基于深度学习的多源数据预测与模式分析

1. 研究主题与背景说明

本次期末作业要求学生利用深度学习模型（如 LSTM、GRU、CNN-LSTM、Transformer 等），针对以下任选一个题目进行研究。可以组队完成，最多三人组队。三人队伍只能获得作业最高分的 80%，二人队伍只能获得作业最高分的 90%，加分项另算。
研究需体现以下特征：
融合多源数据（时序 / 文本 / 图像 / 比赛结果 / 环境数据等）；
使用深度学习模型（推荐 LSTM 作为核心时序预测模块）；
包含完整数据处理、建模、可视化与结果分析流程；
提供清晰的研究问题、实验设计与讨论。

2. 可选研究方向与题目案例

题目一：基于 LSTM 的 Wordle 玩家表现预测（2023 MCM Problem C 改编）
研究问题：
Wordle 是一种文字猜词游戏，玩家需要在限定次数内猜出目标单词。
要求利用玩家历史数据（如猜测次数、词汇分布、时间间隔）构建一个基于 LSTM 的预测模型，预测玩家在下一次游戏中的表现（如猜测步数或是否成功）。

研究任务：
将玩家猜词过程编码为时间序列（词向量 + 猜测反馈模式）。
构建 LSTM 模型预测玩家下一轮猜测的准确度或完成时间。
可扩展任务：利用 Transformer 比较性能差异。

参考模型结构：
Embedding + LSTM + Dense
Word2Vec + BiLSTM + Attention

题目二：干旱条件下植物群落响应的时间序列建模（2023 MCM Problem A 改编）
研究问题：
气候变化导致干旱频发，植物群落的生长与恢复能力受显著影响。
要求利用气象与植被指数数据，构建基于 LSTM 的干旱影响预测模型，评估特定区域植被在未来若干周的响应变化。

可用数据集：
MODIS NDVI 植被指数（NASA）
NOAA 降雨量、气温、蒸散量数据
Soil Moisture Active Passive (SMAP) 土壤湿度数据

研究任务：
整合多源气象与生态数据，构建统一时间序列。
利用多变量 LSTM 模型预测 NDVI 的未来变化趋势。

对比传统 ARIMA 与 LSTM 模型性能差异。

参考模型结构：
多变量输入 LSTM
Encoder-Decoder LSTM

题目三：网球比赛动量分析与结果预测（2024 MCM Problem C 改编）
研究问题：
“势头”是体育比赛中一个关键但主观的概念。要求基于网球比赛逐分数据，利用 LSTM 识别比赛中“势头”变化，并预测下一分或下一局的胜负结果。

研究任务：
构建时间序列（按每一分的得失序列）。
使用 LSTM 预测下一个时刻选手得分概率。
对比 RNN / GRU / Transformer 模型的预测性能。

参考模型结构：
LSTM + Dropout + Dense
GRU + Attention

应用价值：
为体育数据分析、智能赛事预测和策略评估提供支持。

3. 实验设计与技术实现

编程环境：
Python + Jupyter Notebook
PyTorch
Hugging Face Datasets、WandB 实验记录

实验步骤：
数据收集与清洗（时间对齐、缺失值插补）
特征工程（归一化、时间窗口化）
模型构建与超参数调优（LSTM 单元数、序列长度）
性能评估（MAE、RMSE、F1-score、AUC）
可视化与解释（Loss 曲线、预测结果趋势图、动量变化热力图等）

4. 提交内容

报告采用 MCM 论文格式撰写，可以写中文。
交付物清单
1. 代码工程：含模块化源代码及单元测试脚本
2. 数据集：原始数据及预处理后的数据文件（CSV/Parquet 格式），附数据集描述文档，使用 huggingface。
3. 实验报告：MCM 格式 PDF 文档，含研究局限与未来改进方向(代码全部上传至 github，

将链接附在报告中)

5. 拓展与挑战方向（选做）

尝试引入 Transformer / Temporal Fusion Transformer (TFT)
将预测任务拓展为多步时序预测
构建结果展示仪表盘
交互式参数调节（时间窗口、模型类型）

所有交付物需打包为压缩文件，命名格式为“姓名 1-姓名 2-姓名 3-v2.0.zip”，确保文件结构清晰且无几余内容。
上交到 https://box.nju.edu.cn/u/d/ebf5567cab00425d8b43/