**3. STGCN-LSTM耦合预测模型**

**3.1 特征工程**

为了精准刻画不同国家（或地区）在奥运赛场上的综合实力及其时序演化，我们从数据清洗、实体标准化和关键特征构造三个层面入手，构建了“动态国力权重矩阵”并基于图论思路启发性地引入了“PageRank项目影响力”计算方法。具体流程如下所述。

3.1.1 动态国力权重矩阵

(1) 数据准备与清洗

首先，本研究整合了奥运会奖牌数据与国际货币基金组织（IMF）发布的全球经济与人口数据，其中包括每个国家（或地区）在不同年份的GDP（imfGDP）和人口规模（population），并与夏季奥运会奖牌统计表（summerOly\_medal\_counts.csv）进行合并。由于原始数据在国家/地区编码（NOC）与国名之间存在一定程度的不一致或缺漏，我们在数据加载后进行了字符串清洗（如空格、特殊字符的去除）以及基本的缺失值排查。

(2) 时间衰减因子与动态权重公式

为了体现奖牌对国家实力的影响是随时间逐渐衰减的，我们在代码中定义了衰减系数，并引入了指数衰减因子：

.

在此基础上，为每条奖牌记录计算一个“动态国力权重”（Weight），综合考虑了奖牌总数、经济体量与人口规模，核心公式为：

其中，用于平滑不同国家在奖牌数量上的巨大差异；和则分别衡量经济实力与潜在运动人口基数；而确保较早时期的奖牌对当前国力评估的贡献适度衰减。

(3) 矩阵生成与导出

基于上述公式，我们对每个国家（或地区）在每届奥运会上的权重进行了统一计算，并将所有结果整合为一份按年度排列的“动态国力权重矩阵”（dynamic\_weight\_matrix\_2028proj.csv）。该矩阵既能够在时序维度上反映不同国家竞技水平的衰减与累积效应，也为后续的时空特征分析和深度模型训练提供了关键输入。

3.1.2 国家实体标准化与历史映射

(1) 多路径数据源匹配

在构造动态国力矩阵之前，我们先对各数据集关于国家（或地区）的命名进行了实体标准化。如代码所示，通过拆分“夏季奥运奖牌统计”、“主办地信息”和“运动员信息”三个数据源中出现的国名（或队名）字段，使用历史政权映射表（如苏联→俄罗斯、南斯拉夫→塞尔维亚等）和模糊匹配算法（levenshtein距离、FuzzyWuzzy等）对异常或含糊的名称进行纠正与对齐。该过程极大减少了后续多表关联时的歧义，也为PageRank等网络分析方法打下数据基础。

(2) 历史政权时间轴对齐

针对类似前苏联、东德、南斯拉夫等已不复存在或国名已变迁的实体，我们将奖牌年份与历史决裂节点（例如1991年苏联解体）进行比对，如果某国家/地区的奖牌年份晚于其解体或更名时间，则在分配国别代码（NOC码）时自动切换至后继国家。这样能够兼顾国际奥委会（IOC）官方统计与历史现实变化之间的差异，保证了时空维度上国力赋值的相对精确。

3.1.3 PageRank项目影响力计算

为了进一步从“运动项目”这一视角衡量奥运格局，本研究借鉴网页排名算法中的PageRank，将每项运动视为图中的节点，通过国家/地区对奖牌贡献或共同运动员流动为边建立联系，从而度量项目在整个奥运网络中的影响力。

(1) 网络构建

在构建影响力网络时，我们首先对每个比赛项目在不同年份的奖牌数进行加总，随后乘以相应的主办国权重（若某届奥运会该项目获奖者来自于主办国，额外赋予一定倍数的评分），形成一条“项目—国家”加权边。借助networkx等图分析工具，可将这些边转化为有向带权图结构。

(2) PageRank迭代

在该有向图中，我们统一引入一个辅助节点Sports指代所有项目集合，然后反向连接到具体项目节点。应用nx.pagerank时，以weight字段为参考计算各节点的PageRank值。PageRank数值越高，意味着该运动项目更能吸引世界各国的高度参与或竞争（即在网络上有更多“重要节点”指向它），从而反映其相对影响力。

(3) 结果与应用

最终代码将计算好的影响力得分输出为sport\_pagerank\_scores.csv，其中包含各运动项目名称与对应的PageRank分值。该分值可用于：

* 项目投资与资源分配：在备战下一届奥运会时，优先配置资源到影响力与竞争度更高的项目；
* 跨国教练迁移与技术援助：针对高影响力项目重点部署国际教练团队，实现最大化的综合回报。

**3.1.4 小结**

通过上述步骤，本研究在多源异构的奥运数据基础上完成了：

* 动态国力权重矩阵 的精细化构建，兼具时序衰减和经济人口刻画；
* 国家实体的历史标准化处理，满足跨时空和政区更迭的统一编码需求；
* 项目影响力的图论刻画，使得后续对“优势项目集中度”及“潜力项目挖掘”的预测与决策模型更具针对性。

该特征工程不仅在变量选取上兼顾了经济、人口、历史变迁等宏观维度，还通过PageRank强化了对运动项目本身结构性重要度的刻画，为随后章节（见第3.2节“奖牌分布预测”及第4章“三重差分教练效应分析”）的深度定量研究奠定了方法论基础与数据保障。

可视化建议：

1.参考[1]中对国力演变的可视化方式，我们可采用Cartopy或其他地图工具，将同一国家在不同年份的“Weight”映射到地理空间中，形成一部简易的“奥运国力变化地图”。此外，也可使用热力图（Heatmap）来直观展示不同时段、不同国家之间的权重对比（例如，横轴为Year，纵轴为NOC，着色深浅表示Weight值），以突出强势国家与新兴国家的对比。

2.可使用Gephi或Python的networkx库进行可视化，将项目节点通过Force-directed布局展现出来，节点大小由“PageRank分数”决定，边的粗细由国家间重叠度或转移率决定，从而凸显核心运动项目在整个奥运网络中的地位。

**3.2 奖牌分布预测**

在奥运赛事实证研究中，奖牌分布预测既是理论探究的重要目标，也对各国体育管理与战略规划具有直接的应用价值。本节基于深度集成预测（Deep Ensemble）框架，实现了对各国（或地区）在未来奥运周期奖牌总量的点预测与区间预测。与传统单一模型相比，深度集成可通过“多模型并行”有效降低模型偏误并获取更稳健的置信区间。

3.2.1 深度集成预测框架

(1) 模型构成

如本研究所使用的代码中，DeepEnsemble类内集合了四种子模型：

* STGCN-LSTM：融合空间图卷积（GATConv）与时序循环网络（LSTM），用于捕捉国家间的地理/经济关联和奖牌时序演变；
* HistGradientBoostingRegressor (HGBR)：具备对非线性特征较强的学习能力；
* XGBoost：广泛应用于回归与分类任务，在小样本场景中仍保持稳定表现；
* TemporalTransformer：借鉴Transformer的时序注意力机制，对跨届奥运数据进行特征提取。

在训练过程中，这些子模型分别对“国家节点特征”、“边缘属性”（如经贸往来、历史奖牌相似度）等进行学习，形成对未来奖牌数量的独立预测结果。

(2) 非对称置信区间构建

当我们将四个子模型的预测输出进行聚合后，为了更好地刻画真实预测中的非对称性不确定度，本研究采用了分位数方法（Quantile）来构建上下限区间。具体流程如下：

* 针对同一国家的奖牌数预测，分别收集各子模型的预测值；
* 在得到的预测向量基础上，计算 5% 与 95% 分位数，作为下限与上限；
* 将四个预测模型的均值视为“集成预测的中心值”，由此得出形如

预测区间=

* 的非对称区间。

在代码实现中，这一过程体现为对predictions张量在模型维度上执行np.quantile操作，从而获得上下置信区间。该做法避免了简单“对称区间（均值±常数倍标准差）”可能产生的偏误，更能准确反映经济、政策、地缘等多因素带来的不确定性分布。

3.2.2 预测结果与可视化

将深度集成模型应用于2028年奥运周期的数据后，我们得到了各国未来奥运奖牌的点估计与区间估计。可参考图4（示例）中：

横轴为国家（或地区）代码；

纵轴为预测的奖牌数；

画出中心点（平均预测值）及上下限区间带，以直观展示不确定带的非对称变化。

对比历史数据可发现，一些传统强国（e.g., USA、CHN）虽然在预测分布中仍然居于上位，但其置信区间范围也略显扩大，表明在经济、人口和国际竞争加剧的背景下，不确定性较往届有所上升。相反，些新兴经济体由于周期性投入的增强，区间中值有所提升。

深度集成在一定程度上缓解了单一模型的过拟合风险，然而，如果多个子模型自身均在同一特征区间出现偏差（“群体盲区”），集成结果也可能被放大。因此，需要在后续研究中不断优化特征工程与子模型结构，以提升预测精度与稳健性。

**3.3 新兴国家识别**

在奥运赛场上，“首次斩获奖牌”的事件往往标志着一国体育事业的里程碑与崛起。为此，本研究基于“复合泊松过程”构建了对“新兴国家首枚奖牌概率”的预测框架，从而识别出潜在的“新晋奖牌国”并评估其发生概率和时间窗口。

3.3.1 复合泊松过程原理

(1) 模型动机

由于首次夺牌对于过往零奖牌或少量奖牌的国家而言具有稀有事件特征，单纯采用线性回归或一般离散分布（如伯努利或Poisson）难以准确捕捉其达成路径。复合泊松过程（Compound Poisson Process）能够对“稀有但影响重大的事件频度与强度”进行联合建模：

* 泊松过程决定首枚奖牌出现的“时间（或届次）”分布；
* 复合量（Compound）为潜在的“触发强度”，与动态国力、教练资源输入等多重因素高度相关。

(2) 参数设定与估计

在实现代码中，我们根据国家在历史奥运会中零奖牌的持续期、GDP/人口规模、教练投入等，估计复合泊松过程的核心参数（事件发生率）与（复合增益）。该估计过程可理解为：

​

其中λ可能随时间演变，并被教练投入等要素所修正。

3.3.2 首枚奖牌概率预测

在完成复合泊松参数估计后，即可对同类“尚未夺牌或奖牌数极少”的国家进行首枚奖牌概率推断：

* 选定预测时间窗口（如2028年）；
* 计算该国在该时点的累积出现概率 ；
* 考虑“复合因子”α对教练输入等因素进行修正，使概率进一步上调或下调。

图5可视化示例：在世界地图上以颜色深浅表示首枚奖牌的预测发生概率。可以观察到在南亚、中东欧与非洲部分国家区域，出现初次奖牌的概率正在逐步攀升。这与其近年来在人均收入增长或跨国教练团队引入措施相吻合。

3.3.3 研究启示

* 战略意义：识别出新兴奖牌国不仅具有学术意义，也为国际体育组织及各国政策部门提供了提前布局的信号，如跨国教练培训、设施建设等；
* 建模局限：复合泊松模型仍需更丰富的微观数据（运动员流动、运动项目渗透率等）才能更精准地刻画“首枚奖牌”的触发动因；
* 扩展应用：该框架在预测其他稀少但重要的赛事突破（例如首枚冬奥会奖牌、首个世界纪录）时，同样具备可行性。

**3.4小结**

本节从“奖牌分布预测”和“新兴国家首枚奖牌识别”两个维度，结合深度集成与复合泊松模型，探讨了奥运项目竞赛格局可能的未来走向。深度集成为各国的整体奖牌实力提供了系统化的区间预测；而复合泊松过程则专门用于捕捉那些零基础起步的稀有成功事件（首枚奖牌），从而为政策机构预先识别潜在的“黑马国家”提供了量化支撑。后续可在此基础上，引入更多的赛事级微观指标、运动员个人数据与教练全球迁移图谱，进一步细化对奖牌分布与国家崛起的预测精度。