取值详见表 6。

9.2 数据集

使用三个不同的基准数据集，对多种基于机器学习的预测模型进行性能分析和比较。这些数据集包括 Google 集群数据（GCD）中的 CPU 和内存使用跟踪数据 [63]，以及 PlanetLab（PL）虚拟机跟踪数据中的 CPU 使用率[64]。GCD 工作负载提供了集群云应用的行为数据，以及 Hadoop 等大数据分析应用的信息，其中包含对资源（CPU、内存和磁盘输入/输出）的请求与使用信息；这些数据记录了在 29 天内于 12,500 台服务器上执行的 672,300 个作业。每个作业每隔五分钟提供一次 CPU 和内存使用率百分比，据此可得出虚拟机的 CPU 和内存利用率百分比，覆盖的时间段为

预测模型	参数	取值
Evolutionary Neural Network	输入节点	10
	隐藏节点	7
	最大迭代次数	250
	训练数据规模	70%
	变异学习周期	10
	交叉学习周期	10
	种群规模	15
	训练算法	差分进化
Ensemble Learning	#极限学习机网络	[10,100]
	输入节点	[7,70]
	隐藏节点	[5,50]
	种群规模	20
	最大迭代次数	100
	准确率阈值	0.007
	训练数据规模	70%
	训练算法	黑洞优化
Hybrid Learning	输入节点	100
	训练轮数	20-50
	批次大小	16
	激活函数	tan$h$
	训练数据	70%
	训练轮数	500-1000
	批次大小	1-4
	训练算法	Adam优化器
Deep Learning	深度学习库：TensorFlow	0.12.1
	Keras	1.2
	训练数据	70%
	训练轮数	500-1000
	神经元	4-10
	批次大小	1-4
	训练算法	梯度下降
Quantum Learning	输入节点	10
	隐藏节点	7
	输出节点	1
	训练轮数	50
	训练数据	70%
	种群规模	15
	变异学习周期	5
	交叉学习周期	5
	训练算法	量子差分进化

二十四小时。PL 数据包含超过 11,000 台虚拟机的 CPU 利用率数据，这些数据在 2011年3月至4月的十个随机日期内每隔五分钟测量一次。根据不同的预测间隔（如 5、10、20、.....、60 分钟），提取并汇总相应的资源使用量取值。使用公式（1）所述的归一化公式，将这些取值重新缩放至范围 [0, 1] 。表7展示了评估工作负载的统计特征。

表7：评估工作负载的特征

工作负载	持续时间	Jobs	均值（%）	标准差
GCD-中央处理器 (G^C)	10 天	2 M	21.84	13.62
GCD-内存 (G^M)	10 天	2 M	19.55	16.6
PL-CPU (P^C)	10 天	1.5 M	19.77	14.55

9.3 评估指标

使用以下误差指标评估预测模型的预测准确率：

均方误差 ($\mathcal{MSE}$)：它是衡量预测模型准确率的知名指标之一，它会对较大的误差项施加较高的惩罚。模型的得分越接近零，就被认为越准确。该指标的数学表达式见公式（12），其中 m 是工作负载跟踪数据中的数据点数量，$\mathcal{Z}^{Ac}(t)$ 和 $\mathcal{Z}^{Pr}(t)$ 分别是在 t^{th} 时刻的实际工作负载值和预测工作负载值。

$$\mathcal{MSE} = \frac{1}{m}\sum_{t=1}^{m}(\mathcal{Z}^{Ac}(t) - \mathcal{Z}^{Pr}(t))^2 \tag{12}$$

平均绝对误差（$\mathcal{MAE}$）：在均方误差中，较高错误取值的平方可能会获得更高的权重，从而影响预测的准确率。而 $\mathcal{MAE}$对每个错误分量赋予相同的权重，并通过计算实际工作负载（$\mathcal{Z}^{Ac}(t)$）与预测的工作负载（$\mathcal{Z}^{Pr}(t)$）在 t^{th} 时间点之间绝对差值的平均值来衡量预测模型的准确率，如公式（13）所示。该指标会产生一个非负数，用于评估预测结果的准确率；如果其接近于零，则预测结果与实际取值非常相似。

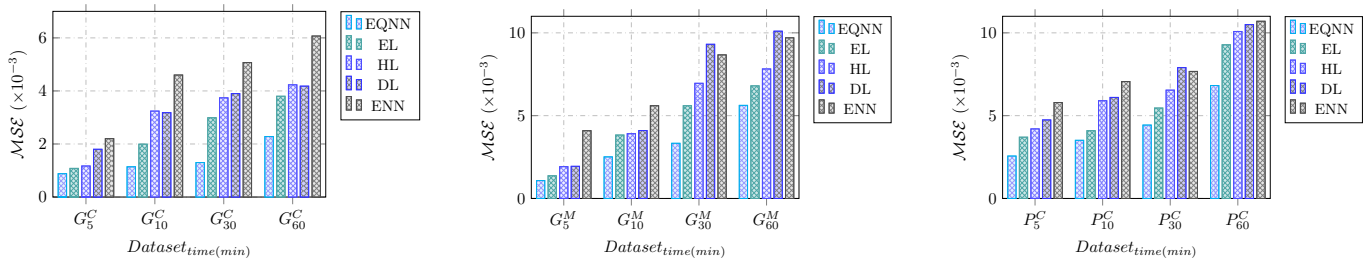
$$\mathcal{MAE} = \frac{1}{m}\sum_{t=1}^{m}|\mathcal{Z}^{Ac}(t) - \mathcal{Z}^{Pr}(t)| \tag{13}$$

9.4 结果

我们对不同预测模型的性能进行了全面研究和比较，包括进化量子神经网络（EQNN） [56],集成学习（EL） [55],混合学习 [50],深度学习（DL） [29], 进化神经网络（ENN）[25]; 研究采用了种类繁多的异构云应用以及虚拟机不断变化的资源利用率。我们评估并比较了不同类型的基于学习的模型在 $\mathcal{MSE}$方面的表现，使用的置信度指标包括 $\mathcal{MAE}$、绝对误差频率（AEF）和训练耗时（TT）。

9.4.1 均方误差

图 9 比较了三类工作负载的均值，包括 GCD-CPU 跟踪数据（图 9a）、GCD 内存跟踪数据（图 9b）和 PL-CPU 跟踪数据（图 9c）。从图中可以看出，$\mathcal{MSE}$ 随预测模型的不同而变化，并且随着预测区间大小从 5 分钟增加到 60 分钟而增大。这是因为随着预测窗口大小增加，可用的训练数据样本数量会减少。随着平均 $\mathcal{MSE}$ 值降低，预测准确率呈现如下趋势：EQNN $<$ EL $\leq$ HL $<$ DL $<$ ENN。此外还可以发现，较短预测区间下的预测误差（$\mathcal{MSE}$）差异较小；随着预测窗口大小增加，训练数据样本数量减少，这种差异也随之增大。因此可以得出结论，对于短期预测，所有类型的预测模型都能达到预期的预测准确率，而性能方面的主要差异出现在长期预测区间。实验结果表明，基于量子学习的预测模型（EQNN）在大多数预测区间和大多数跟踪数据中产生的预测误差最小，而基于进化学习的预测模型（即 ENN）在大多数实验情况下产生的 $\mathcal{MSE}$ 值最高。对于



(a) GCD-CPU 跟踪数据 (b) GCD-Memory 跟踪数据 (c) PlanetLab CPU 跟踪数据

图 9：均方误差

GCD-CPU（图 9a）而言，EQNN 的表现始终优于所有其他方法，这是因为它采用了量子比特和量子叠加效应，能够在学习过程中精确、直观地学习相关性和相关模式。此外，EL 的 $\mathcal{MSE}$ 值低于 HL、DL 和 ENN，这是因为它包含多个基础预测模型，能够自适应地选择每次表现最佳的预测模型，并通过更新相应基础预测器的权重来重新激活学习过程。HL 的预测误差小于 DL，这是因为它在实际预测之前整合了过滤和平滑方法（利用 Savitzky-Golay 滤波器等多种过滤方法，或使用 GRU 提高准确率，同时减轻 LSTM 的缺点）。在 HL 中，两种或更多方法通过协同配合来减轻各自的局限性，从而生成有效的预测模型，以预测各种云工作负载的资源使用量。DL 实验采用基于 LSTM 的预测方法，在包括 G_5^C 至 G_{60}^C 在内的所有情况下，其表现均优于 ENN。图 9b 所示的结果图表明，GCD-Memory 跟踪数据呈现出类似趋势：HL 和 EL 的表现较为接近，即不及 EQNN，但优于基于 DL 和 ENN 的预测模型。基于 DL 的预测模型所得结果表明，除 G_{30}^M 和 G_{60}^M 的情况外，在所有相应实验中，其 $\mathcal{MSE}$ 值均低于 ENN，准确率有所提高。图 9c 展示了 PL CPU 跟踪数据中 $\mathcal{MSE}$ 的比较，其中预测误差值之间的差异虽小但显著，支持了上述性能趋势。此外，还针对所得 $\mathcal{MSE}$ 结果计算了置信度指标，如表 8 所示，其中报告了所有实验情况下的误差范围（EM）和置信区间（CI）。

9.4.2 平均绝对误差

图 10 展示了不同预测模型在各种数据集上获得的 $\mathcal{MAE}$ 对比结果。与 $\mathcal{MSE}$类似，$\mathcal{MAE}$ 值在所有三种评估工作负载中都呈现出相同的性能变化趋势，包括 GCD-CPU（图 10a）、GCD-Memory（图 10b）和 PL-CPU（图 10c）。如上述三个连续柱状图所示，$\mathcal{MAE}$值依次递减：EQNN $<$ EL $\leq$ HL $<$ DL$<$ ENN。此外，观察发现差异在随着预测窗口大小增大，预测误差（$\mathcal{MAE}$）逐渐增大，这是由于训练数据样本数量减少所致。因此可以得出结论，对于短期预测，各类预测模型都能达到预期的预测准确率，而预测性能的主要差异出现在长期预测区间。其原因在于，随着训练样本数量减少，模型无法充分学习模式，相应的预测模型开始低估训练数据集中的相关信息，导致建立必要相关性的能力下降，并因模式学习能力减弱而出现性能下降。

9.4.3 训练时间

图 11 展示了在预测窗口大小为 30 分钟时，各种预测模型在不同工作负载跟踪记录上的学习过程中所耗费训练时间的比较。基于 HL 和 DL 的模型训练耗时相近，且在所有预测模型中最短，这是因为它们采用了基于梯度下降和 Adam优化器的训练算法。相比之下，ENN、EL 和 EQNN 的训练耗时更长，原因是学习过程中采用了进化优化。EL 的训练时间长于 EQNN 和 ENN，这是因为学习过程中同时使用了多个基础预测模型。与此不同，EQNN 和 ENN 的学习过程采用基于单个网络的预测模型，其中 EQNN 比 ENN 耗时更长。这是因为其采用基于量子比特和量子力学的网络权重优化过程，该过程涉及高度复杂的复数运算，用于生成和更新基于量子比特的网络权重。综上，训练时间的趋势为：DL $<$HL$<$ ENN$<$ EQNN $<$ EL。不过，各种预测模型的效率和适用性并不会因此受到影响，因为训练是一项周期性任务，可以在配备充足资源的服务器上并行执行。

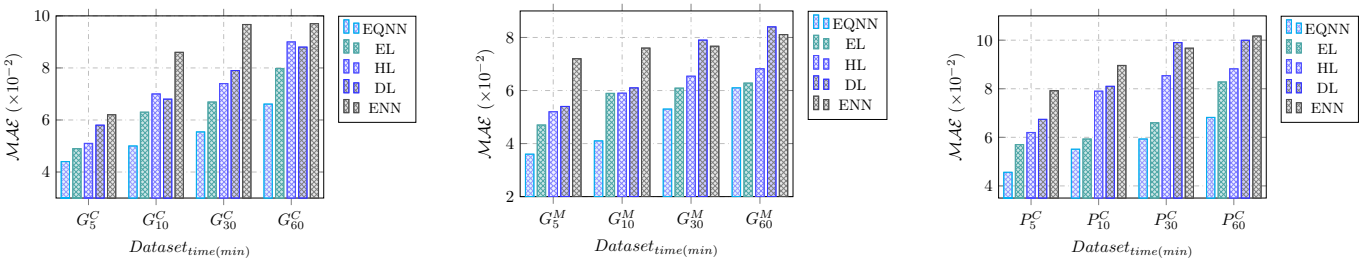
9.4.4 绝对误差频率

通过评估绝对预测误差并比较其在全部三种

表 8：均方误差的置信度指标

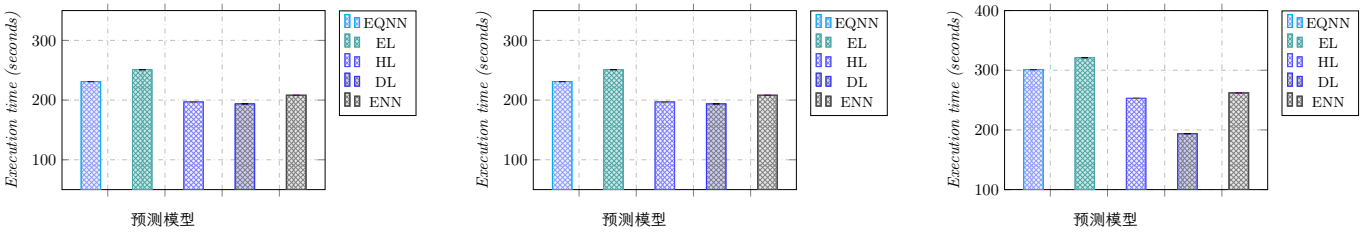
数据集	PWS[a]	指标	EQNN	EL	HL	DL	ENN
GCD-CPU	5	EM	5.3012E-06	1.3824E-06	2.3666E-06	1.8460E-06	2.7320E-06
		CI	8.799E-03 - 8.801E-03	1.079E-03 - 1.080E-03	1.169E-03 - 1.170E-03	1.799E-03 - 1.800E-03	2.199E-03 - 2.200E-03
	10	EM	1.7969E-06	3.6608E-06	5.4890E-06	4.5940E-06	3.2681E-06
		CI	1.398E-03 - 1.140E-03	1.992E-03 - 1.993E-03	3.239E-03 - 3.241E-03	3.895E-03 - 3.900E-03	4.599E-03 - 4.600E-03
	30	EM	2.1890E-06	1.4865E-06	6.8200E-05	4.9826E-06	6.8210E-06
		CI	1.298E-03 - 1.299E-03	2.989E-03 - 2.990E-03	3.733E-03 - 3.747E-03	4.179E-03 - 4.180E-03	5.069E-03 - 5.071E-03
	60	EM	1.1743E-06	1.702E-06	6.476E-06	5.8040E-06	8.2199E-06
		CI	2.280E-03 - 2.282E-03	3.798E-03 - 3.801E-03	4.229E-03 - 4.231E-03	4.179E-03 - 4 ,181E-03	2.39E-05
GCD-Memory	5	EM	5.7890E-05	6.7740E-06	1.7841E-06	8.7341E-06	1.6839E-05
		CI	1.079E-03 - 1.081E-03	1.369E-03 - 1.370E-03	1.919E-03 - 1.920E-03	1.9391E-03 - 1.941E-03	4.090E-03 - 4.093E-03
	10	EM	2.0799E-05	4.0514E-06	4.9616E-06	6.6035E-06	2.5002E-05
		CI	2.507E-03 - 2.512E-03	3.829E-03 - 3.830E-03	3.909E-03 - 3.910E-03	4.099E-03 - 4.101E-03	5.597E-03 - 5.603E-03
	30	EM	6.8200E-05	6.4132E-05	7.1294E-05	5.813E-05	2.5785E-05
		CI	3.329E-03 - 3.331E-03	5.593E-03 - 5.606E-03	6.953E-03 - 6.954E-03	9.307E-03 - 9.310E-03	8.667E-03 - 8.673E-03
	60	EM	4.8290E-05	9.9990E-05	1.5458E-05	2.4786E-06	3.5760E-05
		CI	5.615E-03 - 5.625E-03	6.799E-03 - 6.801E-03	7.818E-03 - 7.822E-03	10.0991E-03 - 10.0996E-03	9.696E-03 - 9.703E-03
PlanetLab CPU	5	EM	4.2810E-06	4.6890E-06	4.6170E-06	6.8910E-06	2.3980E-06
		CI	2.559E-03 - 2.560E-03	3.699E-03 - 3.700E-03	4.199E-03 - 4.200E-03	4.739E-03 - 4.740E-03	5.794E-03 - 5.795E-03
	10	EM	8.4210E-06	9.8140E-06	7.1100E-06	8.7642E-06	1.0890E-05
		CI	3.499E-03 - 3.510E-03	4.092E-03 - 4.093E-03	5.899E-03 - 5.900E-03	6.099E-03 - 6.101E-03	7.058E-03 - 7.061E-03
	30	EM	6.2814E-05	8.6411E-05	1.8230E-05	6.4210E-06	8.6810E-06
		CI	4.423E-03 - 4.436E-03	5.451E-03 - 5.468E-03	6.538E-03 - 6.541E-03	7.899E-03 - 7.900E-03	7.669E-03 - 7.670E-03
	60	EM	8.4210E-05	9.9610E-05	6.4280E-05	2.8136E-05	9.8134E-05
		CI	6.811E-03 - 6.824E-03	9.270E-03 - 9.289E-03	10.075E-03 - 10.088E-03	10.449E-03 - 10.450E-03	10.690E-03 - 10.709E-03

[a] PWS：预测窗口大小，EM：误差范围，CI：置信区间



(a) GCD-CPU 跟踪数据　(b) GCD-Memory 跟踪数据　(c) PlanetLab CPU 跟踪数据

图 10：平均绝对误差



（a）GCD-CPU 跟踪数据　（b）GCD-Memory 跟踪数据　（c）PlanetLab CPU 跟踪数据

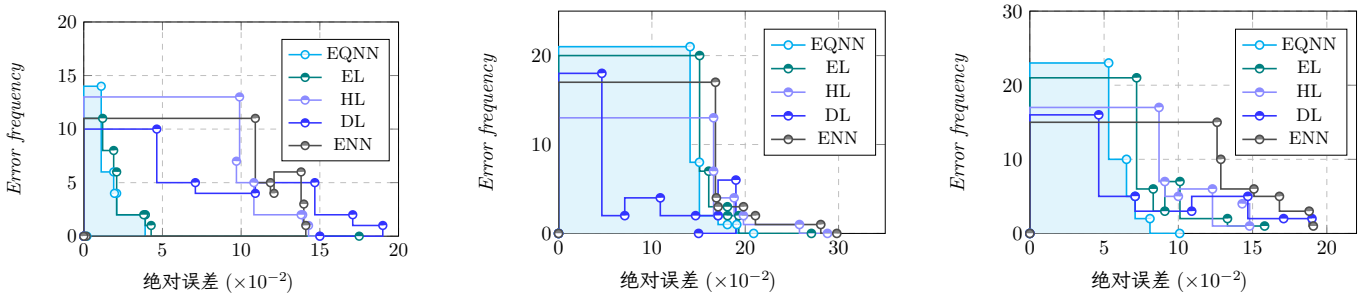
图 11：训练时间开销

工作负载中的频率，对各比较模型取得的预测误差进行了测量和分析。图 12 对比了绝对误差（实际值-预测值）的频率。与其他四种类型的预测模型相比，EQNN 在大多数云工作负载中产生的误差最小。预测模型的绝对误差频率较高，表明其能够持续有效地发挥作用，并稳定地实现较高的预测准确率。GCD-CPU 跟踪数据（图 12a）、GCD-内存（图 12b）和 PL-CPU 跟踪数据（图 12c）所观察到的绝对误差频率呈现出相同趋势：ENN ＜ DL ＜ HL ＜ EL ＜ EQNN。出现这一趋势的原因是，EQNN 使用量子比特种群并结合进化优化，从而更直观地学习模式，进而有效学习范围广泛的动态工作负载

具有最佳准确率的模式。集成学习优于 EQNN，这是因为它利用了多个基础预测模型的学习能力，能够从不同类型的工作负载中准确学习相关信息。其他预测模型的预测错误频率也略低，且处于可接受范围，其变化取决于学习过程中采用的学习能力和优化算法。

9.5 权衡与讨论

所有机器学习算法在针对广泛工作负载进行自适应预测时都存在一定的权衡。同样，各种基于 ENNs 的预测方法之间的关键区别在于学习过程中采用的进化优化



(a) GCD-CPU 跟踪数据 (b) GCD-Memory 跟踪数据 (c) PlanetLab CPU 跟踪数据

图 12：绝对误差频率

算法，这会直接影响预测方法的性能。不同的进化优化算法在种群更新过程和控制参数调优过程中所采用的探索与开发方法各不相同。与需要调优更多超参数的方法相比，需要调优的超参数数量较少的进化优化方法更可取。例如，黑洞学习算法是一种无需参数的算法，其时间和空间复杂度较低，预测准确率也高于需要调优超参数（包括交叉率、变异率、学习率等）的差分进化算法。尽管与基于进化学习的前馈神经网络模型相比，深度学习方法能更快地学习数据样本的自然变化，但它需要更多数据样本进行训练，才能准确估计输出。此外，深度学习方法的计算复杂度较高，需要昂贵的图形处理器和高性能计算设备，从而增加其应用成本。研究发现，深度学习算法与其他机器学习方法（例如用于特征提取的随机森林）结合时表现更好，并通过与其他分类方法协作来生成预测输出。另一方面，混合与集成学习方法需要联合运行多种机器学习算法，与单一机器学习方法相比，会消耗更多的空间和时间。毫无疑问，由于将多种机器学习方法整合到同一平台上，与深度学习和进化学习方法相比，这些方法能够更高效地适应未见数据和广泛的工作负载。在混合与集成学习方法中，集成方法的适应性更强，因为它会考虑所有所选机器学习算法的预测输出，并在生成最终输出时，通过权重优化选择与这些学习算法相关的预测结果。量子神经网络学习方法在所有讨论的方法中效率最高，实验结果验证了它能够准确预测变化的工作负载。在量子神经网络中，源自复数的量子比特具有更高的多样性，相较于实数网络权重值，能够生成更直观的模式，学习复杂关系，并有助于以更高的准确率预测输出。

最后，也是最重要的是，在确定所接近工作负载的异构性时，IaaS、PaaS、SaaS、FaaS 等云服务的多样性具有显著影响，而云服务提供商必须提供无缝的资源质量和资源容量。由于所有云服务模型的资源需求都在动态变化，因此并不存在针对特定云服务选择最佳模型的完美指南。此外，云工作负载还会因多种特征而变化，包括资源容量（即中央处理器、内存、带宽等）的利用率；诸如执行期限等优先级约束；执行成本；在一段时间内作业请求数量的变化程度等。然而，这些特征（除一段时间内作业请求数量的变化程度外）不会对不同预测模型的学习能力产生显著影响，因为相应的数据样本会以周期性方式生成，作为每类工作负载的事件记录，在一个时间戳内报告所有相关信息。这些数据样本用于训练预测模型；对于同一时间段内的估计工作负载，这些模型表现出不同的准确率和训练计算成本。然而，可以根据约束条件的优先级来选择预测模型；例如，为了实现理想的高可用性和期限敏感性，应选择高准确率预测模型，而不论其计算成本和训练时间较高。另一方面，若存在最低执行成本的约束，则计算和训练成本较低的预测模型是降低相应工作负载处理成本的更佳选择。此外，还可根据工作负载的多样性来选择工作负载预测模型，例如高度随机、周期性变化、随机变化、均匀或非均匀多样化等。基于上述对所考虑机器学习模型的特征、设计和能力的讨论，可以推断，EQNN、混合模型和集成模型更适用于高度随机且多样化的工作负载，因为它们更能够学习和处理高度可变且异构的流量数据模式，这类模式涉及大量动态数据、多个具有复杂关系的变量，以及多步时间序列流量数

据。而进化学习模型则更适合用于对在相等时间间隔内顺序记录的周期性且均匀变化的工作负载观测进行处理。

10 结论与未来方向

本文对用于云环境中资源分配与管理的基于机器学习的工作负载预测模型进行了全面调研，并基于性能评估对其进行了比较。文章讨论了工作负载预测方法的运行设计、效用、动机与挑战。基于各种预测模型在概念和运行特征方面的差异，本文提出了机器学习驱动的预测模型分类体系，并对各预测模型所采用的主要预测方法进行了深入讨论。此外，本文在统一平台上实现了所讨论的全部预测模型，以对这些模型的性能进行广泛研究和比较。基于深入研究与性能评估，本文讨论了这些预测模型之间的权衡及其适用性，从而对云工作负载预测模型进行了整体性总结。未来，可以利用可解释人工智能（XAI）方法构建更稳健的工作负载预测模型，并引入可追溯机制，以帮助刻画人工智能驱动的资源管理中的准确性、透明度、公平性和预测结果。此外，还可以通过采用轻量级优化算法优化量子比特网络并降低其计算复杂度，提升量子神经网络预测模型的效率。

致谢

本研究得到印度古鲁格舍特拉国家技术学院、法兰克福歌德大学、奥地利科学基金（FWF）和德国研究基金会（DFG）的资助，项目编号 I 4800-N（ADVISE），资助期限为2020—2023年。我们感谢审稿人和编辑帮助改进本文。

参考文献

[1] D. Saxena、A. K. Singh 和 R. Buyya，“OP-MLB：一种基于在线虚拟机预测的多目标负载均衡框架，用于云数据中心的资源管理”，*IEEE* 云计算汇刊，2021年。[2] D. Saxena、I. Gupta、J. Kumar、A. K. Singh 和 X.Wen，“一种面向云数据中心的安全多目标虚拟机放置框架”，*IEEE* 系统期刊，2021年。[3] R. Mishra、R. Singh 和 T. Papadopoulos，“数字化导向与数据驱动创新的联系：一个 SAP-LAP 关联框架及研究命题”，*IEEE* 工程管理汇刊，2022年。[4] “云计算市场规模、份额及趋势分析报告：按服务类型（基础设施即服务、平台即服务、软件即服务）、部署方式（公有、私有、混合）、企业规模及最终用途划分。”细分市场预测，2022—2030。

[5] Z. Ren、J. Wan 和 P. Deng，“用于复杂设备生命周期管理的机器学习驱动数字孪生”，*IEEE*新兴计算主题汇刊，2022年。[6] D. Saxena、A. K. Singh、C.-N. Lee 和 R. Buyya，“面向绿色云数据中心的可持续安全负载管理模型”，*Scientific Reports*，2023年。[7] W. Song、Z. Xiao、Q. Chen 和 H. Luo，“利用在线装箱实现云的自适应资源配置”，*IEEE*计算机汇刊，第63卷，第11期，第2647–2660页，2013年。[8] D. Saxena 和 A. Singh，“面向云数据中心的安全嵌入式动态资源分配模型”，电子快报，第56卷，第20期，第1062–1065页，2020年。[9] D. Saxena 和 A. K. Singh，“Osc-mc：面向云环境的在线安全通信模型”，*IEEE*通信快报，第25卷，第9期，第2844–2848页，2021年。[10] ——，“Ofp-tm：面向云计算环境高可用性的在线虚拟机故障预测与容错模型”，超级计算杂志，第78卷，第6期，第8003–8024页，2022年。[11] ——，“基于虚拟机重要性排名和资源估算的云应用高可用性管理模型”，*IEEE*服务计算汇刊，2022年。[12] ——，“基于机器学习的云计算环境工作负载预测与资源管理模型”，*arXiv*预印本 *arXiv:2106.15112*，2021年。[13] R. Gupta、D. Saxena、I. Gupta、A. Makkar 和 A. K. Singh，“用于云网络通信的量子机器学习驱动恶意用户预测”，*IEEE*网络快报，2022年。[14] X. Wang、L. Ma、X. Wang、Y. Shi、B. Yi 和 M. Huang，“网络功能虚拟化市场中采用灵活资源配置的真实VNFI采购机制”，*IEEE*云计算汇刊，2022年。[15]D. Saxena 和 A. K. Singh，“面向云数据中心的通信成本感知型资源高效负载均衡（CARE-LB）框架”，计算机科学与通信最新进展，第12卷，第1–00页，2020年。[16]A. K. Singh 和 D. Saxena，“基于密码学和机器学习的联邦云服务环境安全数据共享身份验证”，应用安全研究期刊，第1–24页，2021年。[17]D. Saxena 和 A. K. Singh，“基于智能流量熵学习的云网络负载管理模型”，*IEEE*网络快报，第4卷，第2期，第59–63页，2022年。[18] Y. Xie、L. Pan、S. Yang 和 S. Liu，“一种在亚马逊云市场转售预留型基础设施即服务实例的随机在线算法”，*IEEE*网络科学与工程汇刊，2022年。[19] J. Bi、H. Yuan 和 M. Zhou，“分布式云中多应用整合工作负载的时间预测”，*IEEE*自动化科学与工程汇刊，2019年。[20] H. D. Kabir、A. Khosravi、S. K. Mondal、M. Rahman、S. Nahavandi 和 R. Buyya，“云计算中的不确定性感知决策：基础与未来方向”，*ACM*计算调查（*CSUR*），第54卷，第4期，第1–30页，2021年。[21] D. Saxena 和 A. K. Singh，“一种主动式自动扩缩容

以及基于在线多资源神经网络的节能虚拟机分配框架，适用于云数据中心”，《神经计算》，2020年。[22] D. Saxena、I. Gupta、A. K. Singh 和 C.-N. Lee，“面向云服务高可用性的容错弹性资源管理框架”，*IEEE*网络与服务管理汇刊，2022年。[23] D. Saxena 和 A. K. Singh，“面向云计算环境的智能安全中心化资源高效资源管理模型”，*arXiv*预印本 *arXiv:2210.16602*，2022年。[24] C. Griner、J. Zerwas、A. Blenk、M. Ghobadi、S. Schmid 和 C. Avin，“Cerberus：数据中心拓扑设计中的选择之力——吞吐量视角”，*ACM*计算机系统测量与分析会议录，第5卷，第3期，第1–33页，2021年。[25] J. Kumar 和 A. K. Singh，“利用人工神经网络和自适应差分进化进行云工作负载预测”，未来一代计算机系统，第81卷，第41–52页，2018年。[26]J. Kumar、D. Saxena、A. K. Singh 和 A. Mohan，“用于云数据中心工作负载预测的双相自适应学习神经网络模型”，软计算，第1–18页，2020年。[27] J. Kumar、A. K. Singh 和 R. Buyya，“用于云资源管理的自引导学习工作负载预测模型”，《信息科学》，第543卷，第345–366页，2021年。[28] R. Khorsand、M. Ghobaei-Arani 和 M. Ramezanpour，“云计算环境中多层应用自主资源配置的模糊层次分析法”，《软件：实践与经验》，第48卷，第12期，第2147–2173页，2018年。[29] J. Kumar、R. Goomer 和 A. K. Singh，“基于长短期记忆循环神经网络（LSTM-RNN）的云数据中心工作负载预测模型”，《计算机科学程序汇刊》，第125卷，第676–682页，2018年。[30]X. Tang，“利用并行改进型LSTM神经网络预测大规模计算系统工作负载”，*IEEE*《开放获取》，第7卷，第40 525–40 533页，2019年。[31] J. Gao、H. Wang 和 H. Shen，“利用深度学习预测云数据中心任务故障”，*IEEE*服务计算汇刊，2020年。[32] L. Ruan、Y. Bai、S. Li、S. He 和 L. Xiao，“存储系统中的工作负载时间序列预测：一种基于深度学习的方法”，集群计算，第1–11页，2021年。[33] L. Ruan、Y. Bai、S. Li、J. Lv、T. Zhang、L. Xiao、H. Fang、C. Wang 和 Y. Xue，“通过云特征增强深度学习预测云工作负载转折点”，*IEEE*云计算汇刊，2022年。[34] S. Tuli、S. S. Gill、P. Garraghan、R. Buyya、G. Casale 和 N. Jennings，“START：利用编码器LSTM网络预测并缓解云计算环境中的拖尾任务”，*IEEE*服务计算汇刊，2021年。[35] Q. Zhang、L. T. Yang、Z. Yan、Z. Chen 和 P. Li，“用于工业信息学的高效深度学习云工作负载预测模型”，*IEEE*工业信息学汇刊，第14卷，第7期，第3170–3178页，2018年。[36]Z. Chen、J. Hu、G. Min、A. Y. Zomaya 和 T. El-Ghazawi，“实现对高维、高度可变云工作负载的准确预测

*IEEE*并行与分布式系统汇刊，第31卷，第4期，第923–934页，2019年。[37] C. Peng、Y. Li、Y. Yu、Y. Zhou和S. Du，“基于GRU编码器-解码器的云计算多步主机负载预测”，载于*2018*年第*10*届知识与智能技术国际会议。IEEE，2018年，第186–191页。[38] F. Qiu、B. Zhang和J. Guo，“一种用于云虚拟机工作负载预测的深度学习方法”，载于*2016*年第*17*届*IEEE/ACIS*软件工程、人工智能、网络与并行/分布式计算国际会议（*SNPD*）。IEEE，2016年，第319–324页。[39] W. Zhang、P. Duan、L. T. Yang、F. Xia、Z. Li、Q. Lu、W. Gong和S. Yang，“基于深度置信网络的云计算环境资源请求预测”，《软件：实践与经验》，第47卷，第3期，第473–488页，2017年。[40] Y. Wen、Y. Wang、J. Liu、B. Cao和Q. Fu，“基于深度置信网络和粒子群优化的云资源配置CPU使用率预测”，《并发与计算：实践与经验》，第32卷，第14期，文章编号e5730，2020年。[41] M. Xu、C. Song、H. Wu、S. S. Gill、K. Ye和C. Xu，“ES-DNN：一种基于深度神经网络的云环境多变量工作负载预测方法”，*arXiv*预印本 *arXiv:2203.02684*，2022年。[42]P. Bhagtya、S. Raghavan和K. Chandraseakran，“使用深度神经网络模型对多虚拟机云环境中的工作负载进行分类”，载于第*36*届*ACM*应用计算年度研讨会论文集，2021年，第79–82页。[43] Y. Li、H. Hu、Y. Wen和J. Zhang，“通过深度神经网络实现数据中心运营功耗预测”，载于第*5*届能源高效数据中心国际研讨会，2016年，第1–10页。[44]J. Bi、S. Li、H. Yuan、Z. Zhao和H. Liu，“用于预测云计算系统中任务时间序列的深度神经网络”，载于*2019*年*IEEE*第*16*届网络、感知与控制国际会议（*ICNSC*）。IEEE，2019年，第86–91页。[45] S. Kardani-Moghaddam、R. Buyya和K. Ramamohanarao，“ADRL：一种基于混合异常感知深度强化学习的云资源扩展方法”，*IEEE* 并行与分布式系统汇刊，第32卷，第3期，第514–526页，2020年。[46] M. E. Karim、M. M. S. Maswood、S. Das和A. G. Alharbi，“BHyPreC：一种基于双向长短期记忆网络的混合循环神经网络新模型，用于预测云虚拟机的CPU工作负载”，*IEEE*《开放获取》，第9卷，第131 476–131 495页，2021年。[47] J. Bi、S. Li、H. Yuan和M. Zhou，“用于云系统工作负载与资源预测的集成深度学习方法”，《神经计算》，第424卷，第35–48页，2021年。[48]Z. Chen、Y. Zhu、Y. Di和S. Feng，“使用集成模型和基于减法模糊聚类的模糊神经网络进行云资源需求自适应预测”，《计算智能与神经科学》，第2015卷，2015年。[49] C. Liu、C. Liu、Y. Shang、S. Chen、B. Cheng和J. Chen，“一种基于工作负载

模式辨识的云自适应预测方法”，《网络与计算机应用杂志》，第80卷，第35–44页，2017年。[50] M. N. H. Shuvo、M. M. S. Maswood和A. G. Alharbi，“LSRU：一种基于深度学习的混合新方法，用于预测云数据中心中虚拟机的工作负载”，载于*2020*年*IEEE*第十区研讨会（*TENSYMP*）。IEEE，2020年，第1604–1607页。[51]N. Singh和S. Rao，“面向大规模工作负载预测的集成学习”，*IEEE* 新兴计算主题汇刊，第2卷，第2期，第149–165页，2014年。[52] B. Feng、Z. Ding和C. Jiang，“FAST：一种结合自适应滑动窗口与时间局部性整合的动态云工作负载预测模型”，*IEEE* 服务计算汇刊，2022年。[53] I. K. Kim、W. Wang、Y. Qi和M. Humphrey，“使用CloudInsight预测云应用工作负载，实现预测性资源管理”，*IEEE* 云计算汇刊，2020年。[54]W. Iqbal、J. L. Berral、A. Erradi、D. Carrera等，“用于估算数据中心资源利用率的自适应预测模型”，*IEEE* 网络与服务管理汇刊，第16卷，第4期，第1681–1693页，2019年。[55] J. Kumar、A. K. Singh和R. Buyya，“基于集成学习的虚拟机资源请求预测框架”，《神经计算》，第397卷，第20–30页，2020年。[56] A. K. Singh、D. Saxena、J. Kumar和V. Gupta，“一种用于云工作负载自适应预测的量子方法”，*IEEE* 并行与分布式系统汇刊，2021年。[57] J. J. Prevost、K. Nagothu、B. Kelley和M. Jamshidi，“使用随机模型和神经模型预测云数据中心网络负载”，载于*2011*年第*6*届系统工程国际会议。IEEE，2011年，第276–281页。[58] D. Saxena和A. K. Singh，“动态云环境中的基于自适应学习的工作负载预测”，《国际计算机与应用杂志》，第1–11页，2020年。[59] J. Kumar和A. K. Singh，“使用神经网络和黑洞算法实现云中的动态资源扩展”，载于*2016*年第五届生态友好型计算与通信系统国际会议（*ICECCS*）。IEEE，2016年，第63–67页。[60]W. Zhang和P. Duan，“基于深度置信网络的云资源需求预测”，载于*2015*年*IEEE*第*12*届普适智能与计算国际会议、*2015*年*IEEE*第*12*届自主与可信计算国际会议、*2015*年*IEEE*第*15*届可扩展计算与通信国际会议及其相关研讨会（*UIC-ATC-ScalCom*）。IEEE，2015年，第1043–1048页。[61] D. Chen、X. Zhang、L. L. Wang和Z. Han，“基于分层毕达哥拉斯模糊深度神经网络的云资源需求预测”，*IEEE* 服务计算汇刊，2019年。[62] K. Qazi和I. Aizenberg，“使用复值神经网络进行云数据中心工作负载预测”，载于*2018*年*IEEE*第二届数据流挖掘与处理国际会议（*DSMP*）。IEEE，2018年，第315–321页。[63]J. L. H. C. Reiss、J. Wilkes，“Google集群跟踪数据：格式+架构”，*Google*公司，白皮书，2011年。[64] A. Beloglazov 和 R. Buyya，“云数据中心中虚拟机能源与性能高效动态整合的最优在线确定性算法与自适应启发式方法”，并发与计算：实践与体验，第24卷第13期，第1397–1420页，2012年。



Deepika Saxena 是法兰克福歌德大学的博士后研究助理。她在印度古鲁格舍特拉国立理工学院（NIT）计算机应用系获得博士学位。2014年，她在印度古鲁格舍特拉古鲁格舍特拉大学获得计算机科学与工程硕士学位。她的主要研究兴趣包括神经网络、进化算法、资源管理和云计算安全。



JitendraKumar 是印度蒂鲁吉拉伯利国立理工学院计算机应用系的助理教授。他于2019年在印度古鲁格舍特拉国立理工学院获得博士学位。目前的研究兴趣包括云计算、机器学习、数据分析和并行处理。



Ashutosh Kumar Singh 现任印度库鲁克舍特拉国立理工学院计算机应用系教授。他曾在印度、英国和马来西亚的多所大学从事研究工作，累计超过 20 年。他在印度印度理工学院（BHU）获得博士学位，并在英国布里斯托尔大学计算机科学系完成博士后研究。他还是英国特许工程师。研究领域包括数字电路的验证、综合、设计与测试，数据科学、云计算、机器学习、安全、大数据。



Stefan Schmid是德国柏林工业大学教授，同时在德国弗劳恩霍夫安全信息技术研究所（SIT）兼职，并任职于奥地利维也纳大学计算机科学系。他在苏黎世联邦理工学院获得理学硕士和博士学位，曾在慕尼黑工业大学和帕德博恩大学从事博士后研究，曾任柏林 T-Labs 高级研究科学家、丹麦奥尔堡大学副教授，以及奥地利维也纳大学正教授。他获得了2016年获得IEEE通信学会ITC早期职业奖，2019年获得ERC巩固者资助。