基于分层插值神经网络的疫苗回收率预测技术

刘潇，孙杰，王连虎

金宇保灵生物药品有限公司，内蒙古 呼和浩特 010020

摘　　要： 疫苗回收率的准确预测对于疫苗生产的效率和成本控制具有重要影响。本文依托深度学习原理，提出了一种基于分层插值神经网络的疫苗回收率预测技术。该技术能够依据疫苗生产流程中的实时参数，对物理抗原146S的回收效率进行精确预测，从而为疫苗生产过程提供量化评价标准。研究中，我们采用了金宇保灵生物药品有限公司提供的口蹄疫疫苗生产数据集对网络模型进行训练和验证。实验结果表明，该模型在MSE、RMSE、MAE、MAPE、R^2等评价指标上表现出了良好的预测性能和数据拟合能力。此外，通过对测试集数据的预测分析，模型展示出了与实际回收率数据的高度一致性，验证了其在疫苗生产实际操作中的参考价值。

关 键 词： 疫苗回收率；分层插值神经网络；物理抗原146S

Vaccine Yield Prediction Technique Based on Hierarchical Interpolation Neural Network

Liu Xiao ,Sun Jie ,Wang Lianhu

The Spirit Jinyu Biological Pharmaceutical Co., Ltd,Hohhot, Inner Mongolia 010020

Abstract: Accurately predicting the yield of vaccine has significant implications for the efficiency and cost control of vaccine production. In this paper, relying on the principles of deep learning, we propose a vaccine yield prediction technique based on a hierarchical interpolation neural network. This technique can precisely forecast the recovery efficiency of the physical antigen 146S based on real-time parameters in the vaccine production process, thereby providing a quantitative evaluation standard for the vaccine production process. In our research, we utilized a dataset provided by Jin Yu Bao Ling Biotechnology Co., Ltd. to train and validate the network model. Experimental results demonstrate that the model exhibits good predictive performance and data fitting ability in terms of evaluation metrics such as MSE, RMSE, MAE, MAPE, and R^2. Furthermore, through predictive analysis of the test dataset, the model demonstrates a high degree of consistency with actual yield data, confirming its reference value in practical vaccine production operations.

Key words: vaccine yield; hierarchical interpolation neural network; physical antigen 146S

引言

细胞培育与疫苗生产是生物制药领域的核心，对人类社会的健康和持续发展有着巨大贡献。疫苗的研发与制造对于控制乃至根除全球性疾病大流行，具有至关重要的作用[1]。随着生物科技的飞速发展，新型疫苗的研发已展现出对新出现病原体迅速反应的能力，为全球公共健康安全构筑了坚固屏障。

近年来，中国在工业自动化控制系统领域取得了显著成就。一部分具备实力的疫苗生产企业已经成功实现了生产流程的全面自动化[2]。然而，对于细胞密度的测定、细胞病变程度的评估等关键环节，仍然主要依赖于操作人员的经验和专业知识。由于这些环节受到人为操作方式和主观判断的影响，不同批次之间的差异较大，因此，疫苗的生产在很大程度上仍然依赖于操作者的经验。

本研究依托于深度学习的核心原理，结合疫苗生产的实际应用，提出了一种基于分层插值神经网络的疫苗回收率预测方法。该方法能够依据疫苗生产流程中可获取的参数，对物理抗原146S的回收效率进行较为准确的预测，进而为浓缩抗原溶液的纯化处理提供量化评价标准。通过该标准可对疫苗生产成本进行有效的控制和评估。

* 作者简介：

刘潇，男，蒙古族，硕士学历，高级工程师，研究方向为微生物发酵、制药工程

孙杰，男，汉，本科学历，研究方向为软件工程、项目管理

王连虎，男，汉，本科学历，初级工程师，研究方向为业财信息化建设

致谢：该论文成果受内蒙古自治区科技计划项目（编号：2022YFXM0005）资助

一、研究背景与基础

（一）疫苗质量管理的重要性

在整个疫苗研发、生产、采购到接种的链条中，每个步骤都极为关键。尤其是疫苗的生产阶段，它对环境条件有着极为严格的要求，且生产技术极其复杂。以细胞培养和灭活疫苗的制备为例，其生产流程包含数十个精细步骤，且多数步骤需在无菌条件下进行，并对温度有精确的控制需求。生产过程中任何一个细节的失误，都可能影响到一部分乃至整批疫苗的质量[3]。

当前，中国国家药品检验实验室依据中国合格评定国家认可委员会（CNAS）颁布的《检测和校准实验室能力通用要求》（CNAS CL01）等标准构建了管理体系。这一改进对于保障疫苗在其全生命周期内检验数据的高标准质量、实现疫苗监管的国际标准化具有十分重要的作用，且解决此问题已变得尤为迫切[4]。

构建高效的疫苗生产质量管理体系至关重要，它应覆盖原料采购、变更管理、偏差调查等关键环节，旨在识别和控制潜在风险。这一体系有助于企业迅速发现并应对影响药品质量的隐患，有效预防质量事故，确保药品安全有效[5]。

（二）疫苗回收率

疫苗回收率，是指在疫苗制造、质量检测和分析过程中特定组分（例如抗原、佐剂等）的回收效能。这一指标是评价疫苗生产技术效率和最终产品质量的一个重要参考。它与成本控制、疫苗引发的免疫反应以及产物的一致性紧密相关。一个高的回收率意味着在制备过程中，能够有效地捕获和重复使用更多的疫苗有效成分，从而显著降低原料浪费并提升大规模生产时的成本效益。同时提高特定成分的回收率有助于最大程度地减少生产过程中可能引入的不纯物和污染，这对于维护疫苗的安全性和效能是至关重要的。

以mRNA疫苗的制造为例，其生产过程是一项包含多个关键环节的复杂生物技术工程，主要步骤包括体外转录（IVT）、mRNA的修饰处理，以及mRNA脂质体的包装等。在疫苗生产的全周期中，防止外界污染和杂质的引入是保障疫苗安全性和有效性的核心要素[6]。在此过程中，良好的回收率不仅有助于保持mRNA分子的完整性和生物学效能，同时也有效降低了因操作不当或生产缺陷引入的污染可能性。

（三）分层差值神经网络

分层差值神经网络（HDNN）的设计灵感，来自于模仿生物神经网络的层级化信息处理机制，以及对现有深度学习模型的优化需求。进入21世纪，深度学习领域的突破性进展，尤其是卷积神经网络（CNN）在图像识别上的成功，为HDNN的诞生提供了理论和技术基础。HDNN利用其分层次结构，有效处理和捕捉多尺度数据的复杂特征[7]，广泛应用于图像识别、语音处理和自然语言处理等众多领域。

HDNN通过模拟生物神经系统中的层级结构来处理和学习数据。这种网络设计的核心在于其分层架构，它能够逐级抽象和提炼输入数据的特征，从而捕捉到数据中的复杂关系和模式。随着层级的增加，这些基础特征被进一步组合和抽象，以识别更高级的概念或模式，如物体的形状或场景的类别。

此外，分层结构还有助于减少网络的参数数量，提高计算效率，并增强模型的泛化能力。在每个层级，差分学习机制使得网络能够对输入数据的微小变化做出反应，这增强了模型对新情况的适应性。通过这种方式，分层差值神经网络能够从大量数据中学习到丰富的知识，为解决各种复杂问题提供了强大的工具。

二、实验数据与实验环境

（一）实验数据

在本项研究中，我们所使用的数据集由金宇保灵生物药品有限公司提供。该数据集涵盖了该公司口蹄疫疫苗生产过程中关键参数的实时记录，共包含365条生产记录，每条记录详细记录了一种细胞疫苗的抗原批号、化学批号、毒株类型、培养条件、接毒密度、浓缩过程、146S含量、蛋白含量、内毒素含量、纯化过程、质量控制措施、物理抗原蛋白总量、纯化后内毒素水平、146S回收率以及蛋白与146S的比率等详细信息。

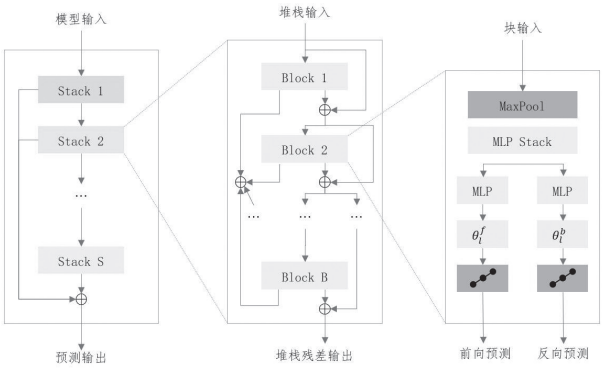
（二）实验软硬件环境

在实验开始之前，必须对实验平台的硬件和软件环境进行适当的设置。在硬件配置上，本研究采用了一台安装有Ubuntu 20.04 LTS操作系统的计算机，该机器配备了32GB的随机存取存储器（RAM）和一款Intel Core i9处理器，以确保强大的中央处理单元（CPU）计算性能；此外，还装备了一款NVIDIA GeForce RTX 3080图形处理单元（GPU），该GPU拥有众多CUDA核心，非常适合进行深度学习的并行计算任务。在软件配置方面，本研究选用了PyTorch深度学习框架，并依据研究需求安装了如pandas和numpy等附加的第三方库。开发过程中，本文在Windows 11操作系统环境下，利用Visual Studio Code作为开发工具，通过远程方式进行操作，并采用Python编程语言来实现模型的编码。

三、实验方法

（一）网络结构设计

1.基于分层差值神经网络的基本结构



>图1 基于分层差值神经网络的疫苗回收率预测模型结构图

本文构建网络结构如图1所示，将指定的过程参数输送给模型

输入即可对模型进行训练得到指定的预测输出参数。该模型框架依托于多层差值神经网络，由多个串接的多层感知器（MLP）组成，并通过 ReLU 激活函数引入非线性特性。这些 MLP 被划分为若干个堆栈（Stacks），每个堆栈由多个块（Blocks）构成，每个块负责挖掘时间序列数据的独特特征。每个块产生两个预测方向：前向预测被累加来形成最终的预测输出，而反向预测则用于对后续块的输入信号进行优化。这些块通过残差连接相互串联，这不仅使得跨堆叠的信息传递和累积成为可能，而且促进了在不同信号频率带中的专业化预测，从而在降低内存需求和计算成本的同时，增强了模型的预测精度和架构的精简性。

2.多速率信号采样

该模块通过应用不同大小的池化核（MaxPool）进行时域下采样，可以将时间序列转换为具有多种时间粒度的序列。具体而言，较大的池化核尺寸倾向于捕获更低频率和更宽时间尺度的成分，而较小的核尺寸则能够捕捉到更高频率和更窄的时间尺度的细节。这样我们就可以利用得到不同核尺寸的池化层得到不同尺度的序列信息，这种多尺度的采样方法在时间序列分析中是一种常见的预处理技术。

这种方法的优势非常明显：首先，下采样能够减少序列的长度，从而降低模型的计算复杂度，提高运算效率。其次，它有助于减少模型参数的数量，这不仅有助于防止过拟合，还能维持模型对原始输入数据的感知能力。通过利用具有不同核尺寸的池化层，模型能够获得并分析时间序列在不同时间尺度上的特征，为构建更为精确和高效的时间序列预测模型提供了一种有效的手段。

3.分层差值计算

分层差值计算模块的过程相当直观，并且与下采样形成了对应关系。在预测阶段，我们进行了一个上采样步骤以恢复到所需的序列长度。结合网络架构图，我们可以更清晰看到：在模型的第一个堆栈中，由于使用了较大的池化核尺寸进行下采样，输入序列变得更为紧凑且具有更大的时间尺度，相应地，预测出的未来序列也较短。为了达到期望的预测范围，必须执行上采样，即通过插值（例如线性或多项式插值）来增加序列的长度，这可能涉及到在序列中插入大量的点，其插值后的预测结果较为平滑，反映了更低频率的成分。

相反，在模型的最后一个堆叠中，较小的池化核尺寸意味着下采样后的序列较长，时间尺度较小，因此预测出的未来序列也较长，这样就减少了所需的插值量，插值结果能够反映了更高频率的成分。实际上，每个堆栈负责处理不同时间尺度的预测任务，最终将这些不同尺度的预测序列通过插值调整到相同的时间粒度，并将它们相加以形成最终的预测结果。这样插值结果反映了更高频率的成分。

本文采用指数递减的方式来选择每个堆叠的池化核尺寸，这样可以确保模型在不同的时间尺度上都能有效地捕捉和预测时间序列的特征。

（二）网络输入数据及参数设置

在本研究中，我们从数据集中精心挑选了10个关键参数，包括细胞类型、转出细胞密度、病毒维持液种类、初始培养密度、细胞代数、沉降时长、细胞复苏密度、所用毒株、深层过滤的最大压力以及浓缩倍数，作为深度学习网络的输入变量。网络的输出则聚焦于三个关键指标：浓缩液还原后的146S浓度、纯化抗原还原后的146S浓度以及146S的回收率。物理抗原146S回收率的准确预测对于实现疫苗生产的高效率和高标准至关重要。

为了验证模型的性能，我们采用了数据集中的292条疫苗生产记录作为训练集来训练模型，而剩余的73条记录则被划分为验证集，用于评估模型的实际效果。在网络训练过程中，我们设定每个隐藏层中神经元的数量为8，并执行了100轮训练。训练采用了0.001的学习率并设置训练批次大小为4，同时选用 MAE 作为损失函数，以优化模型的反馈机制。

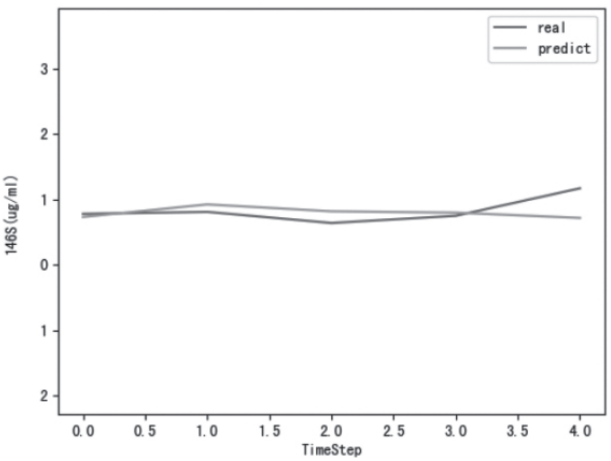
四、结果与分析

表1 两组方法的实验结果比对

	MSE	RMSE	MAE	MAPE	R^2
MLP多层感知机	0.062664	0.250329	0.170503	0.181404	0.090651
本文方法	0.019601	0.140004	0.119836	0.13986	0.381182

两组实验得到的结果如表1所示，实验中分别采用 MLP 多层感知机作为基准模型和我们的方法对数据集进行预测训练。本文评价指标使用了回归算法中常用的评价指标：均方误差（Mean Squared Error，MSE）、均方根误差（Root Mean Squared Error，RMSE）、平均绝对误差（Mean Absolute Error，MAE）、平均绝对百分比误差（Mean Absolute Percentage Error，MAPE）、R^2 决定系数（Coefficient of Determination）。R^2 决定系数是衡量模型拟合数据能力的重要指标，其值域介于0至1之间。R^2 值越接近1，代表模型的拟合效果越佳。

通过对比分析，可以明显观察到，相较于基准模型，本研究提出的方法在五项评价指标上均展现出更优的性能。这表明本方法的预测值与实际观测值之间的偏差更小，误差对预测准确度的影响也相应降低。本方法得到的 R^2 值达到了0.38，相比于原来的0.09有了明显的提升，从而进一步证实了本模型在预测准确性和



>图2 本文方法对146S回收率预测与实际值曲线

数据拟合方面的优势，使其预测结果在实际生产中具有更高的参考价值。

图2展示了利用本模型对部分测试集数据的146S回收率进行预测的实验结果。通过观察图中的曲线我们发现，采用本研究所提出的模型得到的5组预测值与实际回收率数据具有较高一致性，预测误差微小，几乎可以忽略不计。这一结果表明，本研究所构建的预测模型在实际应用中表现出色，其预测结果在疫苗生产的实际操作中具有显著的参考意义和实用价值。

六、结语

本文提出了一种基于分层插值神经网络的疫苗回收率预测方法，旨在提高疫苗生产过程中物理抗原146S回收效率的预测准确性。本研究利用金宇保灵生物药品有限公司提供的生产数据，设计并训练了一个深度学习模型，该模型在多个评估指标上均展现出优于现有方法的性能。本研究的成功实施进一步推动了疫苗生产向更高效率和更低成本的方向发展。未来的工作将集中在模型的进一步优化和在其他疫苗生产流程中的应用探索上，以期达到更广泛的实际应用效果。

参考文献

[1]马磊，杨昭庆，王佑春. 全球疫苗研发现状和展望［J］. 中国药科大学学报，2024，55(1)：115-126.

[2]马相虎，沈谊清，杨月莲，等. 细胞工厂自动化操作系统在水痘疫苗生产中的应用［J］. 中国新药杂志，2014，23(20)：2446-2449.

[3]Lee B Y，Haidari L A. The importance of vaccine supply chains to everyone in the vaccine world［J］. Vaccine，2017，35：4475-4479.

[4]陆明，顾颂青，项新华. 构建满足 WHO-NRA 评估要求的疫苗质量控制实验室质量管理体系的研究［J］. 中国药事，2020，34(12)：1378-1383.

[5]孙京林. 疫苗的质量管理与监管检查［J］. 中国药物评价，2014，31(01)：48-50+60.

[6]张辉，刘建阳，毛群颖，等. mRNA疫苗质量控制进展［J］. 药学进展，2022，46(10)：745-750.

[7]熊桢，郑兰芬，童庆禧. 分层神经网络分类算法［J］. 测绘学报，2000(03)：229-234.