The Design and Implementation of Stock Investment Assistant System Based on Transformer

毕业设计答辩

许振华

计算机学院（国家示范性软件学院）

2024 年 5 月 28 日

金融市场与技术融合：股票预测的新视角

- 经济全球化与信息技术深度结合，重塑金融业态。
- 大数据、AI 等技术推进量化投资、智能投顾等服务，增强市场分析粒度。
- 股票预测面临复杂挑战：多因素交织，高度不确定性。
- 传统模型（统计方法、技术分析）局限性：非线性关系与市场情绪捕捉不足。
- 深度学习，特别是 Transformer 模型，革新金融预测。
- Transformer 通过自注意力机制，优化长序列数据分析，捕捉周期性与异常波动。
- 深度强化学习优化交易策略，提升智能化水平。
- 面临挑战：计算资源需求、实时学习适应性、泛化能力、解释性与合规性。

机器学习在金融预测方面的作用

- 量化交易与风险管理：循环神经网络 (RNN) 与长短期记忆网络 (LSTM)：这些模型擅长处理时间序列数据，如股票价格、交易量，能够捕捉市场的非线性动态和复杂模式，提高对市场走势的预测精度和风险管理能力。
- 卷积神经网络 (CNN) 的应用：CNN 以其在图像处理领域的强大能力，被创新性地应用于股票市场预测。通过将股票价格序列视作图像数据，CNN 能够捕捉序列中的局部特征，类似识别图像中的模式，从而理解并预测价格变化趋势。
- 多层 Encoder 的作用：引入多层 Encoder 旨在增强模型对时间序列数据长期依赖性的理解和利用，这对于具有周期性和持续性影响因素的股票市场至关重要。多层 Encoder 逐步提炼序列的抽象表示，帮助模型捕获更深层次的市场动态。
- 综合模型设计：综合 CNN 的局部特征提取能力和多层 Encoder 的长期依赖建模能力，设计端到端深度学习模型，直接从股票价格序列中学习并预测未来走势，提高了预测的准确性和实用性。

经典案例

- 支持向量机 **(SVM)**: 利用灵活的核函数技术，巧妙地处理非线性关系与高维数据，展现优越的泛化性能，尤其适用于小样本情况；然而，面对大规模数据集，寻找最优超平面的计算开销不容忽视。
- 随机森林 **(Random Forest)**: 作为集成学习的典范，通过构建多棵决策树协同工作，有效处理时间序列中的多因素交互影响，同时提供特征重要性排名，增强模型解释性；但其对噪声的敏感性及在极端非线性问题上的局限性需谨慎考虑。
- 循环神经网络 **(RNN)** 与长短期记忆网络 **(LSTM)**: 专为序列数据设计，LSTM 通过独特的门控机制，有效缓解了长期依赖问题，成为时间序列预测，尤其是如股市数据等复杂序列分析的优选工具；不过，这类模型训练通常要求大量数据支撑及较高的计算资源消耗。
- 卷积神经网络 **(CNN)**: 虽起源于图像识别，但在时间序列分析中找到了新舞台，凭借局部感受野与权值共享机制，高效捕捉序列中的重复模式与周期性特征，特别适合分析具有明确结构特征的时间序列；相对而言，CNN 在处理远距离依赖关系时可能不及 LSTM 等循环结构。

综合模型与混合策略

- 集成学习：结合模型优势，如 LSTM-CNN，统计-机器学习融合。
- 目标：提升预测准确度与稳定性，适应金融市场复杂性。
- 国内外研究聚焦模型创新与混合策略，提升预测效能。
- 全球趋势：深度学习、计算智能技术（强化学习、GANs）的融合应用。

本研究模型创新

- 提出 TECEC 模型，融合全局与局部特征学习。
- 利用 Encoder 的自注意力机制捕捉序列全局依赖，利用卷积神经网络强化局部特征提取。
- 目标：致力于提升模型的泛化性能，深入探究并揭示图像内在数据指标间的微妙联系，诸如多个指标之间的潜在关联，以此优化对市场动态的深刻洞察与增强预测的精确度。
- 未来展望：模型结构优化，结合更多金融指标，实时交易系统应用。

股票市场数据的获取与处理

- 以下是股票指数数据的统计表格：

表 1: 股票原生数据统计

代码	名称	市场标识	发布机构	类别	上市时间
000905.SH	中证 500	SSE	中证指数有限公司	规模指数	07-01-15
000015.SH	红利指数	SSE	中证指数有限公司	策略指数	05-01-04
000852.SH	中证 1000	SSE	中证指数有限公司	规模指数	14-10-17
000927.SH	央企 100	CSI	中证指数有限公司	主题指数	09-03-30
000922.CSI	中证红利	CSI	中证指数有限公司	策略指数	08-05-26
399006.SZ	创业板指	SZSE	深圳证券交易所	规模指数	10-06-01

- 数据来源：TUSHARE
- 数据日期范围：2012-04-01 至 2024-04-01

股票数据的展示和预处理

- 股票 000922.CSI（中证红利）的具体数据样本

表 2: 000922.CSI 数据样本

日期	开盘价	最高价	最低价	收盘价	成交量
2012/4/5	2510.98	2460.273	2516.41	2446.842	13,724,236
2012/4/6	2517.627	2507.342	2521.875	2500.403	10,799,499
2012/4/9	2493.413	2509.527	2514.991	2491.37	7,997,286
2012/4/10	2515.168	2486.785	2515.168	2460.993	9,450,665
2012/4/11	2512.143	2491.379	2531.047	2486.061	9,820,757

- 时间序列处理增强: 通过将日期设置为数据的索引来加强数据的时间结构特性
- 初步数据清洗: 清除了包含缺失值的记录，确保数据分析的基础是完整且准确的数据集
- 技术指标提取: 为了深入分析股票价格动态，文章计算了一系列金融技术指标，具体计算的技术指标包括相对强弱指数 (RSI)、威廉指标 (WILLR)、不同类型的移动平均线 (SMA、EMA、WMA、HMA、TEMA)、趋势强度指标 (ADX) 以及动量指标 (CCI、CMO、ROC) 等
- 多时间窗口分析: 每个技术指标都在多个时间窗口（6 日至 20 日）下进行计算
- 全面覆盖个股: 上述所有技术指标的计算应用于每一个股票数据上，确保了分析的全面性和个体差异的考虑

特征图像示例

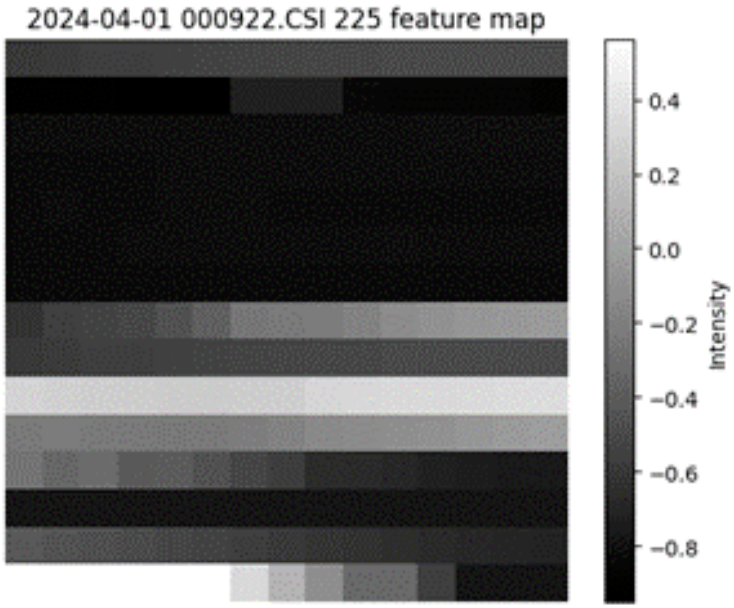

图 1: 2024 年 4 月 1 日中证红利的特征图像实例

我们采用了 **MinMaxScaler** 方法，对除标签和收盘价外的所有 DataFrame 特征实施了归一化，225 个特征，在模型首层处理后被整合成图像形式，作为深度学习流程中的关键输入。

在此基础之上，图像构建过程中的一个核心机制：相关指标的相邻排列及其对最终输出的潜在影响。通过引入 **Transformer** 架构，我们将前期卷积得到的特征进一步通过自注意力机制进行互相关处理，这一创新步骤理论上能够增强模型性能。

综上所述，对于每一天的市场信息，我们生成了一套结构化的输入数据，具体包括 [日期，操作标记，收盘价，以及经过转化的特征图像]。

```
1:  procedure LABELING
2:      设置窗口大小为 11 天
3:      初始化行计数器 counterRow ← 1
4:      while counterRow < 文件中的总天数 do
5:          counterRow ← counterRow + 1
6:          if counterRow > 窗口大小 then
7:              窗口起始索引 windowBeginIndex ← counterRow− 窗口大小
8:              窗口结束索引 windowEndIndex ← windowBeginIndex+ 窗口大小 −1
9:              窗口中间索引 windowMiddleIndex ← (windowBeginIndex + windowEndIndex)/2
10:             for all i ∈ [windowBeginIndex, windowEndIndex] do
11:                 number ← closePriceList[i]
12:                 if number < min then
13:                     min ← number
14:                     minIndex ← i
15:                 else if number > max then
16:                     max ← number
17:                     maxIndex ← i
18:             if maxIndex == windowMiddleIndex then
19:                 结果 ← " 卖出"
20:             else if minIndex == windowMiddleIndex then
21:                 结果 ← " 买入"
22:             else
23:                 结果 ← " 持有"
```

TECEC 架构概览

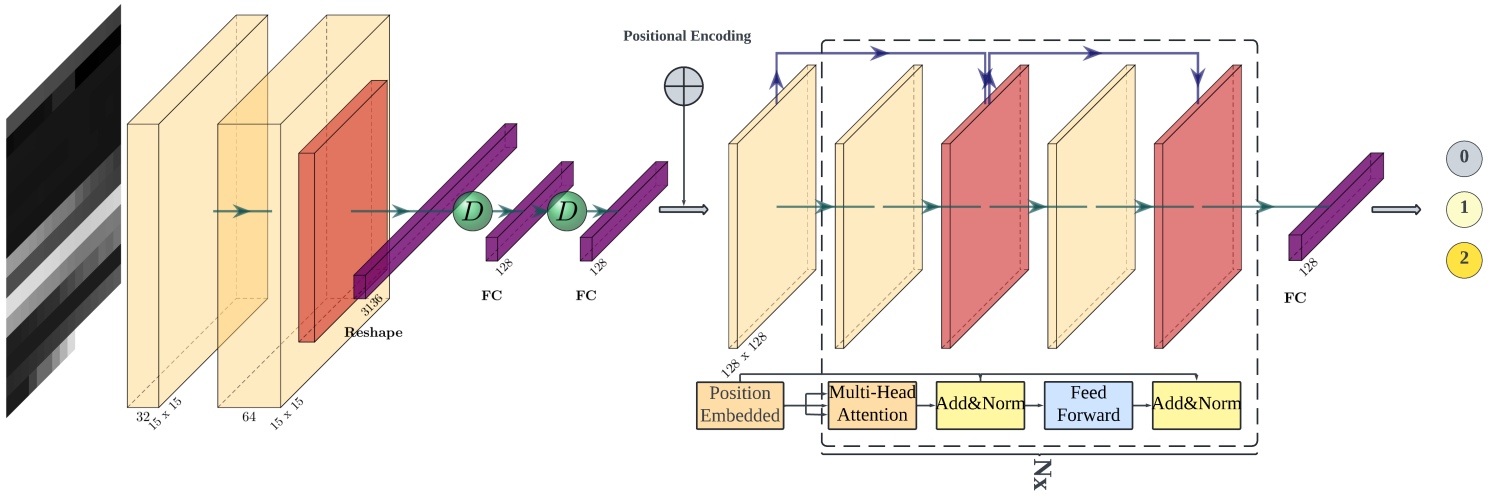

图 2: TECEC 架构图

- 流程始于一个简洁的重塑层（reshape_layer），它将一维的输入数据转换为适合二维卷积处理的格式，即 [1, input_w, input_h]，为后续的空间特征学习奠定基础

超参数的选择

- 实验所用数据集为 000922.CSI
- 测试参数为编码器层数 [1-5]、位置编码的维度 [128,256,512]、前馈神经网络的维度 [128,256,512] 以及多头注意力中的头数 [2,4,8,16] 共 180 种排列组合
- ACCURACY 为模型在训练过程中验证集上最好 10 次准确率的平均值

表 3: 超参数对比实验

LAYERS	D_MODELS	DFFS	HEADS	ACCURACY
1	32	64	16	83.8%
1	32	256	8	82.5%
2	128	256	2	**83.9%**
2	512	512	16	54.3%
3	128	128	16	77.7%
3	512	512	16	75.0%
4	128	512	16	68.7%
4	256	512	16	54.8%
...	...	...	...	...

- 超参数配置实验的结果明确指出，当模型参数设为 LAYERS=2,D_MODELS=128, DFFS=256, HEADS=2 时，系统展现了最为优越的性能，达到了 83.9% 的验证准确率，超越其他所有测试配置，被选定为本研究中 TECEC 模型的标准超参数组合。

对比模型的设计

- TECEC---本文提出的模型
  - CNN—--卷积神经网络
- MLP—--多层感知机，全连接网络
- LSTM—-长短期记忆神经网络
  - RNN—--循环神经网络

实验流程

- 平台与配置：依托 PyTorch，集成数据加载、Adam 优化及交叉熵损失，搭建训练环境，配置日志跟踪。
- 迭代训练：初始化后，利用数据加载器驱动训练与验证循环，多轮（epoch）迭代，每批数据训练并更新权重，实时监控进展。
- 验证与调优：周期性验证模型，记录性能指标，指导参数微调，确保无梯度干扰下准确评估。
- 综合评估：借助混淆矩阵、ROC 曲线和 AUC 值，细致分析模型性能，深挖分类精确率（精确率反映的是模型预测为正例的结果中真正为正例的比例）、召回率（衡量的是模型识别出的正例占所有实际正例的比例）与 F1 分数（F1 分数是 Precision 和 Recall 的调和平均值，旨在提供一个综合的度量）。
- 优化导向：聚焦交叉熵损失减少，提升分类准确率。
- 策略模拟：依据模型预测，执行交易策略模拟（买入、卖出、持有），实证模型实战价值。

TECEC 计算性能

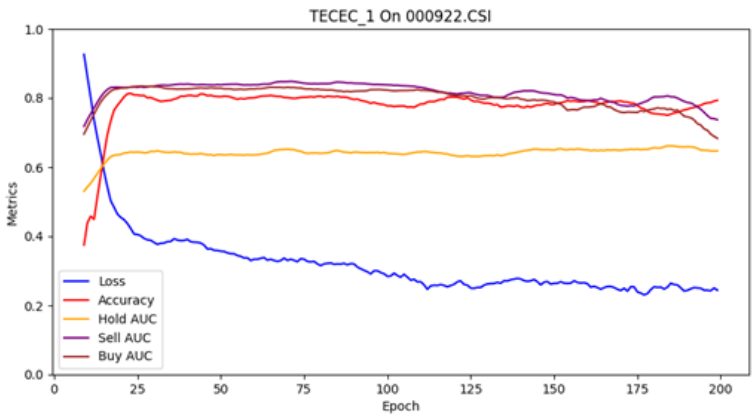

图 3: TECEC 模型性能表现

- 采用 One-vs-Rest (OvR) 策略对于 K 个类别。每次训练时，将一个类别视为正类，其余所有类别合并视为负类，据此计算每个类别的 AUC
- Accuracy 曲线是模型在测试集上的准确率
- 模型训练期间 Loss 值稳步下降，反映出模型逐步优化损失函数的有效性；Accuracy 初呈上扬后略有回调，暗示模型初期快速学习并优化预测能力，但后续遭遇轻微过拟合迹象
- Hold AUC 维持平稳水平，彰显模型在判定持仓决策方面的稳健性；Sell AUC 从低点逐步攀升，体现了模型识别卖出机会能力的提升；而 Buy AUC 初期高位后下滑，可能揭示模型在买入信号判断上存在过拟合或外界干扰问题

TECEC 模型混淆矩阵

表 4-2 TECEC 模型的混淆矩阵（中证红利）

		预测值		
		HOLD	BUY	SELL
实际值	HOLD	22668	2316	1816
	BUY	997	703	0
	SELL	1009	0	491

表 4-3 TECEC 模型的分类报告（中证红利）

总体准确率：0.7954			
	HOLD	BUY	SELL
Precision	0.92	0.23	0.21
Recall	0.85	0.41	0.33
F1 Score	0.88	0.30	0.26

图 4: TECEC（中证红利）

表 4-4 TECEC 模型的混淆矩阵（中证 1000）

		预测值		
		HOLD	BUY	SELL
实际值	HOLD	22147	2416	2237
	BUY	925	775	0
	SELL	732	0	768

表 4-5 TECEC 模型的分类报告（中证 1000）

总体准确率：0.7897			
	HOLD	BUY	SELL
Precision	0.93	0.26	0.24
Recall	0.83	0.51	0.46
F1 Score	0.88	0.34	0.32

图 5: TECEC（中证 1000）

TECEC 模型财务评估

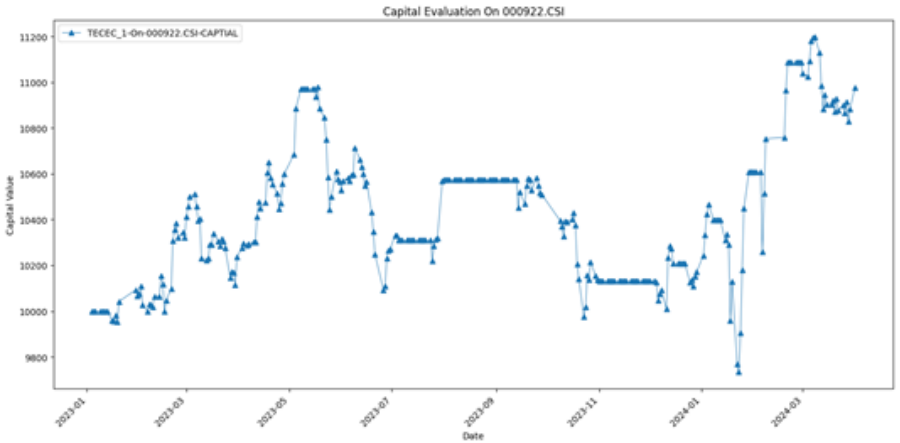

图 6: TECEC 财务评估

- 年化收益率（AR）:9.77%
- 尽管 9.77% 的年化收益率初看似属温和，但置于当前市场环境及对照无风险利率背景分析，此成绩实则彰显出模型超越常规投资策略的稳健增值能力
- 成功交易比例（PoS）:29.79%
- 高达 29.79% 的成功交易比例，标志着模型在近三成的交易决策中精准获利，这一成就远超随机交易的平均水平，充分验证了模型甄别并把握盈利机遇的高效率

其他模型性能

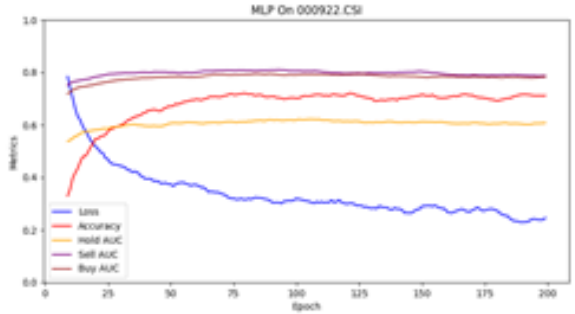

图 7: MLP

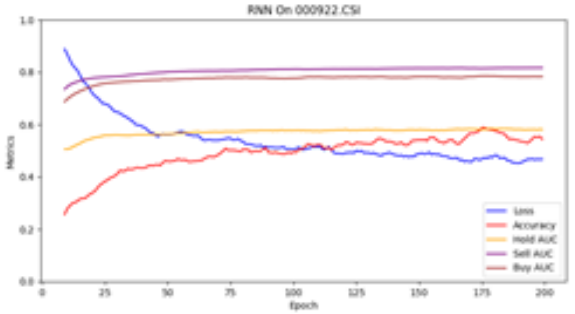

图 9: RNN

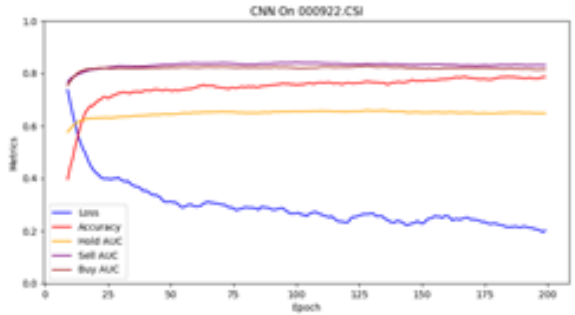

图 8: CNN

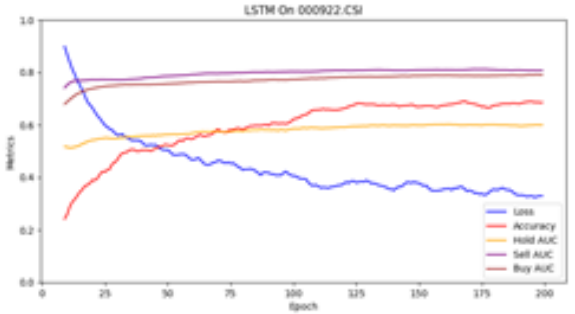

图 10: LSTM

模型性能对比与分析

- 性能亮点：TECEC 模型与基础 CNN 模型展现卓越准确率，稳定维持在约 80
- 决策敏锐性：所有模型于"HOLD" 状态的 AUC 值集中于 0.6 附近，而"SELL" 与"BUY" 决策的 AUC 则聚拢于 0.8 上下，揭示模型在辨别交易时机上的高度敏感性。
- **TECEC** 深入剖析：尽管 TECEC 总准确率显著占先，其分类混淆情形较常见，AUC 未全面顶尖。表明该模型虽宏观预测准确，但在精细交易信号分类上留有提升空间，需进一步细化评估。
- 后续展望：即将呈现的详细测试数据将进一步解析各模型在具体分类任务的表现，指导未来模型优化与策略调整的方向。

各模型的表现汇总

- 实验所用数据集为 000922.CSI
- 测试模型为前文所述五种

表 4: 各模型在 000922.CSI 上的性能表现

测试指标	MLP	CNN	LSTM	RNN	TECEC
分类准确率	0.716	0.780	0.703	0.460	0.806
AUC-持有	0.612	0.640	0.603	0.574	0.670
AUC-购入	0.785	0.807	0.789	0.781	0.801
AUC-卖出	0.785	0.825	0.811	0.816	0.832

- 在分类准确率方面，TECEC 模型以 0.806 的高分位居榜首，显著优于其他模型，如 MLP（0.716）、CNN（0.780）、LSTM（0.703）和 RNN（0.460）。这表明 TECEC 在综合预测准确性上具有明显优势
- 关于持有状态（AUC-持有）的评估，TECEC 模型取得了 0.670 的 AUC 值，单项最高，在所有模型中表现稳健，体现出在“持有”决策场景下良好的区分能力
- 购入（AUC-购入）和卖出（AUC-卖出）决策的评估中，TECEC 模型分别以 0.801 和 0.832 的 AUC 值独占鳌头
- 为了深化对模型效能的理解，接下来我们将扩展至多个股票指数的财务评估，这不仅有助于全面审视模型在实际市场条件下的应用潜力

各模型财务评估

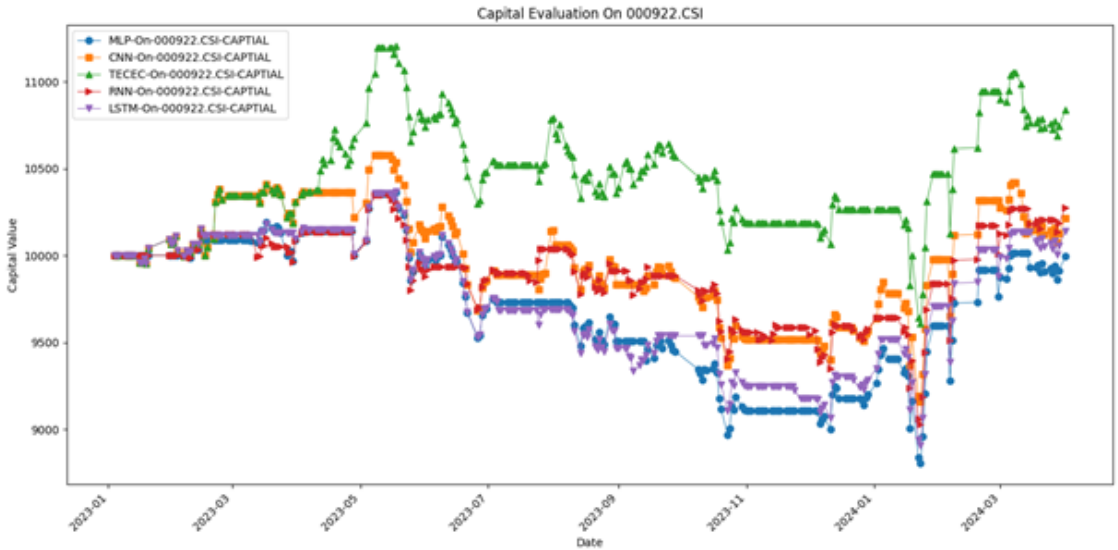

图 11: 各模型财务评估

- **TECEC** 模型卓越表现：在 000922.CSI 指数策略测试中，TECEC 模型凭借其显著的 8.36% 年化收益率（AR）脱颖而出，领先于其他模型。
- TECEC 不仅收益领先，还实现了 45.24% 的成功交易比例，体现了收益与交易质量的高效平衡。
- 相比之下，CNN、LSTM、RNN 及 MLP 模型的年化收益率分别为 2.15%、1.42%、2.75% 及-0.02%，均未能超越 TECEC 模型的收益水平，尤其 MLP 模型在测试期间未能取得正向收益，突显出 TECEC 在策略测试中的显著优势。
- 为进一步验证 TECEC 模型的稳定性和普适性，对其在多元数据集上的表现进行深入分析。

年回报率汇总

- 实验所用数据集为 000922.CSI
- 测试模型为前文所述五种

表 5: 年回报率汇总

测试集	MLP	CNN	LSTM	RNN	TECEC
000700.SZ	54.37%	19.01%	31.42%	52.91%	69.06%
000905.SH	-2.05%	-9.35%	-5.40%	-3.31%	5.55%
000015.SH	2.85%	2.26%	0.14%	-1.34%	-1.12%
000852.SH	-23.67%	-35.67%	-5.40%	-6.67%	5.67%
000905.SH	-15.00%	-33.33%	-15.33%	19.67%	-12.33%
000927.CSI	-32.67%	-3.67%	5.35%	-0.29%	-16.33%
399006.SZ	-11.33%	-16.00%	-19.00%	0.67%	-8.00%
000922.CSI	-0.98%	6.40%	-0.18%	1.96%	9.26%

回报率分析

- 跨市场卓越性：表 4-8 汇总数据显示，TECEC 模型在多测试集，包括沪深 300、中证 500 指数等，展示出极强适应力与盈利潜力，年回报率显著高于 MLP、CNN、LSTM、RNN 等模型。
- 领航表现：特别地，在 000700.SZ 数据集中，TECEC 实现 69.06% 的高额年回报，且在市场不佳如 000905.SH 时，保持正回报，体现其逆境中的防御实力。
- 双面优势：TECEC 在市场上涨时有效增值，下跌时减少损失，证明其在机会捕捉与风险控制的双重能力，符合长期投资战略需求。
- 前沿地位：TECEC 模型在复杂市场条件下的持续稳健表现，巩固了其在财经预测与投资决策的领先地位，指明 Transformer 架构的巨大潜力。
- 未来研究导向：深化 Transformer 模型优化与市场适应性参数调整，目标提升预测精度与投资效益，强化智能决策支持。
- 架构优势对比：与 CNN 等模型对比，Transformer（以 TECEC 为例）在时间序列预测，尤其中国股市复杂情境下，展现更优性能，确保高准确率与稳定收益增长。
- 未来模型构建策略：融合或倾向采用 Transformer 架构，视为提升金融预测模型性能与实用性的核心路径，旨在提供更可靠的投资智能化工具。

TECEC 模型的卓越性能与社会影响

- 卓越预测效能：TECEC 模型在多种评价指标上超越了 MLP、CNN、LSTM、RNN，特别是在分类准确率（80.6%）和买卖决策的 AUC 值（0.801 和 0.832），展示了其在交易策略预测上的精确性。
- 交易节点辨识：模型在识别“买入”与“卖出”时机上的优异表现，强调了其在构建高效交易系统中的重要性。
- 财务验证：实证研究中，TECEC 模型在中证红利指数测试中取得 9.26% 年化收益率和 45.24% 成功交易比例，证明了其盈利转化能力。
- 市场适应性：模型展现的广泛适应性和稳定性，证明其在不同市场条件下的有效性和风险控制机制。
- 社会责任考量：从社会、健康、安全、法律、文化及环境多维度评估，强调了模型对市场效率、投资者心理健康、数据隐私保护、公平性及环保意识的贡献。

未来展望：金融预测模型的创新与整合

- 技术深化与创新：继续优化现有模型，探索多模型融合、多模态数据集成，以及预测系统的实时适应性和风险量化管理。
- 模型精细化与集成：通过自动调参、神经架构搜索强化模型性能，集成学习策略融合 Transformer、CNN、LSTM 等算法优势。
- 多元化信息整合：纳入宏观经济指标、新闻情绪、社交媒体信号，增强跨模态学习框架，提升模型全面预测能力。
- 实战应用与风险管理：实现模型自我更新，快速响应市场变化，量化预测不确定性，融入风险管理策略。
- 可持续发展与社会责任：注重 AI 技术的环境影响，开发低碳算法，利用可再生能源，纳入 ESG 指标，推动绿色金融市场建设。

Thanks!