

# 卷积神经网络能否预测中国股市 ——基于大单净流入率的改进方法

方雨桥<sup>1</sup>, 郑行健<sup>2</sup>, 李峰<sup>\*</sup>

<sup>1, 2, \*</sup> 上海交通大学, 上海高级金融学院

## 摘要

基于Jiang et al. (2023), 本文提出一种简单却高效使用卷积神经网络 (CNN) 在中国股票市场中预测收益率的方法。通过将股票根据交易者大单净流入率高低分组, 用全连接层作为输出层训练不同模型, CNN 能够更好地挖掘不同组别内投资者潜在交易模式, 从而显著提高策略效果。进一步的异质性研究发现, 分组训练后的 CNN 能够识别出散户和机构投资者对于色彩、量价关系、图形技术指标的不同行为模式, 不仅增强了预测能力, 也提高了模型的可解释性。稳健性检验显示, 无论选择何种分组指标, 改进后的 CNN 策略组合都会有显著的因子异象  $\alpha$ 。本文提出的基于投资者行为模式分组的训练方式也能广泛应用于其它类似机器学习模型, 提升模型拟合能力, 进而实现更高的超额收益。

**关键词:** 卷积神经网络, 股票预测, 大单净流入率, 技术分析

# 目录

<b>1 引言</b>	<b>3</b>
<b>2 文献综述</b>	<b>5</b>
<b>3 模型与数据</b>	<b>6</b>
3.1 卷积神经网络结构 . . . . .	6
3.2 研究数据与训练方法 . . . . .	9
<b>4 预测结果</b>	<b>11</b>
4.1 基准结果 . . . . .	11
4.2 Fama French 五因子 . . . . .	15
4.3 基本技术指标 . . . . .	16
4.4 其他分组指标 . . . . .	16
<b>5 异质性研究</b>	<b>18</b>
5.1 颜色视觉效应 . . . . .	18
5.2 神经网络的注意力分布 . . . . .	19
5.3 图型技术指标 . . . . .	23
<b>6 结论</b>	<b>25</b>
<b>A 大单净流入率截面相关性及自相关性</b>	<b>33</b>
<b>B 像素图构造</b>	<b>33</b>
<b>C 模型预测结果的稳健性检验</b>	<b>34</b>
<b>D 基本技术指标定义及收益率</b>	<b>34</b>
<b>E 颜色特征指标描述性统计</b>	<b>40</b>
<b>F 光滑梯度类别激活映射 ++ 模型</b>	<b>40</b>
<b>G 图型技术指标构造</b>	<b>47</b>

# 1 引言

K 线图 (*Candlestick Chart*) 是一种在金融市场中常用的图表类型，用于反映一段时间内资产价格的变动情况。在来自业界的经验中，大部分个人投资者和少数专业的机构投资者往往都会参考 K 线图走势，运用技术分析以执行交易决策 (赵鹏, 2015; McBride, 2001; Sindreu, 2020)。K 线图虽然并没有显性地包含每只股票的基本面信息或量化因子信息，但依旧具有非常强的交易引导作用。然而，如何使视觉上的 K 线图被数据化，一直是个悬而未决的难题，直到 O'Shea and Nash (2015) 首次引入一种新的机器学习方法——卷积神经网络 (*convolution neural network, CNN*)，并由 Jiang et al. (2023) 应用在股票市场中，“图”的重要性才逐步得到学术界的证实。

虽然 Jiang et al. (2023) 的研究发现基于 CNN 的交易策略在成熟的美国市场中具有显著的预测能力，但是在由散户占据主导地位、且更依赖于 K 线图和技术分析的中国市场中却并不显著。CNN 作为一种强大的深度学习模型，在中国市场中应当能够通过深度挖掘投资者非理性交易行为，获得较发达市场中更高的超额收益 (Leippold et al., 2022)。作为一种高度依赖于投资者交易行为的机器学习策略，CNN 在训练和测试时，需要样本 K 线图中的交易都服从同一种交易模式，即不同股票在拥有相同的 K 线图模式时具有相同的未来收益率。然而，本文认为这一假设在中国市场中并不成立，从而导致不同模式之间存在较大的泛化误差而降低模型的预测效果。

在实际交易中，不同投资者在相同的 K 线图形态下往往会有不同的行为，进而影响未来的股票收益率走向。例如，在市场整体暴跌时，散户往往会由于恐慌而大幅抛售，导致股价进一步下跌；而专业的机构投资者则可能会逆向投资，在攫取流动性溢价的同时甚至会推动股价上涨。于是，由散户主导交易的股票和由机构主导交易的股票在未来具有了不同的价格走势，即使它们呈现出相同的 K 线图形态。考虑到中国市场投资者结构复杂，散户占比高 (郑振龙 and 孙清泉, 2013)，加之中国股市的诸多限制如卖空约束 (苏冬蔚 and 倪博, 2018)、 $T+1$  (张兵, 2020)、涨跌停板制 (汪天都 and 孙谦, 2018)，中国股票市场的完整程度 (*Market Completeness*) 要弱于美国市场。因此，不同股票隐含不同交易模式的情况变得更加不可忽略，使得直接使用 CNN 可能会由于泛化误差过大而效果不佳。那么，是否可以通过区分这种隐式交易模式从而提高 CNN 的预测表现？

为了探究不同隐式交易模式的区分是否能提升 CNN 在中国股票市场中的预测能力，本文以每个时间截面中每只股票的大单净流入率作为该股票当日的交易群体的代理变量，并据此进行分组训练。实证研究发现，通过将股票按大单净流入率从低至高分为 5 组训练，训练的效果得到了大幅提升，并且交易模式越单一的组别训练效果提升越显著：对于大单率最小 (Q1) 和最大 (Q5) 组，训练提升效果最好，相较于随机混合组的多空组合收益提升高达每周 0.28% 和 0.39%。与之相对地，交易模式比较复杂、不同投资者都混合在一起的中间组 (Q3) 甚至有了更差的表现，每周的收益率降低了 0.12%。除 Q3 外，相较于没有分组的结果而言，多空组合的周度收益率最低提升了 0.11%、最高提升了 0.39%。均等持仓其余的四组 (Q1、Q2、Q4、Q5) 后，市值加权的投资组合的夏普比率从 0.75 提升到 2.88，多头投资组合的夏普比率从 0.74 提升到 1.87。

通过比较基于大单净流入率分组前后策略收益率的变化，本文发现 CNN 能够识别投资者的潜在交易模式，并且能够获得超出 Jiang et al. (2023) 的投资组合表现。基于因子回归的研究结果发现，CNN

策略的收益率并不能被传统的风险因子如 Fama French 五因子、动量与反转因子所解释。此外，在我们研究的众多技术指标中，短期动量、动量震荡、交易量指数、历史波动率、真实震荡幅度和交易量加权平均收盘价指标都对 CNN 资产组合的收益率有预测能力，且不同交易组别在不同技术指标上有不同因子载荷，而这也从侧面证明了技术指标在中国市场中的实践意义。对于投资者聚集程度越高的组别（如 Q5），风险因子载荷和技术指标的特征越显著。

为了厘清预测能力提升的机制，我们进一步探索不同组别之间 CNN 学习内容的异质性，换言之，CNN 是否能在不同分组中挖掘出投资者交易特征的异质性。为此，本文还对输入 K 线图的颜色特征、CNN 学习时的注意力特征、以及 Lo et al. (2000) 提出的八个图型技术指标进行了研究。具体而言，我们发现 CNN 捕捉了散户投资者的交易模式——当 K 线图上有大面积的绿色时，散户组（Q1 和 Q2）的 CNN 交易策略能够有更显著的超额收益，而这一现象对于机构组（Q4 和 Q5）并不显著。这是由于散户往往表现出追涨杀跌、依赖 K 线图的视交易行为 (Odean, 1998, 1999; Barber and Odean, 2008; Barber et al., 2009, 2022)，而机构普遍依赖更加复杂的技术指标进行交易决策。在 K 线图更深层的信息上，例如，对于 K 线图上不同区域的注意力、图型技术指标如头肩顶形态、三角底上，机构比散户具有更高的敏感性，机构交易行为在复杂技术指标特征中的表现更为明显。这些异质性结果表明，对于同样的 K 线图，散户与机构的判断和交易行为仍旧有很大的不同，Jiang et al. (2023) 将其放在同一批样本中试图学习出相同的网络参数，会使得网络模型的失真，从而导致在中国市场中的预测结果不显著。

在研究 CNN 学习到的因子、技术指标以在不同组别上存在显著异质性时，本文的实证结论除了证明 CNN 能精准高效地挖掘出投资者交易模式异质性，大幅提升 Jiang et al. (2023) 的预测结果之外，也为技术分析在金融市场中的预测能力提供了来自机器学习领域的的新证据。在之前的研究中，学者们基于构造因子或线性回归的传统方法对于技术分析是否能预测股票收益尚未达成共识 (Park and Irwin, 2007)，部分研究支持技术分析有预测能力 (Caginalp and Laurent, 1998; Lo et al., 2000; Osler, 2003; Murray et al., 2024)，但也有研究认为技术分析并无显著预测能力 (Marshall et al., 2006; Bajgrowicz and Scaillet, 2012)。就中国市场中技术分析是否能帮助预测股票收益的研究中，本文在李斌 et al. (2017) 和林耀虎 et al. (2022) 的基础之上，证明了股票走势的“形态”，如“头肩”、“三角顶”、“矩形顶” (Lo et al., 2000)，都对股票收益率确实具有一定的预测能力。本文的异质性结果也说明了交易金额更大的机构投资者同样也关注股票 K 线图，并非完全基于基本面风格、拥有充分信息和认知资源的理性投资 (Bailey et al., 2011; Akepanidtaworn et al., 2023; Ben-Rephael et al., 2017; 李斌 et al., 2019)。通过尝试更多不同种类的因子或技术指标，结合更先进的分析算法，本文在一定程度上进一步打开了神经网络的“黑箱子”。

本文采用的这种分类训练以提升效果的方式，不同于传统上更改卷积神经网络结构的方式 (Lu et al., 2020; Liu et al., 2017)，更加依赖于模型背后的行为金融学含义。因此，这种方式在具有相当的稳健性的同时也有很强的迁移属性——并不仅仅局限于卷积神经网络，也可以运用于其他机器学习模型 (姜富伟 et al., 2022; Gu et al., 2020; Rossi, 2018; Chen et al., 2024)。与本文最接近的 Lu and Wu (2022) 将 Jiang et al. (2023) 中的卷积神经网络进行改进，将输入的黑白 K 线图转为彩色 K 线图，并加上了复杂的注意力机制，将多头资产组合夏普比率提高到了 1.15。Lu and Wu (2022) 采用的注意力机制在 CNN 模型上增加了五层的全连接层用于处理图像信息之外的数据信息，增大了模型的参数量的输

入的数据量从而增加了信息的重复度和训练负担，在技术上更加复杂的同时却导致泛用性上有所不足。本文采取的方式并没有引入其他参数或增加卷积神经网络需要的数据量，在大幅提升预测能力时没有增加额外算力要求，同时也能够得出 (Lu and wu, 2022) 中散户在图像上具有更高的注意力的结论。

本文剩余部分如下：章节2介绍了神经网络在金融领域的应用以及交易者行为差异相关文献；章节3介绍了本文采用的神经网络模型和误差机理；章节3.2介绍了本文使用的数据以及相关性质；章节4介绍了每个股票池中，卷积神经网络模型的预测效果；章节5解释了卷积神经网络在学习不同股票池时的异质性来源；章节6总结本文。

## 2 文献综述

相较于其他机器学习算法，神经网络以其优秀的处理复杂模式、自动提取特征、泛化等能力被广泛应用在金融预测方面。李斌 et al. (2017) 使用多种机器学习算法处理技术指标并发现神经网络的表现优于线性模型、支持向量机 (*support vector machine, SVM*) 等模型的表现。李斌 et al. (2019) 使用人工神经网络 (*artificial neural network, ANN*) 和循环神经网络 (*recurrent neural network, RNN*) 进行基本面量化投资，表现显著优于了 OLS、全连接神经网络、岭回归、LASSO 等线性机器学习算法。卷积神经网络作为神经网络中的一种，其优势在于提取图像中的局部特征，这是其他神经网络无法实现的功能。在金融预测方面，被普遍用于处理各种各样的金融图像。Chen et al. (2016) 将台湾股指期货收盘价的时间序列数据转换成不同的图像，包括格拉姆角场 (*Gramian angular field, GAF*)、移动平均映射 (*moving average mapping, MAM*)、双移动平均映射 (*double moving average mapping, DMAM*)、以及蜡烛图 (*candlestick figure*)。不仅是图像处理，卷积神经网络也广泛用于非图像特征的拟合方面。王刚 et al. (2023) 直接使用时间序列上欧美、美日、美元人民币汇率收盘价构造移动窗口下的矩阵输入到神经网络中，得到了较好的预测效果。

卷积神经网络与其他机器学习模型有很强的联动性，其中尤以与长短期记忆 (*long short term memory, LSTM*) 模型的结合为主。在 CNN-LSTM 模型下，景楠 et al. (2020) 使用价格数据和技术指标数据预测沪铜期货高频价格；汪刘凯 et al. (2023) 使用常见供应链金融质押物的混频数据，相较于普通的神经网络提高了泛用性。与本文更加接近的是 Liu et al. (2017)，其在量化投资和股票选择策略中使用 CNN-LSTM，其表现要优于普通的动量策略。不仅是美国金融市场，其他国家的市场也已经有神经网络的相关研究。Lu et al. (2020) 使用中国 A 股上证指数数据，包括日频的开盘价、最高价、最低价、和收盘价，以及交易量等量价数据进行预测，预测结果显著优于多参数线性规划 (*multiparametric linear programming, MLP*)、CNN、RNN、LSTM 和 CNN-RNN。类似地，陈凯杰 et al. (2022) 将 CNN-LSTM 改进到 CNN-BiLSTM 以预测股票市场收益率变化。在实际应用中，CNN 可以连接其他不同的机器学习模型，而并非局限于 LSTM，例如 RNN 或 SVM (Zhang et al., 2018; Cao and Wang, 2019)，也能很好的预测股票未来收益率。

在使用神经网络预测资产价格的研究中，研究者们始终在通过改进神经网络的模型结构或者数据的输入形式来提升股票预测能力，而这种纯技术层面的改进没有办法克服金融层面带来的问题。本文参考散户与机构的异质性将样本进行分类从而使得训练样本具有同质性。赵涛 and 郑祖玄 (2002); 张乐

and 李好好 (2008) 指出，散户虽然数量多，但占据的市场份额并不大，散户投资金额小，不能进行大规模投资。更重要的是，散户容易受到市场噪音与短期行情的影响而做出盲目的买卖决策，所以也不会积累某一方向的头寸从而影响市场。Kirilenko et al. (2017) 研究了各种机构投资者的行为以及持仓调仓情况。基本面投资者、投机投资者、高频投资者以及做市商都属于机构投资者。基本面和投机投资者会积累头寸；高频投资这和做市商虽然不会积累头寸，但会有非常高的调仓频率。其行为与散户有显著差异。即使同为机构投资者，也会有很强的异质性。不同投资者的行为是显著不同的，一般而言，机构投资者由于兼备信息渠道广，信息处理能力强更接近完全市场假设下的理性投资者 (尹海员 and 朱旭, 2022)。同时，对于本文采用的分类准则——大单净流入率，已有文献研究其导致的股票市场的异质性。许泳昊 et al. (2022) 发现“大单异象”的存在，即大单买入量与预期收益的显著负相关关系，并且这种关系在低机构持仓的股票中更为显著。这一发现强调了不同大单净流入率对未来收益的影响。万谍 and 杨晓光 (2019) 发现小、中、大交易订单具有不同的信息含量，其中中单的信息含量最高，证明了具有不同订单大小的股票的异质性。

### 3 模型与数据

#### 3.1 卷积神经网络结构

本文采用经典的卷积神经网络结构，如图1所示。卷积神经网络 CNN 模型的输入为“嵌入”后的 K 线图，输出则为模型预测的股票未来上涨的概率。我们通过训练 CNN 模型，使得其能够理解 K 线历史走势和未来上涨的关系并做出预测。

具体来看，模型的结构中左侧五个立方体为一个输入层和四个卷积层，其中，由于输入为 RGB 三色图，因此输入层的尺寸为  $3 \times 96 \times 180$ ，而四个卷积层深度分别为 64、128、256、512。在每次卷积前，输入张量的边缘会被填充以使得卷积后的输出尺寸等于输入尺寸，再进入  $2 \times 1$  的池化层。因此，每层宽度均为 180，而长度每次都为池化前的一半，即 48、24、12、6。右侧的三个柱状体为两个全连接层和输出层。第一个全连接层为把最后一个卷积层平铺后的向量，尺寸为  $512 \times 6 \times 180 = 552960$ 。第二个全连接层的尺寸为 1024。由于本文采用的样本标签为二分类，所以输出层的尺寸为 2，输出值为  $(1 - \hat{y}, \hat{y})$ ， $\hat{y}$  为模型认为该股票未来五天会涨的概率。本文选择交叉熵作为模型的损失函数：

$$\text{CrossEntropy} = -y \log \hat{y} - (1 - y) \log (1 - \hat{y}) \quad (1)$$

其中  $y$  为样本标签，即未来五天收益率的涨 ( $y = 1$ ) 或跌 ( $y = 0$ )。本文考虑到传统投资组合往往是由空头和多头构成，导致股票相较于截面的收益率大小比起真实收益率具有更高的信噪比。于是，本文将截面上每一个股票池的 5 日累计收益率  $R_{i,t}$  市值加权平均，得到该股票池的市场 5 日收益率  $R_{m,t}$ 。然后将每一只股票的 5 日累计收益率与股票池的市场 5 日收益率作差，得到超额收益率  $R_{i,t}^e$ 。同时，从模型训练的角度来说，还需要保持 0,1 标签的平衡。因此，每张输入的像素图的标签为

$$y_{i,t} = \begin{cases} 1 & R_{i,t}^e > \text{median}(R_{i,t}^e); \\ 0 & R_{i,t}^e \leq \text{median}(R_{i,t}^e). \end{cases} \quad (2)$$

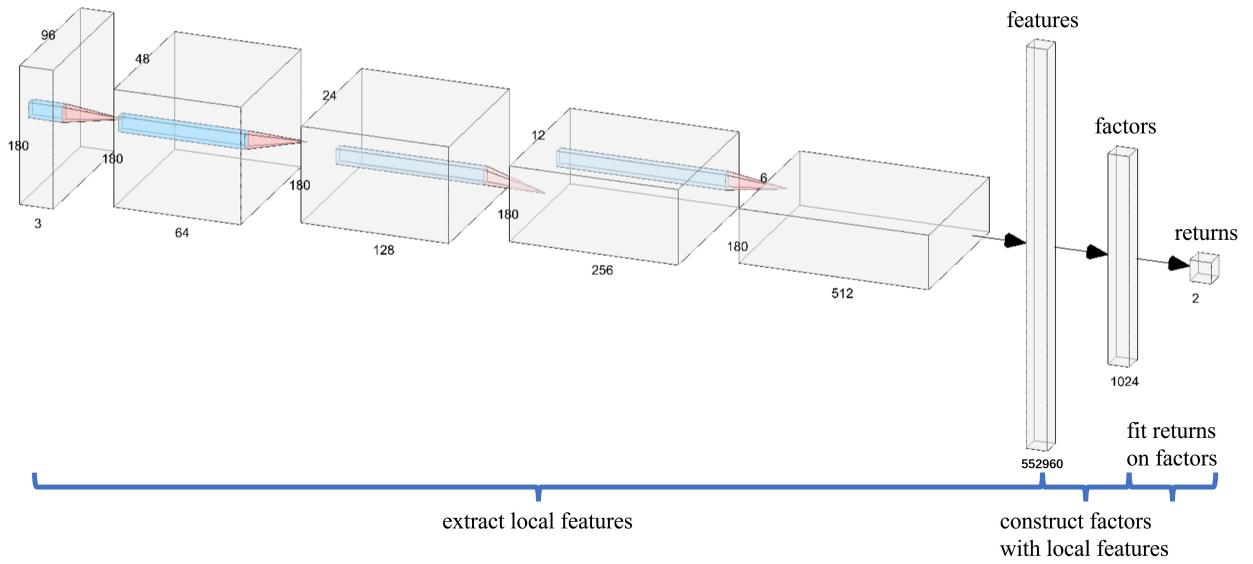


图 1: 卷积神经网络结构

图中绘出了本文采用的卷积神经网络结构。左侧的五个立方体为一个输入层和四个卷积层，右侧的三个柱状体为两个全连接层和输出层。两个全连接层分别代表从图中提取的 552960 个局部特征用这些特征构造的 1024 个因子。输出层代表因子通过线性的神经网络和激活函数后的股票涨跌概率预测。

在整个模型中，卷积神经网络的不同部分负责了不同的功能。第一个输入层是将我们肉眼看到的 K 线图转化成机器能够理解的嵌入表达 RGB 像素图。作为神经网络的输入，开盘价、高价、低价、收盘价、移动均线、和交易量需要被转化成机器能够读取如图2的 RGB 像素图，其具体构造方式见附录B。图2中，左图是转换后的日内像素图，右图是与其相对应的日内收益率表。左图中，线段颜色越深，意味着当日股价涨跌幅度越大。此外，由于本文采用三通道图像处理技术，使得移动均线和交易量上也有 RGB 色彩，图像中蕴含的信息相较于 Jiang et al. (2023) 中的信息更加丰富。

后续四个卷积层则用于提取局部特征，每一层的深度代表着卷积核的数量，不同的卷积核代表不同的特征。由于卷积本身是一种求相关性的操作，因此四个卷积层最终得到的结果为输入的 RGB 图关于这些特征的相关性分布。将这些相关性的值平铺成第一个全连接层，代表着从图中得到的 552960 个初始特征。从第一个全连接层到第二个全连接层，代表这些初始特征通过线性运算并通过非线性激活函数被构造成 1024 个因子。从第二个全连接层到输出层，代表着收益率未来 5 天上涨概率与这 1024 个因子的拟合。

本文分组训练 CNN 模型后能提升模型预测能力的底层逻辑源自于最基本的资产定价理论 (*capital asset pricing model, CAPM*)：假定市场中有  $K$  个因子包含在价量信息中，对于每个截面上的  $N$  只股票，股票  $i$  在  $t$  时刻的超额收益率为：



图 2: 像素图

图中给出了输入像素图的样例 (a) 以及其色谱对应的 RGB 值和日内收益率 (b)。像素图 (a) 的上 75% 为价格像素图, 下 25% 为交易量像素图。色谱 (b) 右侧第一列为 RGB 向量三个元素对应的值, 也为卷积神经网络三通道输入每个通道的输入值, 第二列为颜色对应的日内收益率, 即收盘价相较于开盘价的增长率。

$$R_{i,t}^e = \beta_i^T f_t + \varepsilon_{i,t} \quad (3)$$

$f_t$  便是 CNN 模型第二个全连接层网络中得到的 1024 个因子。其中,  $\sigma_i^2 = \text{Var} [\varepsilon_{i,t}]$ ,  $t$  时刻的因子为  $f_t = [f_{1,t}, f_{2,t}, \dots, f_{K,t}]^T$ , 股票  $i$  的因子载荷为  $\beta_i = [\beta_{i,1}(\theta_i), \beta_{i,2}(\theta_i), \dots, \beta_{i,K}(\theta_i)]^T$ 。 $f_t$  是标准正交化后的因子, 即对于  $f = [f_1, f_2, \dots, f_T]$ ,  $ff^T = I_{K \times K}$ 。因子载荷  $\beta_i = \beta(\theta_i)$  是状态变量  $\theta_i$  的函数。在最小化均方误差 (*mean square error, MSE*) 估计下, 目标函数为:

$$\min_{\beta} \frac{1}{N \times T} \sum_{i,t} (\beta_i^T f_t - R_{i,t}^e)^2 \quad (4)$$

在  $\mathbb{E}[f_t \varepsilon_{i,t}] = 0, \forall i$  的假设下, 4 的一阶条件为  $\hat{\beta} = \frac{\sum_i \beta_i}{N} = \bar{\beta}$ 。在训练集、验证集、测试集同分布、 $\varepsilon_{i,t}$  时间序列上独立同分布、 $\beta_{i,k}$  互相独立的假设下, 最优化后的 *MSE* 可以用于衡量模型预测收益率的表现:

$$MSE = \mathbb{E}[\sigma_i^2] + \sum_{k=1}^K \text{Var}[\beta_{i,k}] \quad (5)$$

[Lu et al. \(2021\)](#) 指出, 机器学习的误差由三部分构成, 第一部分为优化误差 (*optimization error*), 即模型在优化时无法获取到真正的最优化的参数; 第二部分为泛化误差 (*generalized error*), 即样本具

有不同模型时，用相同的模型去描述所表现出的误差；第三部分为逼近误差 (*approximation error*)，即数据或模型本身含有的噪音，真模型曲线也无法经过所有的样本点带来的误差。在不考虑优化误差的情况下，逼近误差为股票池中，每只股票在时间序列上的模型残差的方差，即  $\sigma_i^2 = \text{Var}[\varepsilon_{i,t}]$ ，这种误差内生地来自于每只股票，是伴随着每只股票内生存在的特质误差；泛化误差为股票池中，每只股票的每个因子载荷在截面上的方差的总和，即  $\text{Var}[\beta_{i,k}]$ ，这种误差可以通过调整训练样本、提升样本集中特征相似度而降低。

在中国股市中，由于市场内交易主体异质性强、不同板块之间准入限制较高，导致不同股票的 K 线图往往是由不同投资者交易形成，而这也使得由 CNN 模型提取得到的交易特征拥有很强的异质性。例如，机构持股较高、高股价股票和散户持股较高、低股价股票之间有截然不同的交易特征。换言之，前者和后者所拥有的因子  $\beta_k$  不尽相同，如果将两者归为一类训练会使得模型泛化误差  $\text{Var}[\beta_{i,k}]$  过高且模型训练效果过差。在本文中，降低模型预测泛化误差主要通过对样本采取按大单净流入率分组的方式实现，并且后续的实证结果显示，分组训练对于模型有显著的效果提升。

### 3.2 研究数据与训练方法

本文使用中国 A 股主板 2013 年 1 月 1 日到 2022 年 12 月 31 日共十年的数据，将时间序列数据转化为图像数据后，共有 2345 个交易日。股票的日频交易数据来自于 CSMAR，主要包括开盘价、高价、低价、收盘价、交易量、流通市值和总市值。大单净流入率数据来自万得，单笔交易高于 100 万元的订单被定义为机构单，单笔交易高于 20 万的订单被定义为大户单，大单净流入率包含了大户单和机构单，对于  $t$  时刻股票  $i$  的大单净流入率被定义为

$$\text{LargeOrder}_{i,t} = \frac{\sum_k \mathbf{1}_{\text{Amount}_{i,t,k} \geq 20} \times \text{Vol}_{i,t,k}}{\sum_k \text{Vol}_{i,t,k}} \quad (6)$$

其中， $k$  为逐笔订单， $\text{Amount}_{i,t,k}$  为每一笔订单的交易金额， $\text{Vol}_{i,t,k}$  为每一笔订单的交易量。本文旨在通过大单净流入率划分判断交易者构成，因此本文所指大单净流入率均为绝对值而不考虑流入与流出。

在进行神经网络训练前，本文将研究样本分为训练集、验证集、和测试集，如图3所示。其中，训练集为第 1 到 1024 个交易日，验证集为第 1085 到 1340 个交易日，测试集为第 1401 到 2345 个交易日。为了避免数据集的相互重叠，每个集合之间设置了 60 个交易日的暂停区间。对于每个时间  $t$ ，截面上的股票根据大单净流入率的绝对值大小被分为 5 组，并随机采样一组作为对照 (Random 组)。每一组股票在训练集、验证集、和测试集上的描述性统计如表1所示，表中汇报内容均为组内均值。在不同组别中，虽然大单净流入率从低至高的差异非常显著，但是各组平均收益率、价格、或市值变量并无明显差异，这意味着这些变量与大单净流入率没有显著相关关系，由此可以排除规模或是价格、动量等因素带来的内生性问题。

作为本文重要的分组变量，本文需要首先证明大单净流入率不同于投资率  $INV$ 、账面市值比  $B/M$ 、净资产收益率  $ROE$ 、总资产收益率  $ROA$ 、杠杆率  $Leverage$ 、以及公司规模  $Size$ 。附录 A 表 9 给出了七个变量在时间截面上的相关系数的均值。可以看到，大单净流入率与其他六个变量间的相关关系显

表 1: 股票数据描述性统计

Training							Validation						
	Ret (%)	Large Order(%)	Price (RMB)	Capital (Billion)	#	Ret (%)	Large Order(%)	Price (RMB)	Capital (Billion)	#			
Q1	0.24	1.36	15.16	14.60	278	-0.71	1.36	15.68	15.35	435			
Q2	0.25	4.18	15.07	14.39	278	-0.72	4.30	15.50	14.84	435			
Q3	0.29	7.27	14.91	14.25	278	-0.70	7.57	15.10	15.45	435			
Q4	0.29	11.13	14.81	14.11	278	-0.72	11.81	14.29	15.76	435			
Q5	0.29	18.95	14.20	14.03	280	-0.79	20.91	12.46	15.68	437			
Random	0.27	8.60	14.85	14.29	278	-0.72	9.17	14.50	15.25	435			
Testing							Overall						
	Ret (%)	Large Order(%)	Price (RMB)	Capital (Billion)	#	Ret (%)	Large Order(%)	Price (RMB)	Capital (Billion)	#			
Q1	0.30	1.19	18.13	17.22	541	0.04	1.33	15.88	15.31	339			
Q2	0.32	3.71	17.33	17.20	541	0.05	4.11	15.63	15.07	339			
Q3	0.33	6.56	17.11	17.82	541	0.08	7.19	15.40	15.25	339			
Q4	0.35	10.27	16.36	18.57	541	0.08	11.10	15.01	15.39	339			
Q5	0.31	18.46	13.85	18.19	542	0.05	19.29	13.74	15.25	341			
Random	0.35	8.04	16.66	17.78	541	0.07	8.61	15.14	15.22	339			

表中汇报了每一组在训练集、验证集、测试集、以及全时间段里的平均周度收益率、大单净流入率、价格、市值、以及样本数在时间序列上的均值。从 Q1 到 Q5，大单净流入率依次递增。随机组作为对照，从每个截面上随机采样 1/5 的股票。

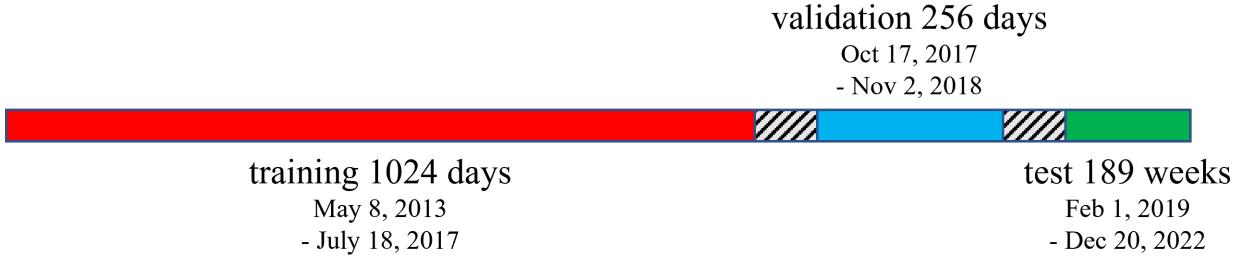


图 3: 训练集、验证集、测试集划分

图为训练集、验证集、测试集的示意图。训练集包含 1024 天，从 2013 年 5 月 8 日到 2017 年 7 月 18 日；验证集包含 256 天，从 2017 年 10 月 17 日到 2018 年 11 月 2 日。模型预测的是未来五天的收益率，因此策略为周度调仓，测试集为 189 周，从 2019 年 2 月 1 日到 2022 年 12 月 20 日。图中阴影部分为暂停区间，用于避免不同数据集之间的数据重叠和泄漏。

著小于与其他六个变量间的关系。因此，我们在根据大单净流入率排序时，能够排除由其他变量潜在的交互影响。此外，为了判断在每个五天的持仓周期中，股票的投资者构成是否稳定，本文还检验了大单净流入率的自相关性，对每只股票  $i$  进行回归：

$$LargeOrder_{i,t} = \sum_{lag=1}^{\tau} \beta_{i,lag} \times LargeOrder_{i,t-lag} + FixedEffect_i + \varepsilon_{i,t} \quad (7)$$

其中， $FixedEffect$  为固定效应， $\beta_{t-1} \sim \beta_{t-5}$  分别为 1 ~ 5 个时间滞后的系数。 $\bar{\beta}_{t-6 \sim t-20}$  为 6 ~ 20 个滞后系数的平均值。 $\tau = 1$ 、 $\tau = 5$  和  $\tau = 20$  的回归结果如附录 A 表 10，方括号中的  $t$  值为样本截面均值。回归结果显示，个股的大单净流入率有显著的正自相关性，并且这种自相关性随着滞后的增大而递减。最大的自相关系数为前一天的自相关系数，为 10.36% ~ 11.93%。这种正自相关性意味着在每一个长达五天的持仓周期内，股票的大单净流入率都具有一定程度的稳定性，这也从根本上保证了分组后股票交易群体结构和他们交易行为的稳定性。

## 4 预测结果

### 4.1 基准结果

表 2 汇报了各股票池神经网络预测收益率的表现。Q1 ~ Q5 为五个大单净流入率依次增大的股票池，在每个股票池中，根据神经网络的输出概率，我们将股票按照模型预测的未来五天中上涨概率从低（L）到高（H）分为十组，并且在表中汇报每组各个股票收益率的市值加权平均相较于整个股票池的周度超额收益率。此外，在 H - L 列汇报了多头 H 组，空头 L 组的投资组合的收益率。Improve 列为 Q1 ~ Q5 股票池的多空组合相较于没有按大单净流入率区分的 Random 组股票池多空组合提升的收益率。Sharpe 列汇报了每个股票池的年化夏普比率。

表中实证结果显示，除了 Q3 和 Random 组外，其他四组均有显著的多空收益率。大单净流入率最高的组 Q5 实现最高的周度收益率 0.59% 和年化夏普比率 2.37。Q1 与 Q2 组有较为接近的周度收

表 2: 资产组合周度收益率 (%)

	L	2	3	4	5	6	7	8	9	H	H - L	Improve	Sharpe
Q1	-0.28 [-2.80]	-0.18 [-2.39]	-0.07 [-1.00]	-0.14 [-1.96]	-0.03 [-0.37]	0.05 [0.86]	0.07 [1.08]	0.07 [0.99]	0.13 [1.67]	0.20 [2.19]	0.48 [3.38]	0.28 [1.65]	1.74
Q2	-0.33 [-2.76]	-0.13 [-1.67]	-0.00 [-0.04]	0.02 [0.34]	0.12 [1.97]	0.07 [1.18]	0.15 [2.64]	0.14 [2.05]	0.17 [2.16]	0.19 [2.13]	0.51 [3.42]	0.32 [1.93]	1.76
Q3	0.07 [0.71]	-0.09 [-1.31]	-0.08 [-1.12]	-0.03 [-0.46]	0.01 [0.13]	0.04 [0.70]	0.06 [0.96]	0.10 [1.82]	0.12 [1.80]	0.15 [1.57]	0.08 [0.51]	-0.12 [-0.75]	0.26
Q4	-0.17 [-1.94]	-0.14 [-1.88]	-0.06 [-0.94]	0.03 [0.50]	0.16 [2.50]	0.05 [1.01]	0.15 [2.33]	0.07 [1.31]	0.16 [2.51]	0.14 [1.48]	0.31 [2.56]	0.11 [0.64]	1.31
Q5	-0.36 [-3.41]	-0.22 [-2.97]	-0.15 [-2.15]	-0.02 [-0.33]	0.01 [0.19]	0.06 [0.88]	0.00 [0.05]	0.14 [1.94]	0.12 [1.68]	0.23 [2.35]	0.59 [4.61]	0.39 [2.29]	2.37
Random	-0.07 [-0.68]	-0.10 [-1.35]	-0.10 [-1.49]	-0.03 [-0.49]	-0.03 [-0.50]	-0.00 [-0.03]	0.08 [1.51]	0.09 [1.45]	0.16 [2.08]	0.13 [1.49]	0.20 [1.45]		0.75
Q1,2,3,4,5	-0.21 [-3.50]	-0.15 [-3.52]	-0.07 [-2.12]	-0.03 [-0.96]	0.05 [1.82]	0.06 [1.98]	0.09 [2.83]	0.10 [3.54]	0.14 [3.83]	0.18 [3.79]	0.39 [4.69]	0.20 [1.55]	2.41
Q1,2,4,5	-0.28 [-4.50]	-0.17 [-3.58]	-0.07 [-1.98]	-0.03 [-0.88]	0.07 [2.01]	0.06 [1.88]	0.09 [2.86]	0.10 [3.26]	0.15 [3.65]	0.19 [3.73]	0.47 [5.59]	0.28 [2.09]	2.88

表中汇报了各股票池基于 K 线图识别的卷积神经网络预测收益率的表现。在每个股票池中，根据神经网络的输出概率，将股票分为十组，从 L 到 H，股票在未来五天中涨的概率逐渐增大。表中每组的值为各个股票收益率的流通市值加权平均相较于整个股票池的周度超额收益率。H - L 列汇报了多头 H 组、空头 L 组的投资组合的收益率。Improve 列为 Q1~Q5 股票池的多空组合相较于 Random 股票池多空组合提升的收益率。Sharpe 列汇报了每个股票池的年化夏普比率。Q1,2,4,5 为等权持仓 Q1、Q2、Q4、Q5 后的结果，排除了 K 线图不具备预测信息的 Q3 组。Q1,2,3,4,5 为等权持仓 Q1、Q2、Q3、Q4、Q5 后的结果。

益率 0.48%，0.51% 和年化夏普比率 1.74，1.76。Q4 组次之，周度收益率为 0.31%，年化夏普比率为 1.31。这四组相较于 Jiang et al. (2023) 中，中国股票市场 0.66 的夏普比率，有非常显著的提升。相较之下，Random 组虽然有 0.2% 的周度收益率，但并不显著，夏普比率为 0.75，与 Jiang et al. (2023) 的结果较为接近但略有提升。Q3 几乎没有任何收益率，基于 K 线图识别的卷积神经网络对其没有预测能力。造成 Random 组与 Q3 组的表现过差的原因是由于在这两组中，投资者结构高度复杂，不仅有大量散户，也有机构的参与，导致 K 线图模式混乱且不提供任何有效的交易信息。考虑到公式 5 中针对神经网络预测误差的分解，本文的分组操作缩小的误差即为泛化误差  $\sum_{k=1}^K \text{Var} [\beta_{i,k}]$ 。在散户与机构有不同交易行为的情况下，其交易行为的差异会导致不同股票的  $\beta$  有更大的方差，从而增加了预测的泛化误差、降低了预测效果。因此，随着大单净流入的增加，投资者在不同组别中的聚集程度从高到低、再由低到高，故尔模型的整体预测效果呈现出 U 型。股票交易者构成居于中位的 Q3 的收益率几乎为 0，比未分组前的训练结果还减小了 0.12%，而两端的散户和机构有很好的预测提升效果。

为了进一步体现减小泛化误差后带来的收益率效果提升，本文计算了等权持仓 Q1、Q2、Q3、Q4、Q5 后的结果，多空投资组合的收益率为 0.39%，相较于未分组训练的结果提升了 0.20%，不同股票池在不同模型下的对冲效果减小了投资组合的方差从而提高了夏普比率，夏普比率有了进一步的提升，达到 2.41。在不考虑 Q3 的情况下，多空投资组合的收益率为 0.47%，提升了 0.28%，夏普比率为 2.88。

我们还在图 4 中汇报六个股票池多空组合的累计收益率曲线。分组来看，Q1 和 Q2 具有很高的相关性，说明当大单比例较低时，神经网络模型对大单净流入率并不敏感，这意味着此时神经网络拟合的因素较为相似。Q3 和 Q4 的模型在 2020 年六月后开始失效，说明机器学习的模型存在过拟合或者该股票池存在模式转换的情况。附录 C 图 5 汇报了六个股票池中，各自的从 L 至 H 共 10 个投资组合的累积收益率曲线。其中，Q1, Q2, Q5 的十个不同组之间具有明显的价差，意味这模型拟合程度非常优秀。相较之下，Q3, Q4, Random 的十个不同组收益率曲线之间存在明显的交叉，说明此时模型的预测结果并不稳健。

为了检验分组学习预测结果的稳健性，本文按照大单净流入率将股票分成十组，为了维持每个资产组合中股票数量与表 2 相同，每个股票池中，按照卷积神经网络将股票分成五组，并构造多空组合，结果如附录 C 表 11 所示。D10 具有最高的周度收益率和年化夏普比率，中间的 D6、D7、D8 的周度收益率和年化夏普比率最低，结果同基准保持一致。值得一提的是，表 1 显示，即使是大单净流入率最高的组，平均大单净流入率依旧不超过 21%，也从侧面说明了中国市场由散户占据主导地位的交易结构。另外，机构往往会拆分自己的订单以隐藏自己的投资行为，这种拆分订单的行为会导致大单净流入率在衡量机构和大户的占比时偏小。然而，只要机构和大户在拆分自己的订单时，拆分方式与每只股票真实的大单净流入率无关，那么每只股票的真实大单净流入率均等地偏小，不会对分组的结果产生显著的影响。同时，为了检验随机组的稳健性，本文进行重复试验，在每个截面上随机选取五分之一的股票并重复五次，得到五个样本集并进行神经网络的训练和检验，结果如附录 C 表 12。对于每一个随机组，周度收益率均在 0.2% 左右，Sharpe 比率位于 0.59-0.82 之间，结果同样稳健，排除了随机性对于训练效果的影响。

本文中使用的卷积神经网络参数总量为 568,822,402 个，作为 32 比特浮点型参数，在计算机中占有 2169.89MB。对于如此复杂的隐式模型，要厘清其究竟学习到什么指标是十分困难的。尽管如此，本

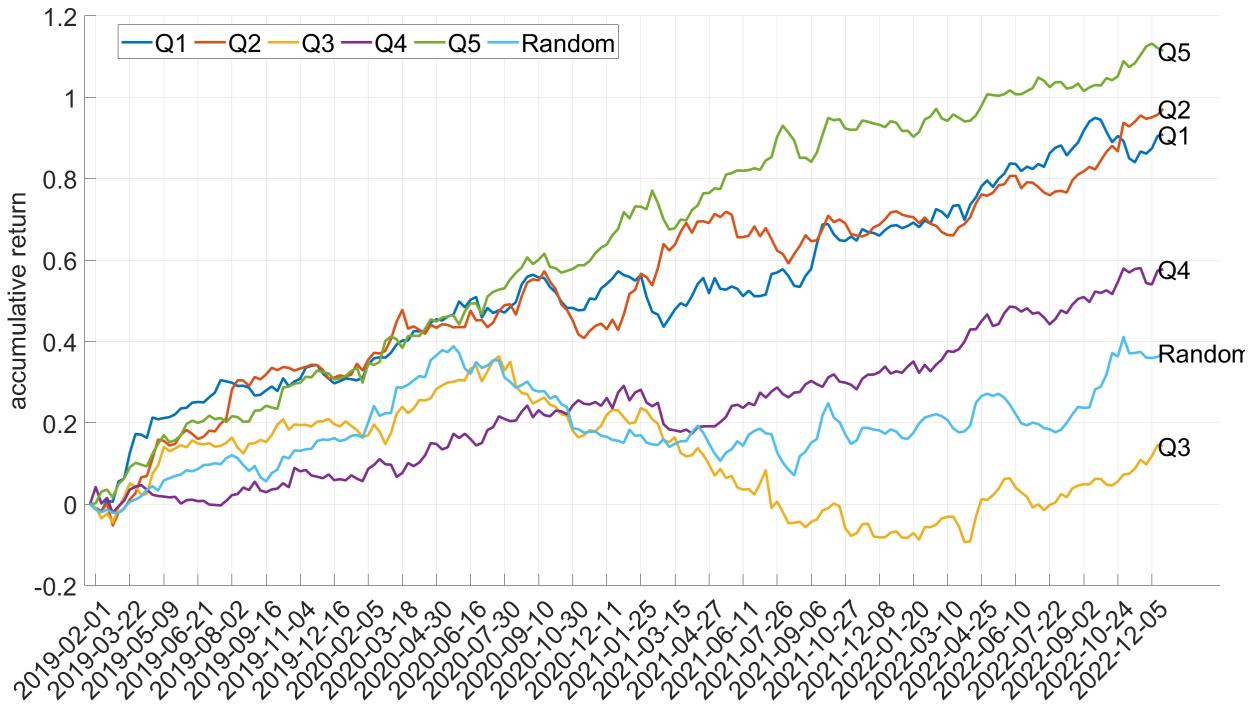


图 4: 每个股票池多空投资组合累积收益率曲线

图中给出了大单净流入率组 Q1~Q5 以及 Random 股票池的多空投资组合累积收益率折线图。每个股票池的多空组合均为多头神经网络预测涨概率最高的 10%、空头最低的 10% 的股票，权重为流通市值加权。

文仍旧在本章节和下一章异质性分析中做出些尝试。章节4.2与4.3分别给出了关于 Fama French 五因子与基础技术指标的研究。这些研究的目的主要有两个：1) 根据资产组合的  $\alpha$  判断 CNN 学习到的资产组合的收益率来源究竟是常用的一些风险因子或技术指标，还是其从 K 线图中挖掘出的新的技术指标。2) 根据风险因子和技术指标的  $\beta$  判断 CNN 的学习内容是否到底包含了哪些风险因子和技术指标。

## 4.2 Fama French 五因子

本文从 CSMAR 上获得了中国市场上 Fama French 五因子的时间序列数据，并将根据卷积神经网络构造的多空资产组合对 Fama French 五因子进行回归：

$$Ret_{i,t} = \alpha_i + \beta_{i,1} (R_t^m - R_t^f) + \beta_{i,2} SMB_t + \beta_{i,3} HML_t + \beta_{i,4} RMW_t + \beta_{i,5} CMA_t + \varepsilon_{i,t} \quad (8)$$

其中  $i$  代表不同的股票池，Q1~Q5 以及 Random 组。 $\alpha$  为因子异象， $R_t^m$  为市场风险溢价因子， $SMB_t$  为规模因子， $HML_t$  为账面市值比因子， $RMW_t$  为盈利能力因子， $CMA_t$  为投资模式因子。五个因子均为流通市值加权平均的投资组合时间序列。表3汇报了 Q1~Q5 以及 Random 组在 Fama French 五因子模型上的回归情况。所有组的因子异象  $\alpha$  相较于其原本的多空组合收益率 H - L 均无显著减小，

表 3: Fama French 五因子模型因子载荷

	H - L(%)	$\alpha$ (%)	$R^m - R^f$	$SMB$	$HML$	$RMW$	$CMA$	$R^2$
Q1	0.48*** [3.38]	0.53*** [3.47]	-0.14** [-2.17]	-0.18* [-1.82]	-0.32*** [-2.68]	0.09 [0.37]	0.46 [1.64]	0.07
Q2	0.51*** [3.42]	0.58*** [3.55]	-0.11 [-1.42]	-0.23** [-2.09]	-0.08 [-0.64]	0.09 [0.40]	0.26 [0.94]	0.06
Q3	0.08 [0.51]	0.10 [0.64]	-0.16** [-1.99]	-0.08 [-0.69]	-0.15 [-1.10]	0.12 [0.53]	0.11 [0.40]	0.05
Q4	0.31** [2.56]	0.31*** [3.05]	0.04 [0.67]	-0.18** [-2.38]	-0.25** [-2.08]	0.01 [0.05]	0.12 [0.63]	0.07
Q5	0.59*** [4.61]	0.56*** [5.20]	0.03 [0.39]	0.01 [0.10]	-0.30* [-1.86]	0.37** [2.15]	0.21 [0.73]	0.10
Random	0.20 [1.45]	0.22 [1.34]	-0.03 [-0.43]	-0.12 [-1.22]	0.02 [0.18]	-0.17 [-0.77]	-0.09 [-0.31]	0.01

表中汇报了大单净流入率 Q1~Q5 组和 Random 组在 Fama French 五因子上回归的因子载荷：

$$Ret_{i,t} = \alpha_i + \beta_{i,1} (R_t^m - R_t^f) + \beta_{i,2} SMB_t + \beta_{i,3} HML_t + \beta_{i,4} RMW_t + \beta_{i,5} CMA_t + \varepsilon_{i,t}.$$

表中，H - L 为神经网络多空组合的收益率， $\alpha$  为因子异象， $R_t^m$  为市场风险溢价因子， $SMB_t$  为规模因子， $HML_t$  为账面市值比因子， $RMW_t$  为盈利能力因子， $CMA_t$  为投资模式因子。五个因子均为流通市值加权平均的时间序列。

例如 Q5 组的超额周度  $\alpha$  为 0.59%，而组合收益率为 0.56%，说明卷积神经网络的收益率来源并非是 Fama French 五因子。此外，尽管输入图像中并不包含基本面信息，但基本面信息依旧会反映在价格数据中，从而被卷积神经网络捕捉到。因此可以看到 Q1、Q3 具有显著的市场因子载荷；Q1、Q2、Q4 具有显著的规模因子载荷；Q1、Q4、Q5 具有显著的账面市值比因子载荷；Q5 具有显著的盈利能力因子载荷。此外，Random 组不在任何一个因子上显著，这主要源自于不同股票在该组之中过于混杂，股票的 K 线图特征在其中表现特殊难以被风险因子捕捉，导致没有任何一个因子能在其中有显著表现。

### 4.3 基本技术指标

本文选择了十二个常用的基本技术指标分别在各个组内构造投资组合。由于输入的图中只包含 60 天量价信息，所以我们只使用窗口期内的数据。除了表征反转 (reversal) 的 5 天累计收益率  $cum5$  和表征动量 (momentum) 的 60 天累计收益率  $cum60$ ，其他的十个技术指标分别为：动量震荡指标  $MomOs$ 、高低价比率指标  $HLRatio$ 、相对强弱指数  $RSI$ 、交易量指标  $VolInd$ 、威廉指数  $Williams$ 、开收价差指标  $OCDiff$ 、历史波动率指标  $HisVol$ 、真实振幅指标  $RealOsMag$ 、最大回撤指标  $MaxBack$ 、以及交易量加权移动均价指标  $MAVol$ ，具体定义见附录D。在计算出每只股票在每个时间截面的技术指标后，仿照 Fama French 五因子的构造方式，本文在每个截面上根据技术指标排序，并用最高的 1/3 与最低的 1/3 的股票构造流通市值加权的多空投资组合，作为该股票池中此技术指标的因子值，这些因子的周度收益率见附录D表14所示。对于单一技术指标，除  $cum5$  (Q2、Q3)， $MomOs$  (Q1)， $VolInd$  (Q1、Q2) 外，均没有显著收益率。

我们将依据卷积神经网络构造的投资组合对技术指标构造的投资组合进行回归：

$$Ret_{i,t} = \alpha_i + \sum_{j=1}^{12} \beta_{i,j} \times Tech_{i,j,t} + \varepsilon_{i,t}, \quad (9)$$

其中， $i$  代表不同的股票池， $j$  代表不同的技术指标， $t$  代表不同的时间截面， $Tech_{i,j,t}$  为不同截面上不同股票池中的 12 个不同的技术指标对应的多空组合。表4给出了 Q1~Q5 以及 Random 组关于技术指标的回归结果。尽管十二个技术指标并没有显著的收益率，但卷积神经网络依旧可以学习到这些技术指标，除了  $HLRatio$ 、 $RSI$  (除 Q5 组) 没有显著的因子载荷外，其他技术指标均在不同股票池中有显著的因子载荷，体现出了卷积神经网络强大的学习能力。以 Q5 组为例，开收价差指标  $OCDiff$  相对强弱指标  $RSI$  均同组合收益显著正相关；类似地，Q1、Q2、Q4、Random 股票池的收益率都得到了一定程度上的解释。由于每个因子在不同股票池中不同均有显著的异质性，所以我们在章节5中对其做出了进一步探究。

### 4.4 其他分组指标

除了大单净流入率外，本文尝试了其他分组指标，包括机构持股比例、基金持股比例、换手率和收盘价。

对于机构持股占比与基金持股占比这两个指标而言，我们依照万得持仓类型分类进行划分。机构持股主要包含一般法人、基金、券商、信托公司、QFII、保险公司、非金融类上市公司、财务公司、银

表 4: 基本技术指标因子载荷

	H - L	$\alpha$	cum	cum	MoM	HL	RSI	Vol	Will-	OC	His	Real	Max	MA	$R^2$	Power
	(%)	(%)	5	60	$O_s$	Ratio	$I_{nd}$	iams	Diff	Vol	$OsMag$	Back	Vol			
Q1	0.48***	0.40***	-0.22***	-0.33*	0.20**	0.39	0.12	-0.07	-0.18	0.06	-0.19	-0.40	-0.29	0.51*	0.20	0.19
	[3.38]	[3.02]	[-2.93]	[-1.88]	[2.37]	[1.38]	[0.67]	[-0.70]	[-1.50]	[0.68]	[-1.48]	[-1.29]	[-1.03]	[1.75]		
Q2	0.51***	0.46***	0.02	0.20	0.09	-0.22	-0.24	-0.10	0.16	0.04	-0.36***	-0.81***	0.63***	0.78***	0.20	0.11
	[3.42]	[3.16]	[0.23]	[1.06]	[1.06]	[-1.25]	[-1.17]	[-0.97]	[1.35]	[0.31]	[-2.87]	[-3.13]	[3.10]	[3.57]		
Q3	0.08	0.10	0.04	0.23	0.27***	0.22	-0.14	-0.24***	0.12	-0.16	-0.26**	0.11	-0.08	0.12	0.20	-0.21
	[0.51]	[0.70]	[0.53]	[1.31]	[2.77]	[0.95]	[-0.72]	[2.87]	[0.76]	[-1.35]	[2.11]	[0.45]	[-0.38]	[0.55]		
Q4	0.31***	0.21***	-0.06	-0.35***	0.22***	0.07	-0.02	-0.21***	-0.21***	0.12	-0.14	0.46***	0.03	-0.33*	0.24	0.45
	[2.56]	[2.28]	[-1.02]	[-2.31]	[3.43]	[0.45]	[-0.14]	[-3.12]	[-2.03]	[1.12]	[-1.14]	[2.31]	[0.16]	[1.69]		
Q5	0.59***	0.59***	-0.16***	-0.51***	-0.00	-0.04	0.30*	-0.06	-0.02	0.21**	-0.01	-0.07	0.01	0.29	0.21	-0.01
	[4.61]	[5.25]	[-2.06]	[-3.23]	[-0.03]	[-0.17]	[1.69]	[-0.87]	[-0.16]	[2.06]	[-0.13]	[-0.34]	[0.02]	[1.47]		
Random	0.20	0.12	-0.09	-0.40**	0.12	0.23	0.37	-0.23***	-0.14	-0.10	-0.03	-0.60	-0.20	0.62*	0.18	0.58
	[1.45]	[0.83]	[-1.11]	[-2.06]	[1.17]	[1.52]	[1.54]	[-3.10]	[-1.12]	[-0.78]	[-0.27]	[-1.66]	[-1.02]	[1.81]		

表中汇报了每个股票池中，累积神经网络的资产组合关于每个技术指标因子的因子载荷，回归方程为：

$$Ret_{i,t} = \alpha_i + \sum_{j=1}^{12} \beta_{i,j} \times Tech_{i,j,t} + \varepsilon_{i,t},$$

其中  $Tech_{i,j,t}$  为技术指标因子时间序列。表中，H - L 为不同股票池中累积神经网络筛选出的多空组合的周度收益率。 $\alpha$  为该多空组合对其他十二个技术指标组合投影后的因子异象。Power 为每个股票池的累积神经网络投资组合被基本技术指标解释的占比。

行、阳光私募、券商集合理财、企业年金、社保基金、基金管理公司、陆股通等。这些持股类型同时包含了流通股和非流通股，因此机构持股占比是相较于总股本的比率。反之，对于基金而言，其持股更偏向于流通股，因此，基金持股占比是相较于流通股的比率。相较于大单净流入率反映的流量信息，持股占比更多的是存量信息。除此之外，[鲁臻 and 邹恒甫 \(2007\)](#) 研究发现成交量大的股票有更强的反转、更弱的动量；[韩豫峰 et al. \(2014\)](#) 根据股票换手率进行分组并研究其各自的基本趋势策略表现，在策略收益率时间序列的一、二、三阶矩和夏普比率上发现了很强的同质性。因此，为了研究其在视觉识别和非线性领域上的同质性和异质性，本文尝试了换手率（交易金额与流通市值的比值）分组指标。考虑到中国 A 股市场的整手买卖规则使得资金量不足的投资者无法购买价格较高的股票，本文也尝试了收盘价作为分组指标。

附录C表13给出了四种不同分类指标下卷积神经网络的学习结果，不同的分类指标对卷积神经网络均有不同程度的提升，证明了异质性样本的区分对于卷积神经网络学习的重要性。

在众多分类指标中，本文选择大单净流入率，更多的是因为其与交易者行为的对应关系，在便于解释 CNN “黑箱子”的同时，还可以进一步探究散户与机构的行为差异，使 CNN 更具有金融学意义。

## 5 异质性研究

本文通过排序大单净流入率并分组得到不同的股票池，大幅提高了卷积神经网络的预测能力。这种方法虽然简单，但是背后的逻辑是通过将有类似交易模式的股票合并在一起进行训练，从而挖掘出投资者隐藏着共同的交易模式。为了进一步探究投资者在 K 线图背后的行为模式，打开机器学习模型内在的黑箱子，理解 CNN 到底通过分组训练学习到了什么，本章节从颜色视觉效应、注意力分布与图形结构技术指标三个角度出发，进一步研究研究异质性与机制。

### 5.1 颜色视觉效应

投资行为容易受到各种情绪的影响，这种情绪多来自于视觉上的敏感性。[姜富伟 et al. \(2021\); 吴武清 et al. \(2020\)](#) 使用机器学习方法，挖掘了文本情绪对股票市场的影响。本文针对 K 线图，认为对于投资者而言，颜色同样也能够显著影响投资情绪。

为了研究这种视觉效应，本文构造了用于描述一张图红绿颜色特征的指标。 $ColorInfoRatio$  为 K 线图区域占全图的比例，主要由每日的高低价差之和与 60 天窗口期的高低价差之比决定；能够反映颜色的  $RedRatio$  和  $GreenRatio$  则代表更为真实的 K 线图，即视觉上红色和绿色区域占整张 K 线图的面积比例，其中， $RedRatio$  为红色区域占全图的比例。 $GreenRatio$  为绿色区域占全图的比例；此外，

我们还定义平均红色深浅  $RedAvg$  以及平均绿色深浅  $GreenAvg$  如下:

$$RedAvg_{i,t} = \frac{\sum_{h=1}^H \sum_{w=1}^W \mathbf{1}_{R_{i,t,h,w}=255} \times \left(1 - \frac{G_{i,t,h,w}}{255}\right)}{\sum_{h=1}^H \sum_{w=1}^W \mathbf{1}_{R_{i,t,h,w}=255}},$$

$$GreenAvg_{i,t} = \frac{\sum_{h=1}^H \sum_{w=1}^W \mathbf{1}_{G_{i,t,h,w}=255} \times \left(1 - \frac{R_{i,t,h,w}}{255}\right)}{\sum_{h=1}^H \sum_{w=1}^W \mathbf{1}_{G_{i,t,h,w}=255}} \quad (10)$$

$RedAvg$  和  $GreenAvg$  可以理解为平均红绿颜色深浅程度, 相较于 60 天累积收益率  $cum60$ , 这种颜色统计方式由于考虑了每天有色区域的面积, 因此更多体现了视觉效应加权的累积收益率而非单纯的窗口累积收益率。五个颜色特征指标的描述性统计如附录E表15所示。对于 Q1~Q5 以及 Random 组, 五个颜色特征指标均具有接近的值, 说明不同的股票池中股票本身体现的颜色信息并没有异质性。换言之, 颜色本身并没有包含特殊的信息, 而更多地体现在了不同交易者在看到颜色后的异质性判断上。根据五个颜色特征指标, 我们多头前  $1/3$  的股票、空头后  $1/3$  的股票以构造颜色因子, 其收益率如附录E表16所示, 所有的颜色因子都没有显著的收益率。我们将卷积神经网络得到的多空组合对该因子进行回归:

$$Ret_{i,t} = \alpha_i + \sum_{j=1}^n \beta_{i,j} \times Color_{i,j,t} + \sum_{j=1}^m \gamma_{i,j,t} \times FF5_{j,t} + \varepsilon_{i,t}, \quad (11)$$

其中  $Color_{i,j,t}$  为第  $i$  个股票池、 $j$  类型、在  $t$  时刻的颜色特征指标因子时间序列。控制变量  $FF5_{j,t}$  为 Fama French 五因子, 回归结果如表5所示。

首先, 在信息比率  $ColorInfoRatio$  中, 随着大单净流入率的增大, 从 Q1 到 Q5 信息比率的因子载荷单调减小,  $\beta$  从 0.20 单调减小到-0.03, 并且显著性逐渐减弱,  $t$  从 3.42 单调减小到-0.37。就其实际意义来看, 信息比率实际上反映了相对高低价差的大小, 即每天高低价差与 60 天观察期内的高低价差的比率, 这一结果意味着散户更加偏好于相对高低价差较大的股票。将信息比率拆分为红色(上涨)信息比率与绿色(下跌)信息比率后, 因子载荷显示出散户对于相对高低价差的偏好主要体现在股票下跌时期, Q1 中  $\beta = 0.19 (t = 3.46)$ , Q2 中  $\beta = 0.26 (t = 3.82)$ , 与散户的抄底行为相吻合。

在第三列中, 我们汇报  $RedAvg$  和  $GreenAvg$  这两个由面积加权的 60 天累计涨收益率和跌收益率(忽略隔夜误差)的因子回归结果, 在  $RedAvg$  上, Q1、Q2、Q3 有更大的因子载荷,  $\beta = -0.22$ 、 $-0.2$  和  $-0.45 (t = -2.81, -2.10 \text{ 和 } -2.92)$ 。这一结果意味着散户在交易时, 更容易受到视觉效应的影响, 对于有更长红柱的 K 线图更敏感, 更容易导致他们的出售股票行为; 而机构在交易时, 更专注于 60 天窗口期的整体收益率表现。

## 5.2 神经网络的注意力分布

章节5.1仅考虑了颜色带来的视觉效应的效果, 这是一种最容易被投资者所理解的注意力效果。然而, 有很多注意力并不仅仅通过颜色体现, 而机器学习模型往往能够通过挖掘非线性的潜在交易模式, 得到信息含量更高的注意力指标。为此, 本章节尝试使用光滑梯度类别激活映射 ++ (*smooth gradient*

表 5: 颜色特征指标因子载荷

	H - L	ColorInfoRatio				RedRatio and GreenRatio				RedAvg and GreenAvg			
		$\alpha$ (%)	Info Ratio	$R^2$	$\alpha$ (%)	Red Ratio	Green Ratio	$R^2$	$\alpha$ (%)	Red Avg	Green Avg	$R^2$	
Q1	0.48*** [3.38]	0.52*** [3.54]	0.20*** [3.42]	0.12	0.50*** [3.50]	0.09	0.19*** [1.52]	0.13	0.47*** [3.46]	-0.22*** [3.35]	0.07 [0.76]	0.11	
Q2	0.51*** [3.42]	0.55*** [3.36]	0.19** [2.25]	0.09	0.53*** [3.23]	0.08	0.26*** [1.01]	0.11	0.51*** [3.82]	-0.25** [3.36]	0.06 [0.51]	0.11	
Q3	0.08 [0.51]	0.10 [0.61]	0.19* [1.75]	0.08	0.08 [0.47]	0.10	0.20	0.10	0.11 [1.54]	-0.45*** [0.74]	0.44*** [2.10]	0.15	
Q4	0.31** [2.56]	0.30*** [3.02]	0.06 [0.76]	0.07	0.30*** [3.05]	0.10*	-0.06	0.10	0.32*** [1.80]	-0.10 [0.96]	-0.10 [3.10]	0.10 [2.92]	
Q5	0.59*** [4.61]	0.56*** [5.21]	-0.03 [-0.37]	0.11	0.54*** [4.93]	-0.04 [-0.54]	0.07 [1.11]	0.12 [5.05]	0.55*** [5.05]	-0.23* [-1.91]	0.18* [1.90]	0.14	
Random	0.20 [1.45]	0.20 [1.23]	0.11 [1.49]	0.02	0.19 [1.20]	0.08 [1.14]	0.10 [0.98]	0.03 [1.40]	0.23 [-1.10]	-0.15 [1.10]	0.08 [0.76]	0.02	

表中汇报了每个股票池中，累积神经网络的资产组合关于颜色特征指标因子的因子载荷，回归方程为：

$$Ret_{i,t} = \alpha_i + \sum_{j=1}^n \beta_{i,j} \times Color_{i,j,t} + \sum_{j=1}^m \gamma_{i,j,t} \times FF5_{i,j,t} + \varepsilon_{i,t},$$

其中  $Color_{i,j,t}$  为第  $i$  个股票池、 $j$  类型、在  $t$  时刻的颜色特征指标因子时间序列。控制变量  $FF5_{i,j,t}$  为 Fama French 五因子。

*class activation mapping++*, 简称 *SmoothGradCAM++*) 来分析神经网络在输入的像素图上的注意力分布, 这种分布通过热力图显示出来 (Omeiza et al., 2019)。

在实际交易中, 机构投资者会依赖成千上万的技术指标做出投资决策, 而我们没有办法一一构造出来逐个检验。而 *SmoothGradCAM++* 的方法却可以通过技术手段在使用量价数据构造的技术指标中, 挖掘出量价数据的各个部分具有的权重。因此, 相较于表象上的颜色, 神经网络的注意力更多的是挖掘了深层次的数据所具有的预测能力, 在无法构造出所有的技术指标的情况下, 其可以代替表征技术指标对量价数据的依赖程度。从这种技术的原理来看, 本节采用的 *SmoothGradCAM++* 是一种检测输出结果针对图上每个位置敏感性的方法, 这种敏感性是通过偏导数度量的, 对于敏感性高的地方, 我们可以说神经网络在这个地方有更高的注意力, 其具体方法原理在附录F介绍, 具体热力图的样例如附录F图6所示。从图 (a) 可以看出, 有些地方没有 K 线数据, 但仍旧吸引了神经网络的部分注意力, 并且这些注意力呈现出噪音的分布。故尔, 就图像整体而言, K 线数据具有更高的亮度, 即更高的注意力, 说明神经网络在学习 K 线图时, 确实学习到了 K 线数据。从图 (b) 可以看出, 即使是在同一条 K 线数据上, 神经网络的注意力也有不同分布。

在实际使用光滑梯度类别激活映射 ++ 方法的过程中, 首先我们在计算注意力时添加的随机扰动引入了部分噪音, 这些噪音在无 K 线图区域可以统计出来, 在 K 线图区域减去这些噪音的均值, 从而得到去除噪音后的 K 线图注意力。然后, 我们汇报了 K 线图上各部分的注意力均值, 以体现其在神经网络上所占权重大小, 如表6所示。需要注意的是, 神经网络在学习 K 线图时, 除了学习 K 线图本身, 还会学习 K 线图的边界, 因此在统计信息分布时, 可能还需要 K 线图的信息区域, 可能还需要考虑 K 线图的信息区域边界。在汇报神经网络注意力分布的描述性统计表表6时, 我们选取了不同 K 线图的不同区域作为分析对象。其中, Panel A 中只有 K 线图区域被定义为信息区域, Panel B 中除了 K 线图区域, K 线图的边界也被定义为信息区域并参与计算注意力值。

结果显示, 在表6的 Panel A 中 =, Q5 作为具有最高多空组合收益率的组, 神经网络在价量信息上有最高的注意力, 为 6.37%, 其次为 Q1 组, 为 5.52%, 这两个组别作为交易者成分更纯粹的组, 神经网络在 K 线图上的注意力最高, 也意味着神经网络在其中学习到了更多信息。特别地, Q1、Q3、Q5 中, 开盘价比收盘价有更高的注意力; 除 Q5 外, 开收价的注意力都要高于高低价。说明在神经网络学习时, Q5 组对于日内高低价差比其他四组有更高的敏感性, 而其他组有更多的注意力在日内收益上。Panel B 中的描述性统计结果显示, 当 K 线图的边界被包含进来后, 信息区域的注意力得到了显著的提升, 约 0.5~0.8%。由此我们可以判断, 神经网络在读图时, 除了读 K 线本身, 也会读取 K 线数据的边界来构造特征。考虑边界后, 价格信息的注意力和交易量信息的注意力变得平衡, 说明神经网络在学习交易量信息时比价格信息采用了更多的边缘检测的方式。考虑边界后, 结果是稳健的。Q5 具有最高的注意力, 为 7.13%, 其次为 Q1 组, 为 6.17%。值得注意的是, 当前的 K 线图边缘是 K 线信息外的一个像素宽度, 当我们将这个定义拓宽到两个像素宽度时, 注意力大小减小了, 说明噪音被包含了进来。由此可见, 神经网络在学习 K 线图时, 定义的边缘检测只有一个像素的宽度。

基于这一方法, 我们进一步研究神经网络的预测表现与神经网络在图上注意力的关系, 探索神经网络关注到了 K 线图中的哪些因素。据此, 我们采用与章节5.1类似的方式, 依据注意力高低将股票在截面上排序, 并多头前 1/3 的股票和空头后 1/3 的股票构造资产组合作为注意力因子值, 其收益率如附

表 6: 注意力分布描述性统计 (%)

	Price+ Volume	Price	Volume	Open	Close	Open + Close	High + Low
Panel A: Information Region							
Q1	5.52	5.71	5.20	5.92	5.80	5.86	5.66
Q2	5.38	5.58	5.05	5.73	5.76	5.75	5.52
Q3	5.40	5.56	5.13	5.73	5.59	5.66	5.53
Q4	5.25	5.45	4.91	5.56	5.73	5.65	5.38
Q5	6.37	6.49	6.18	6.93	6.07	6.50	6.50
Random	5.40	5.58	5.10	5.72	5.73	5.73	5.53
Panel B: Information Region including Edges							
Q1	6.17	6.14	6.24	5.84	6.75	6.29	5.84
Q2	6.04	6.02	6.10	5.59	6.76	6.18	5.70
Q3	6.02	5.99	6.07	5.66	6.57	6.12	5.73
Q4	5.88	5.86	5.92	5.37	6.67	6.02	5.53
Q5	7.13	7.07	7.25	7.19	7.20	7.19	6.82
Random	6.02	5.99	6.08	5.58	6.71	6.14	5.68

表中汇报了注意力均值在像素图上的分布情况。Price 为价信息，Volume 为量信息，Open 为开盘价信息，Close 为收盘价信息，High 为最高价信息，Low 为最低价信息。Panel A 中，只有 K 线图区域被定义为信息区域。Panel B 中，除了 K 线图区域，K 线图的边界区域也被定义为信息区域。

录F表17所示, Q5 的量价、价格、交易量、开盘价、开收价、高低价注意力因子均有显著收益率。由此可见, 卷积神经网络确实能够挖掘到机构背后根据量价信息的决策逻辑。我们进一步将卷积神经网络的资产组合对这些因子进行回归:

$$Ret_{i,t} = \alpha_i + \sum_{j=1}^n \beta_{i,j} \times Attention_{i,j,t} + \sum_{j=1}^m \gamma_{i,j,t} \times FF5_{i,j,t} + \varepsilon_{i,t}, \quad (12)$$

其中  $Attention_{i,j,t}$  为第  $i$  个股票池、 $j$  类型、在  $t$  时刻的注意力因子时间序列。控制变量  $FF5_{i,j,t}$  为 Fama French 五因子。回归结果如表7所示。作为资产组合表现最好的组, Q5 在各个因子指标中整体上具有最高和最显著的因子载荷,  $\beta = 0.37 (t = 4.07)$ , 显著高于其他四组。其收益率的主要来源为价格信息、开盘价以及高低价上的注意力, 并且在价格信息和开盘价上均有最高和最显著的因子暴露,  $\beta = 0.45 (t = 2.46)$  以及  $\beta = 0.39 (t = 3.61)$ 。收益率表现次之的 Q2, 收益率来源主要为量价信息和开收价信息上的注意力, 其中在交易量信息和收盘价信息上有最高和最显著的因子暴露,  $\beta = 0.34 (t = 2.46)$  以及  $\beta = -0.24 (t = -2.59)$ 。收益率表现再次的 Q1, 其收益率来源主要为量价信息、开盘价和高低价上的注意力, 其在高低价上有最高和最显著的因子载荷  $\beta = 0.31 (t = 2.57)$ 。收益率表现较差的 Q3 和 Q4, 在各个部分的注意力上因子载荷表现较差。由此可以判断, 卷积神经网络在不同的股票池中, 会把注意力放在量价信息的不同部分, 同时这些注意力带来的收益率也是不同的。为了检验结果的鲁棒性, 本文对考虑边界的注意力因子进行了同样的分析, 结果如附录F表18所示。考虑边界后, Q5 在量价信息和开收价上依旧有最高和最显著的因子载荷,  $\beta = 0.45 (t = 4.38)$  以及  $\beta = 0.33 (t = 3.15)$ 。

### 5.3 图型技术指标

传统方法往往仅使用线性回归和构建投资组合的方式研究技术分析, 而其它更为常见的机器学习模型如 SVM、LSTM、XGBoost 往往只能依赖于时间序列数据来构建技术分析指标。对于技术分析而言, 将图形的“走势”转化为量化指标往往过于模糊和主观, 在提取信息特征的时候也会丢失大多信息。而基于 CNN 在处理图像方面的天然优势, 我们得以进一步研究图形技术分析对于股价的预测能力。

据此, 本文采用了Lo et al. (2000) 的方式根据收盘价定义了 8 种 K 线图型结构: 头肩 (*header and shoulder, HS*)、反头肩 (*inverse header and shoulder, IHS*)、扩展顶 (*broadening top, BTOP*)、扩展底 (*broadening bottom, BBOT*)、三角顶 (*triangle top, TTOP*)、三角底 (*triangle bottom, TBOT*)、矩形顶 (*rectangle top, RTOP*)、矩形底 (*rectangle bottom, RBOT*)。附录G给出了 8 种结构的定义方式。然而, 我们并不能单纯使用收盘价进行定义极值点, 因为许多极值点过于局部, 实际上是噪音的体现。因此需要去掉这些噪音得到真正的价格趋势变动, 从而得到极值点。

Lo et al. (2000) 使用高斯核估计 (*Gaussian kernel estimation*) 方法进行降噪, 也就是高斯平滑 (*Gaussian smoothing*) 法。高斯平滑方法的基本原理也在附录G中介绍。高斯平滑方法是一种局部方法, 即每一个点的高斯核估计结果由其前后  $[-r, r]$  个点决定, 以此来剔除噪音。为了检验其稳健性, 本文采用了另一种方法去除噪音, 即经验模态分解 (*empirical mode decomposition, EMD*)。经验模态分解是一种提取整体模态的方法, 从高频开始提取成分, 逐步获取到低频成分, 剔除前面的高频成分后, 剩下的部分即为平滑后的结果。王文波 et al. (2010) 最早使用 EMD 方法进行模态分解, 以提高 RNN 的

表 7: 注意力分布因子载荷

	H - L	Price				Price				Open				Open + Close	
		+ Volume		and Volume		+ Volume		and Volume		Open		Close		and High + Low	
		(%)	α (%)	Price (%)	R <sup>2</sup> (%)	α (%)	Price (%)	Volume	R <sup>2</sup> (%)	α (%)	Open	Close	R <sup>2</sup>	α (%)	Open
Q1	0.48*** 0.52***	0.23*	0.11 0.53***	0.20	0.08	0.11 0.54***	0.30***	0.00	0.13 0.54***	0.00	0.13	0.54***	0.00	0.01	0.31** 0.15
	[3.38] [3.61]	[1.68]	[3.64]	[1.56]	[0.51]	[3.68]	[2.77]	[0.01]	[4.04]	[0.15]	[2.57]				
Q2	0.51*** 0.57***	0.06	0.06 0.62***	-0.21*	0.34**	0.09 0.50***	0.24**	-0.24**	0.10 0.55***	-0.15	0.15	0.15	-0.15	0.15	0.07
	[3.42] [3.56]	[0.53]	[3.80]	[-1.70]	[2.46]	[3.20]	[2.52]	[-2.59]	[3.35]	[-0.93]	[0.85]				
Q3	0.08	0.10	0.13	0.06	0.12	0.01	0.17	0.07	0.09	-0.01	0.06	0.06	0.11	-0.11	0.24** 0.08
	[0.51] [0.65]	[0.98]	[0.76]	[0.08]	[1.16]	[0.62]	[0.62]	[-0.09]	[0.57]	[0.73]	[-0.95]	[0.73]	[-0.95]	[0.73]	[2.37]
Q4	0.31** 0.31***	0.15	0.08 0.31***	0.13	0.03	0.08 0.32***	0.00	0.00	0.14	0.08 0.31***	0.05	0.05	0.18	0.09	
	[2.56] [3.05]	[1.37]	[3.04]	[1.04]	[0.38]	[3.07]	[0.03]	[0.03]	[1.40]	[3.05]	[0.48]	[3.05]	[0.48]	[3.05]	[1.65]
Q5	0.59*** 0.47***	0.37***	0.18 0.44***	0.45**	0.05	0.23 0.42***	0.39***	0.13	0.24 0.44***	0.34***	0.34***	0.34***	0.23** 0.24		
	[4.61] [4.09]	[4.07]	[3.71]	[2.46]	[0.30]	[3.54]	[3.61]	[1.13]	[3.71]	[2.42]	[2.05]				
Random	0.20	0.22	0.11	0.02	0.24	-0.18	0.31***	0.05	0.24	0.16	-0.15	0.03	0.21	0.02	0.03 0.01
	[1.45]	[1.35]	[0.99]	[1.55]	[-1.42]	[2.69]	[1.53]	[1.21]	[-1.12]	[1.32]	[0.18]	[0.18]	[0.18]	[0.18]	[0.20]

表中汇报了每个股票池中，卷积神经网络的资产组合在注意力因子上的因子载荷，回归方程为：

$$Ret_{i,t} = \alpha_i + \sum_{j=1}^n \beta_{i,j} \times Attention_{i,j,t} + \sum_{j=1}^m \gamma_{i,j,t} \times FF5_{i,j,t} + \varepsilon_{i,t},$$

其中  $Attention_{i,j,t}$  为第  $i$  个股票池、 $j$  类型、在  $t$  时刻的注意力因子时间序列。控制变量  $FF5_{i,j,t}$  为 Fama French 五因子。

预测精度。附录G图7给出了判断的结果示例，可以看出，两种方法会得到不同的曲线导致了不同的极值点选取，从而得到不同的技术指标。整体而言，由于在使用经验模态分解时，仅剔除了第一阶高频模态，因此经验模态分解平滑曲线比高斯平滑曲线更敏感。附录G中表19给出了两种平滑方法在各个股票池中的技术指标分布情况，方括号中为所占股票池比例。两种方法中，具有特定技术指标的股票数量几乎一致。HS、IHS、BTOP、BBOT、TTOP、TBOT 所占比例在两种平滑方法下几乎相同。但 RTOP 中，高斯平滑方法的数量要显著高于经验模态分解方法，RBOT 则相反。技术指标的分布与大单净流入率有明显的关系，Q5 有更多的 HS、TTOP、RTOP 和更少的 IHS、BBOT、TBOT、RBOT。

在实证检验中，我们将具有每个技术指标和无技术指标的股票按流通市值加权平均得到资产组合，并多头每个技术指标的资产组合，空头无技术指标的资产组合，得到该技术指标的因子。附录G表20给出了因子的收益率，大部分图型技术指标的投资组合并没有显著的收益率。附录G表21给出了两种方法得到的因子的相关性，方括号中为相关性的 p 值。对于大部分图型技术指标，两种平滑方法没有高度的重合度，可以看出，这种图型技术指标的判断较为依赖平滑方法，判断结果并不稳健，该结论在 Murray et al. (2024) 中也有提及。为了研究神经网络能否学习到各种图型技术指标，本文将卷积神经网络的资产组合对图型技术指标因子进行回归：

$$Ret_{i,t} = \alpha_i + \sum_{j=1}^n \beta_{i,j} \times PatternTech_{i,j,t} + \varepsilon_{i,t}, \quad (13)$$

表8给出了各个股票池神经网络资产组合在高斯平滑方法的图型技术指标因子上因子载荷，经验模态分解方法的结果作为稳健性检验在附录G表22中。Q5 学习到了最多的图型技术指标，包括 HS、BTOP、BBOT、TTOP 以及 RTOP。图型技术指标对 Q5 有较高的解释力度， $R^2$  达到了 18%，表明了机构投资者对于技术分析的决策过程更加依赖。以 HS (头肩顶形状) 为例，回归因子载荷为 0.17 ( $t = 2.38$ )，意味着 1% 的头肩顶形状能够提升 0.17% 的 CNN 策略收益率。Q1 (TTOP)，Q2 (BTOP、RTOP)，Q3 (HS) 仅学习到 1~2 个图型技术指标。Q4 与 Random 几乎没有学习到任何技术指标。

## 6 结论

Jiang et al. (2023) 使用卷积神经网络对美国股市进行预测，得到了较好的预测效果，然而对中国股市的预测仅获得了 0.66 的夏普比率。本文从减小模型泛化误差的角度出发，根据大单净流入率在截面上将股票池分为五组，对 CNN 模型进行分组训练。实证结果显示分组训练后能够大幅提升预测效果，最多能有效提升模型的夏普比率至 2.37。同时，随着构投资者在交易过程中占比的逐步增多，投资组合的未来收益率和大单净流入率之间呈现出 U 型关系。对于投资者集中程度更高的投资组合 Q1 和 Q5 而言，卷积神经网络拥有非常强大的预测能力，而对于充斥着不同机构和个人投资者、市场交易主体成分复杂的 Q3 和 Random 组而言，卷积神经网络的预测效果较显著更差。

此外，本文还检验了卷积神经网络投资组合在 Fama French 五因子和十二个基本技术指标上的投影情况，得到了显著的因子异象，证明了其非线性的数据挖掘带来的显著收益率。同时，不同组别的投资组合在相同因子和技术指标上不同的因子暴露体现出不同组别的异质性。

表 8: 图型技术指标因子投影

	H - L	$\alpha(\%)$	HS	IHS	BTOP	BBOT	TTOP	TBOT	RTOP	RBOT	$R^2$
Q1	0.48*** [3.38]	0.47*** [2.94]	-0.04 [-0.84]	0.04 [0.78]	-0.05 [-1.24]	0.00 [0.09]	0.16*** [3.25]	-0.00 [-0.00]	0.02 [0.39]	-0.08 [-1.48]	0.07
Q2	0.51*** [3.42]	0.48*** [2.97]	0.01 [0.17]	-0.03 [-0.50]	-0.12* [-1.87]	0.05 [1.13]	-0.02 [-0.40]	-0.03 [-0.75]	0.17*** [2.64]	-0.05 [-0.94]	0.07
Q3	0.08 [0.51]	0.05 [0.35]	0.10* [1.74]	-0.09 [-1.18]	-0.03 [-0.67]	-0.06 [-1.60]	0.04 [0.81]	-0.02 [-0.62]	-0.05 [-1.06]	-0.07 [-0.93]	0.06
Q4	0.31** [2.56]	0.25** [2.55]	-0.02 [-0.45]	-0.04 [-1.32]	-0.02 [-0.56]	-0.02 [-1.10]	-0.01 [-0.18]	-0.03 [-1.20]	0.02 [0.62]	-0.03 [-0.68]	0.03
Q5	0.59*** [4.61]	0.61*** [6.05]	0.17** [2.38]	-0.08 [-1.35]	-0.10*** [-2.66]	-0.06** [-2.02]	0.06* [1.76]	-0.03 [-0.93]	0.14*** [2.98]	-0.06 [-1.61]	0.18
Random	0.20 [1.45]	0.18 [1.11]	0.14 [1.64]	-0.05 [-0.88]	-0.05 [-1.09]	-0.00 [-0.14]	0.00 [0.04]	-0.01 [-0.25]	-0.00 [-0.10]	0.02 [0.65]	0.04

表中汇报了每个股票池中，卷积神经网络的资产组合在高斯平滑方法构造的因子时间序列上的因子载荷，回归方程为：

$$Ret_{i,t} = \alpha_i + \sum_{j=1}^n \beta_{i,j} \times PatternTech_{i,j,t} + \varepsilon_{i,t}, \quad (14)$$

其中  $PatternTech_{i,j,t}$  为图型技术指标因子时间序列。

为了进一步探究不同投资者构成带来的 K 线图交易异质性，本文还进一步研究了 K 线图的颜色特征指标、神经网络的注意力指标和图型技术指标。实证结果显示，虽然 K 线图的颜色特征指标本身并没有包含特殊信息，然而不同类型的投资者却会对颜色做出不同反映，例如散户组会对绿色的 K 线图更为敏感，而机构组的反映则并不显著。同时，本文也探索了卷积神经网络在每张 K 线图上的注意力分布，并对开盘价、收盘价、高低价、量信息、和价信息分别进行了统计，发现不同股票池中神经网络具有不同的注意力分布。卷积神经网络对注意力因子的投影结果说明机构对于注意力的敏感性更强。最后，本文采用了 Lo et al. (2000) 使用的八个图型技术指标构造了图型技术指标因子，发现仅有大单净流入率最高的机构组学习到了这些图型技术指标。异质性研究的整体结果显示，散户将更多注意力放在 K 线图的颜色信息上，而机构能够挖掘 K 线图中的技术形态等更深层次的信息。

本文提出的方法虽然简单，但是却非常有效，并且能够运用到所有基于机器学习挖掘投资者行为模式的投资策略中。除了分析了卷积神经网络的学习内容外，本文还使用颜色、注意力、和图型技术指标分析了散户和机构在根据 K 线图做决策时的行为异质性，为未来卷积神经网络模型在金融中的应用提供借鉴意义，也为之后研究者们将图像、文本、视频、音频、数字结合在一起的多模态深度学习模型预测未来股价收益率做出铺垫。

## 参考文献

- Klakow Akepanidtaworn, Rick Di Mascio, Alex Imas, and LAWRENCE DW SCHMIDT. Selling fast and buying slow: Heuristics and trading performance of institutional investors. *The Journal of Finance*, 78(6):3055–3098, 2023.
- Warren Bailey, Alok Kumar, and David Ng. Behavioral biases of mutual fund investors. *Journal of financial economics*, 102(1):1–27, 2011.
- Pierre Bajgrowicz and Olivier Scaillet. Technical trading revisited: False discoveries, persistence tests, and transaction costs. *Journal of Financial Economics*, 106(3):473–491, 2012.
- Brad M Barber and Terrance Odean. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The review of financial studies*, 21(2):785–818, 2008.
- Brad M Barber, Yi-Tsung Lee, Yu-Jane Liu, and Terrance Odean. Just how much do individual investors lose by trading? *The Review of Financial Studies*, 22(2):609–632, 2009.
- Brad M Barber, Xing Huang, Terrance Odean, and Christopher Schwarz. Attention-induced trading and returns: Evidence from robinhood users. *The Journal of Finance*, 77(6):3141–3190, 2022.
- Azi Ben-Rephael, Zhi Da, and Ryan D Israelsen. It depends on where you search: Institutional investor attention and underreaction to news. *The Review of financial studies*, 30(9):3009–3047, 2017.
- Gunduz Caginalp and Henry Laurent. The predictive power of price patterns. *Applied Mathematical Finance*, 5(3-4):181–205, 1998.
- Jiasheng Cao and Jinghan Wang. Stock price forecasting model based on modified convolution neural network and financial time series analysis. *International Journal of Communication Systems*, 32(12):e3987, 2019. ISSN 1074-5351. doi: <https://doi.org/10.1002/dac.3987>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/dac.3987>.
- J. F. Chen, W. L. Chen, C. P. Huang, S. H. Huang, and A. P. Chen. Financial time-series data analysis using deep convolutional neural networks. In *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, pages 87–92, 2016. doi: 10.1109/CCBD.2016.027.
- Luyang Chen, Markus Pelger, and Jason Zhu. Deep learning in asset pricing. *Management Science*, 70(2):714–750, 2024.
- Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

Jingwen Jiang, Bryan T. Kelly, and Dacheng Xiu. (re-)imag(in)ing price trends. *The Journal of Finance*, n/a(n/a), 2023. ISSN 0022-1082. doi: <https://doi.org/10.1111/jofi.13268>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.13268>.

Andrei Kirilenko, Albert S Kyle, Mehrdad Samadi, and Tugkan Tuzun. The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998, 2017. ISSN 0022-1082. doi: <https://doi.org/10.1111/jofi.12498>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12498>.

Markus Leippold, Qian Wang, and Wenyu Zhou. Machine learning in the chinese stock market. *Journal of Financial Economics*, 145(2):64–82, 2022.

Shuanglong Liu, Chao Zhang, and Jinwen Ma. Cnn-lstm neural network model for quantitative strategy analysis in stock markets. In *International Conference on Neural Information Processing*, 2017.

Andrew W. Lo, Harry Mamaysky, and Jiang Wang. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *The Journal of Finance*, 55(4):1705–1765, 2000. ISSN 0022-1082. doi: <https://doi.org/10.1111/0022-1082.00265>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/0022-1082.00265>.

Huahao Lu and huihang wu. Can stock price image predict future return? evidence from chinese stock market. *SSRN Electronic Journal*, 2022. doi: 10.2139/ssrn.4171663.

Lu Lu, Xuhui Meng, Zhiping Mao, and George Em Karniadakis. Deepxde: A deep learning library for solving differential equations. *SIAM Review*, 63(1):208–228, 2021. doi: 10.1137/19m1274067. URL <https://pubs.siam.org/doi/abs/10.1137/19M1274067>.

Wenjie Lu, Jiazheng Li, Yifan Li, Aijun Sun, and Jingyang Wang. A cnn-lstm-based model to forecast stock prices. *Complexity*, 2020:6622927, 2020. ISSN 1076-2787. doi: 10.1155/2020/6622927. URL <https://doi.org/10.1155/2020/6622927>.

Ben R Marshall, Martin R Young, and Lawrence C Rose. Candlestick technical trading strategies: Can they create value for investors? *Journal of Banking & Finance*, 30(8):2303–2323, 2006.

Sarah McBride. Technical analysts, seen as quacks, fight the image. *Wall Street Journal*, 2001. URL <https://www.wsj.com/articles/SB99842746074105545>.

Scott Murray, Yusen Xia, and Houping Xiao. Charting by machines. *Journal of Financial Economics*, 153:103791, 2024.

Terrance Odean. Are investors reluctant to realize their losses? *The Journal of finance*, 53(5):1775–1798, 1998.

Terrance Odean. Do investors trade too much? *American economic review*, 89(5):1279–1298, 1999.

Daniel Omeiza, Skyler Speakman, Celia Cintas, and Komminist Weldermariam. Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models. *arXiv preprint arXiv:1908.01224*, 2019.

Keiron O’Shea and Ryan Nash. An introduction to convolutional neural networks. *ArXiv e-prints*, 2015.

Carol L Osler. Currency orders and exchange rate dynamics: An explanation for the predictive success of technical analysis. *The Journal of Finance*, 58(5):1791–1819, 2003.

Cheol-Ho Park and Scott H Irwin. What do we know about the profitability of technical analysis? *Journal of Economic surveys*, 21(4):786–826, 2007.

Alberto G Rossi. Predicting stock market returns with machine learning. *Georgetown University*, 2018.

Jon Sindreu. Today’s irrational stock market: A moment in the sun for technical analysts? *Wall Street Journal*, 2020. URL <https://www.wsj.com/articles/todays-irrational-stock-market-a-moment-in-the-sun-for-technical-analysts-11592571858>.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

R. Zhang, Z. Yuan, and X. Shao. A new combined cnn-rnn model for sector stock price analysis. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, volume 02, pages 546–551, 2018. ISBN 0730-3157. doi: 10.1109/COMPSAC.2018.10292.

万谍 and 杨晓光. 价格跳跃前大中小单的行为特征和信息含量. *管理科学学报*, 22(10):37–54, 2019. ISSN 1007-9807.

吴武清, 赵越, 闫嘉文, and 汪寿阳. 分析师文本语调会影响股价同步性吗?——基于利益相关者行为的中介效应检验. *管理科学学报*, 23(09):108–126, 2020. ISSN 1007-9807.

姜富伟, 孟令超, and 唐国豪. 媒体文本情绪与股票回报预测. *经济学 (季刊)*, OL(04):1323–1344, 2021.

姜富伟, 薛浩, and 周明. 大数据提升了多因子模型定价能力吗?——基于机器学习方法对我国 a 股市场的探究. *系统工程理论与实践*, pages 2037–2048, 2022.

尹海员 and 朱旭. 机构投资者信息挖掘、羊群行为与股价崩盘风险. *管理科学学报*, 25(02):69–88, 2022. ISSN 1007-9807. doi: 10.19920/j.cnki.jmsc.2022.02.004.

张乐 and 李好好. 我国证券市场中的噪声交易研究——基于一个“机构噪声交易者—散户噪声交易者模型”的分析. 中国管理科学, 16(S1):340–345, 2008. ISSN 1003-207X. doi: 10.16381/j.cnki.issn1003-207x.2008.s1.123.

张兵. 中国资本市场的 t+1 交易制度研究: 隔夜收益率视角. 管理世界, 36(12):26–35+51+36, 2020. ISSN 1002-5502. doi: 10.19744/j.cnki.11-1235/f.2020.0181.

景楠, 史紫荆, and 舒毓民. 基于注意力机制和 cnn-lstm 模型的沪铜期货高频价格预测. 中国管理科学, OL:1–13, 2020. ISSN 1003-207X. doi: 10.16381/j.cnki.issn1003-207x.2020.0342. URL <https://kns.cnki.net/kcms/detail/11.2835.G3.20200813.1428.005.html>.

李斌, 林彦, and 唐闻轩. Ml-tea: 一套基于机器学习和技术分析的量化投资算法. 系统工程理论与实践, 37(5):1089–1100, 2017. ISSN 1000-6788. doi: 10.12011/1000-6788(2017)05-1089-12. URL [https://sysengi.cjoe.ac.cn/CN/10.12011/1000-6788\(2017\)05-1089-12](https://sysengi.cjoe.ac.cn/CN/10.12011/1000-6788(2017)05-1089-12).

李斌, 邵新月, and 李玥阳. 机器学习驱动的基本面量化投资研究. 中国工业经济, 8(08):61–79, 2019. ISSN 1006-480X. doi: 10.19581/j.cnki.ciejournal.2019.08.004. URL <https://link.cnki.net/doi/10.19581/j.cnki.ciejournal.2019.08.004>.

林耀虎, 刘善存, and 杨海军. 一种基于机器学习和蜡烛图的股市投资策略研究. 计量经济学报, 2: 126–140, 2022. ISSN 2096-9732.

汪刘凯, 张小波, 闫相斌, and 王未卿. 基于混频数据驱动神经网络模型的波动率预测研究. 系统工程理论与实践, 43(12):3488–3507, 2023. ISSN 1000-6788. URL <https://kns.cnki.net/kcms/detail/11.2267.N.20230907.1847.html>.

汪天都 and 孙谦. 传统监管措施能够限制金融市场的波动吗?. 金融研究, 459(9):177, 2018. URL [http://www.jryj.org.cn/CN/abstract/article\\_460.shtml](http://www.jryj.org.cn/CN/abstract/article_460.shtml).

王刚, 陈红, 马敬玲, and 王珏. 基于多尺度 1d-cnn 和注意力机制的汇率多步预测研究. 系统工程理论与实践, pages 1–17, 2023. ISSN 1000-6788. URL <https://kns.cnki.net/kcms/detail/11.2267.N.20231211.1149.012.html>.

王文波, 费浦生, and 翟旭明. 基于 emd 与神经网络的中国股票市场预测. 系统工程理论与实践, 30(06): 1027–1033, 2010. ISSN 1000-6788.

苏冬蔚 and 倪博. 转融券制度、卖空约束与股价变动. 经济研究, 53(03):110–125, 2018. ISSN 0577-9154.

许泳昊, 徐鑫, and 朱菲菲. 中国 a 股市场的“大单异象”研究. 管理世界, 38(07):120–136, 2022. ISSN 1002-5502. doi: 10.19744/j.cnki.11-1235/f.2022.0102.

赵涛 and 郑祖玄. 信息不对称与机构操纵——中国股市机构与散户的博弈分析. 经济研究, 07:41–48+91, 2002. ISSN 0577-9154.

赵鹏. 基金经理教你如何选择基金经理. 清华金融评论, 2015. URL <http://www.thfr.com.cn/wap/index-wap2.php?p=15253>.

郑振龙 and 孙清泉. 彩票类股票交易行为分析: 来自中国 a 股市场的证据. 经济研究, 48(05):128–140, 2013. ISSN 0577-9154.

陈凯杰, 唐振鹏, 吴俊传, 杜晓旭, and 蔡毅. 基于分解-集成和混频数据采样的中国股票市场预测研究. 系统工程理论与实践, 42(11):3105–3120, 2022. ISSN 1000-6788. URL <https://kns.cnki.net/kcms/detail/11.2267.n.20220915.1547.009.html>.

韩豫峰, 汪雄剑, 周国富, and 邹恒甫. 中国股票市场是否存在趋势?. 金融研究, 3, 2014. ISSN 1002-7246. URL <https://cstj.cqvip.com/Qikan/Article/Detail?id=49172126>.

鲁臻 and 邹恒甫. 中国股市的惯性与反转效应研究. 经济研究, 42(9):11, 2007.

## 附录 A 大单净流入率截面相关性及自相关性

正文章节3.2中，计算的大单净流入率与投资率  $INV$ 、账面市值比  $B/M$ 、净资产收益率  $ROE$ 、总资产收益率  $ROA$ 、杠杆率  $Leverage$ 、以及公司规模  $Size$  的截面相关性如表9所示；大单净流入率自

表 9: 大单净流入率指标与主要基本面指标相关性 (%)

	<i>LargeOrder</i>	<i>INV</i>	<i>B/M</i>	<i>ROE</i>	<i>ROA</i>	<i>Leverage</i>	<i>Size</i>
<i>LargeOrder</i>	100.00	-0.16	-0.23	0.10	-0.66	0.96	1.33
<i>INV</i>		100.00	5.80	1.74	0.37	8.83	4.08
<i>B/M</i>			100.00	0.62	-6.13	14.57	8.26
<i>ROE</i>				100.00	43.71	-7.94	6.76
<i>ROA</i>					100.00	-37.71	4.90
<i>Leverage</i>						100.00	14.53
<i>Size</i>							100.00

表中汇报了大单净流入率  $LargeOrder$  与投资率  $INV$ 、账面市值比  $B/M$ 、净资产收益率  $ROE$ 、总资产收益率  $ROA$ 、杠杆率  $Leverage$ 、公司规模  $Size$  的相关关系。该相关系数为每个时间截面上相关系数的均值。

相关性如表10所示。

## 附录 B 像素图构造

将量价数据转化为像素图时，使用收盘价计算 20 天移动均价，作为量价数据的输入之一，以此仿照真实的 K 线图。每张图包含 60 个交易日，每个交易日包含 3 个像素宽，因此图宽 180 个像素。高设定为 96 个像素，上方 72 个像素为价格数据构造的像素图，下方 24 个像素为交易量数据构造的像素图。将价格数据转化为图时，60 天的窗口期内，所有价格的最大值映射到最上面的像素点，所有价格的最小值映射到最下面的像素点，中间的价格由线性插值得到其像素点位置，每天开盘价的像素点位于三个像素点的左侧，收盘价位于右侧，最高价和最低价位于中间，移动均价像素点位于中间，左右两侧的移动均价像素点由插值得到。将交易量数据转化为图时，60 天窗口期内的最大交易量映射到最上方，最下方像素点对应交易量为 0，交易量位于中间的像素点。将交易价格像素图与交易量像素图上下拼接，构成了 60 天内的输入像素图，如图2 (a) 所示。Jiang et al. (2023) 认为，日内的收益率已经体现在开盘价和收盘价的相对位置中，因此不需要再添加额外的信息体现日内收益率。而 Lu and Wu (2022) 用额外的两个输入通道，将图片由原来的黑白变为 RGB 彩色图，提高了预测效果。事实上，使用额外的两个通道体现日内交易信息是对神经网络施加的约束。在输入层进行卷积时，如果只使用一个通道、即黑白图，所有输出通道均为自由状态，如果使用三个通道，即彩色图，则已经有两个通道被指定学习

表 10: 大单净流入率自相关性 (%)

	<i>FixedEffect</i>	$\beta_{t-1}$	$\beta_{t-2}$	$\beta_{t-3}$	$\beta_{t-4}$	$\beta_{t-5}$	$\bar{\beta}_{t-6 \sim t-20}$
$\tau = 1$	-2.88	11.93					
	[-158.80]	[88.19]					
$\tau = 5$	-2.52	10.67	5.11	3.61	2.23	1.10	
	[-152.93]	[88.46]	[74.25]	[56.93]	[36.00]	[17.46]	
$\tau = 20$	-2.15	10.36	4.81	3.25	1.86	0.55	0.82
	[-136.94]	[84.82]	[68.62]	[49.86]	[29.83]	[8.55]	[50.50]

表中汇报了大单净流入率的自相关性回归结果:

$$LargeOrder_{i,t} = FixedEffect_i + \sum_{lag=1}^{\tau} \beta_{i,lag} \times LargeOrder_{i,t-lag} + \varepsilon_{i,t}$$

*FixedEffect* 为固定效应,  $\beta_{t-1} \sim \beta_{t-5}$  分别为 1 ~ 5 个时间滞后的系数。 $\bar{\beta}_{t-6 \sim t-20}$  为 6 ~ 12 个滞后系数的平均值。方括号中的 t 值为样本截面均值, 即  $\sqrt{N} \frac{\text{mean}_i(\beta)}{\text{std}_i(\beta)}$ 。 $\tau = 1$ 、 $\tau = 5$ 、 $\tau = 20$  分别为滞后一天、滞后一天到五天、滞后一天到二十天的回归结果。

特定信息。这种约束事实上先验地认为日内收益率对未来收益率具有一定预测能力, 也降低了过拟合的程度。因此, 本文采用 RGB 三通道输入, 每个输入通道的值和颜色的对应关系如图2 (b) 所示。

## 附录 C 模型预测结果的稳健性检验

图5给出了正文章节4中按照大单净流入率组分类后, 每组内根据 CNN 输出的涨跌概率分成的十个资产组合的收益率曲线。

表11给出了正文章节4中按照大单净流入率分十组后, 每组的训练结果。

表12给出了正文章节4中在每个时间截面上, 随机抽样五个股票池并分别训练的每个股票池的训练结果。

表13给出了正文章节4.4中, 分别按照机构持仓比率、基金持仓比率、换手率、以及收盘价四个分组指标分组后的训练结果。

## 附录 D 基本技术指标定义及收益率

正文章节4.3中采用的除 *cum5* 和 *cum60* 外的十个技术指标定义如下:

### 1. 动量震荡 (*Momentum Oscillator*)

$$MomOs_{i,t} = \frac{1}{\tau} \sum_{s=t-\tau+1}^t \frac{close_{i,s} - low_{i,s}}{high_{i,s} - low_{i,s}}$$

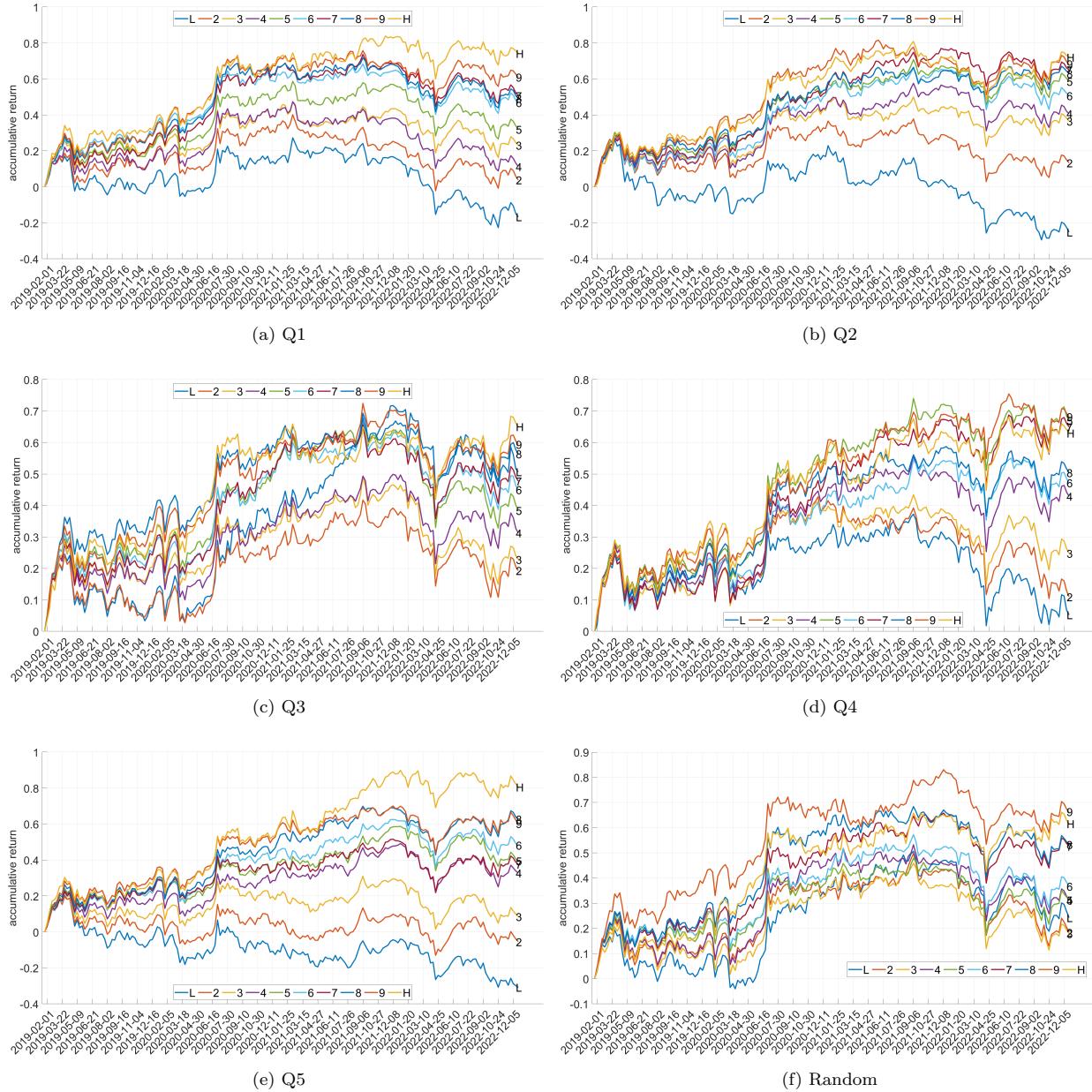


图 5: 每个股票池中 10 个投资组合的累积收益率曲线

图中给出了 Q1~Q5、Random 六个股票池中，根据神经网络预测概率分成的十个资产组合的累积收益率曲线。从 L 到 H，神经网络预测概率逐渐增大。

表 11: 资产组合周度收益率（%，按大单净流入率分十组）

	L	2	3	4	H	H - L	Improve	Sharpe
D1	-0.20 [-1.99]	-0.11 [-1.24]	0.01 [0.09]	0.05 [0.63]	-0.02 [-0.22]	0.18 [1.33]	0.00 [0.02]	0.68
D2	-0.17 [-1.63]	-0.05 [-0.59]	0.11 [1.41]	0.08 [0.91]	0.21 [2.08]	0.38 [2.70]	0.20 [1.12]	1.39
D3	-0.11 [-1.13]	-0.02 [-0.27]	0.11 [1.49]	0.17 [2.35]	0.23 [2.38]	0.34 [2.50]	0.16 [0.93]	1.28
D4	-0.11 [-0.96]	-0.10 [-1.40]	-0.02 [-0.24]	0.03 [0.35]	0.23 [2.34]	0.34 [2.20]	0.16 [0.86]	1.13
D5	-0.16 [-1.51]	-0.07 [-0.94]	0.03 [0.48]	0.12 [1.71]	0.19 [2.13]	0.35 [2.54]	0.17 [1.01]	1.31
D6	0.06 [0.62]	0.08 [0.97]	0.01 [0.15]	0.07 [0.99]	0.04 [0.40]	-0.02 [-0.15]	-0.20 [-1.10]	-0.08
D7	-0.11 [-1.07]	-0.07 [-0.88]	-0.07 [-0.96]	-0.05 [-0.63]	0.11 [1.29]	0.22 [1.69]	0.04 [0.24]	0.87
D8	0.02 [0.26]	0.07 [0.94]	0.07 [0.93]	0.16 [2.04]	0.16 [1.66]	0.13 [1.05]	-0.04 [-0.27]	0.54
D9	-0.17 [-1.74]	0.01 [0.11]	0.04 [0.53]	0.04 [0.50]	0.14 [1.56]	0.31 [2.31]	0.13 [0.77]	1.19
D10	-0.26 [-2.28]	-0.06 [-0.67]	-0.04 [-0.48]	0.06 [0.65]	0.25 [2.40]	0.50 [3.98]	0.32 [2.09]	2.05
Random	0.00 [0.05]	0.01 [0.17]	0.02 [0.30]	0.11 [1.52]	0.18 [2.04]	0.18 [1.36]	0.10 [0.81]	0.70

表中汇报了各股票池基于 K 线图识别的卷积神经网络预测收益率的表现。在每个股票池中，根据神经网络的输出概率，将股票分为五组，从 L 到 H，股票在未来五天中涨的概率逐渐增大。表中每组的值为各个股票收益率的市值加权平均相较于整个股票池的周度超额收益率。H - L 列汇报了多头 H 组，空头 L 组的投资组合的收益率。Improve 列为 D1~D10 股票池的多空组合相较于 Random 股票池多空组合的超额收益率，最后一行为等权持仓 D1~D10 股票池后，相较于 Random 组的超额收益率。Sharpe 列汇报了每个股票池的年化夏普比率。

表 12: 资产组合周度收益率 (%)，随机抽样并重复五次)

	L	2	3	4	5	6	7	8	9	H	H - L	Sharpe
Random1	-0.07 [-0.68]	-0.10 [-1.35]	-0.10 [-1.49]	-0.03 [-0.49]	-0.03 [-0.50]	-0.00 [-0.03]	0.08 [1.51]	0.09 [1.45]	0.16 [2.08]	0.13 [1.49]	0.20 [1.45]	0.75
Random2	-0.03 [-0.32]	-0.05 [-0.62]	0.05 [0.90]	0.05 [0.78]	-0.02 [-0.30]	-0.06 [-1.01]	0.02 [0.27]	0.06 [0.93]	0.00 [0.02]	0.16 [1.85]	0.19 [1.46]	0.75
Random3	0.01 [0.12]	0.07 [0.89]	-0.10 [-1.64]	-0.02 [-0.32]	-0.00 [-0.02]	-0.02 [-0.42]	0.06 [1.20]	0.05 [0.81]	0.14 [2.00]	0.19 [2.02]	0.18 [1.35]	0.69
Random4	0.06 [0.59]	0.02 [0.33]	0.02 [0.31]	-0.06 [-0.99]	-0.05 [-0.81]	0.03 [0.69]	-0.04 [-0.67]	0.04 [0.75]	0.17 [2.69]	0.21 [2.27]	0.16 [1.15]	0.59
Random5	-0.09 [-1.00]	-0.05 [-0.68]	-0.12 [-1.93]	-0.02 [-0.27]	0.05 [0.90]	0.11 [1.86]	0.03 [0.54]	0.09 [1.55]	0.01 [0.22]	0.13 [1.45]	0.23 [1.60]	0.82

表中汇报了五个随机组的卷积神经网络预测表现。在每个截面上随机抽取五分之一的样本并重复五次得到每个 Random 组的股票池。

## 2. 高低价比率 (High-Low Ratio)

$$HLRatio_{i,t} = \frac{\max_{s=t-\tau+1}^t high_{i,s}}{\min_{s=t-\tau+1}^t low_{i,s}}$$

## 3. 相对强弱指数 (Relative Strength Index)

$$RSI_{i,t} = 1 - \frac{1}{1 + RS_{i,t}}, \quad RS_{i,t} = \frac{\sum_{s=t-\tau+1}^t |Ret_{i,s}^{up}|}{\sum_{s=t-\tau+1}^t |Ret_{i,s}^{down}|}$$

## 4. 交易量指数 (Volume Indicator)

$$VolInd_{i,t} = \frac{volume_{i,t}}{\frac{1}{\tau} \sum_{s=t-\tau+1}^t volume_{i,s}}$$

## 5. 威廉指数 (Williams Index)

$$Williams_{i,t} = \frac{\max_{s=t-\tau+1}^t high_{i,s} - close_{i,t}}{\max_{s=t-\tau+1}^t high_{i,s} - \min_{s=t-\tau+1}^t low_{i,s}}$$

## 6. 开收价差

$$OCDiff_{i,t} = \sum_{s=t-\tau+1}^t (close_{i,s} - open_{i,s})$$

表 13: 资产组合周度收益率（其他分类指标，%）

	L	2	3	4	5	6	7	8	9	H	H - L	Improve	Sharpe
Panel A: Institution Holding Ratio													
Q1	-0.29 [-1.90]	-0.08 [-0.64]	0.06 [0.51]	0.10 [0.78]	0.14 [1.17]	0.14 [1.18]	0.18 [1.50]	0.20 [1.59]	0.18 [1.46]	0.18 [1.35]	0.47 [3.47]	0.27 [1.67]	1.78
Q2	-0.10 [-0.76]	0.01 [0.08]	0.03 [0.24]	0.13 [1.15]	0.07 [0.59]	0.12 [1.11]	0.18 [1.56]	0.26 [2.22]	0.20 [1.74]	0.26 [2.14]	0.37 [3.05]	0.17 [1.02]	1.57
Q3	-0.34 [-2.68]	-0.13 [-1.18]	-0.09 [-0.85]	-0.09 [-1.01]	0.02 [0.24]	-0.02 [-0.18]	0.02 [0.20]	0.15 [1.60]	0.20 [2.02]	0.23 [1.90]	0.57 [3.89]	0.37 [2.09]	2.00
Q4	-0.20 [-1.94]	-0.10 [-1.32]	-0.04 [-0.62]	-0.04 [-0.75]	-0.03 [-0.51]	0.08 [1.48]	0.08 [1.35]	0.08 [1.59]	0.11 [1.80]	0.20 [2.30]	0.40 [2.75]	0.21 [1.28]	1.42
Q5	-0.13 [-1.34]	-0.11 [-1.74]	-0.09 [-1.51]	-0.03 [-0.50]	-0.03 [-0.44]	-0.01 [-0.10]	0.05 [0.84]	0.04 [0.66]	0.06 [1.00]	0.12 [1.53]	0.24 [2.02]	0.05 [0.31]	1.04
Random	-0.07 [-0.68]	-0.10 [-1.35]	-0.10 [-1.49]	-0.03 [-0.49]	-0.03 [-0.50]	-0.00 [-0.03]	0.08 [1.51]	0.09 [1.45]	0.16 [2.08]	0.13 [1.49]	0.20 [1.45]	0.21 [1.68]	0.75
Panel B: Fund Holding Ratio													
Q1	-0.42 [-2.46]	-0.15 [-0.97]	-0.08 [-0.53]	-0.02 [-0.13]	0.03 [0.21]	0.04 [0.26]	0.06 [0.47]	0.17 [1.24]	0.26 [2.01]	0.29 [2.13]	0.71 [6.89]	0.51 [3.42]	3.54
Q2	-0.31 [-2.05]	-0.21 [-1.66]	-0.10 [-0.87]	-0.06 [-0.55]	0.03 [0.24]	0.10 [0.88]	0.15 [1.26]	0.14 [1.21]	0.22 [1.89]	0.25 [1.94]	0.56 [4.13]	0.36 [2.23]	2.12
Q3	-0.26 [-1.93]	-0.15 [-1.36]	-0.11 [-1.00]	-0.09 [-0.89]	-0.04 [-0.35]	0.06 [0.59]	-0.02 [-0.17]	0.06 [0.58]	0.11 [1.20]	0.14 [1.24]	0.40 [2.73]	0.20 [1.20]	1.40
Q4	-0.13 [-1.34]	-0.12 [-1.62]	-0.08 [-1.09]	-0.02 [-0.26]	0.02 [0.24]	-0.02 [-0.28]	0.00 [0.03]	0.02 [0.27]	0.03 [0.55]	0.10 [1.16]	0.23 [1.99]	0.03 [0.20]	1.03
Q5	-0.04 [-0.45]	0.02 [0.30]	0.03 [0.37]	0.00 [0.06]	0.11 [1.44]	0.11 [1.40]	0.06 [0.66]	0.10 [1.16]	0.13 [1.43]	0.29 [2.71]	0.34 [2.77]	0.14 [0.87]	1.43
Random	-0.07 [-0.68]	-0.10 [-1.35]	-0.10 [-1.49]	-0.03 [-0.49]	-0.03 [-0.50]	-0.00 [-0.03]	0.08 [1.51]	0.09 [1.45]	0.16 [2.08]	0.13 [1.49]	0.20 [1.45]	0.25 [1.99]	0.75

续表13: 资产组合周度收益率 (其他分类指标, %)

	L	2	3	4	5	6	7	8	9	H	H - L	Improve	Sharpe
Panel C: Turnover													
Q1	-0.19 [-2.23]	-0.08 [-0.94]	-0.14 [-1.82]	-0.10 [-1.25]	-0.03 [-0.40]	0.02 [0.22]	0.01 [0.11]	0.06 [0.71]	0.15 [1.63]	0.06 [0.55]	0.25 [2.41]	0.05 [0.32]	1.24
Q2	0.07 [0.77]	0.01 [0.12]	0.03 [0.53]	0.07 [1.07]	0.04 [0.75]	0.13 [2.15]	0.20 [3.23]	0.13 [2.12]	0.17 [2.66]	0.21 [2.48]	0.14 [1.07]	-0.06 [-0.31]	0.55
Q3	-0.08 [-0.85]	0.01 [0.13]	0.11 [1.42]	0.02 [0.21]	0.08 [0.95]	0.12 [1.43]	0.21 [2.33]	0.11 [1.28]	0.26 [2.76]	0.22 [1.90]	0.30 [2.31]	0.11 [0.57]	1.19
Q4	-0.10 [-0.73]	-0.01 [-0.12]	-0.03 [-0.23]	0.03 [0.29]	0.03 [0.29]	0.07 [0.63]	0.13 [1.15]	0.03 [0.26]	0.10 [0.79]	0.21 [1.38]	0.31 [2.14]	0.11 [0.59]	1.10
Q5	-1.01 [-4.02]	-0.59 [-2.71]	-0.49 [-2.35]	-0.36 [-1.78]	-0.30 [-1.50]	-0.26 [-1.36]	-0.11 [-0.54]	-0.07 [-0.35]	0.01 [0.07]	0.10 [0.51]	1.11 [5.34]	0.91 [4.52]	2.75
Random	-0.07 [-0.68]	-0.10 [-1.35]	-0.10 [-1.49]	-0.03 [-0.49]	-0.03 [-0.50]	-0.00 [-0.03]	0.08 [1.51]	0.09 [1.45]	0.16 [2.08]	0.13 [1.49]	0.20 [1.45]	0.22 [1.66]	0.75
Panel D: Close Price													
Q1	-0.13 [-0.92]	-0.04 [-0.31]	0.07 [0.53]	0.10 [0.82]	0.14 [1.19]	0.15 [1.29]	0.13 [1.10]	0.14 [1.16]	0.15 [1.27]	0.17 [1.46]	0.30 [2.62]	0.10 [0.66]	1.35
Q2	-0.35 [-2.62]	-0.16 [-1.44]	-0.10 [-0.99]	-0.04 [-0.41]	-0.07 [-0.73]	-0.04 [-0.40]	0.02 [0.18]	0.04 [0.41]	0.04 [0.38]	0.07 [0.66]	0.43 [3.53]	0.23 [1.53]	1.81
Q3	-0.23 [-2.17]	-0.16 [-1.56]	-0.10 [-0.96]	-0.03 [-0.28]	-0.02 [-0.25]	-0.02 [-0.16]	0.03 [0.33]	0.12 [1.30]	0.16 [1.65]	0.20 [1.82]	0.43 [3.92]	0.23 [1.44]	2.02
Q4	-0.18 [-1.60]	-0.12 [-1.38]	-0.09 [-1.12]	-0.04 [-0.45]	0.10 [1.24]	0.05 [0.59]	0.05 [0.60]	0.16 [2.00]	0.21 [2.71]	0.25 [2.31]	0.42 [3.34]	0.23 [1.45]	1.72
Q5	0.01 [0.12]	-0.02 [-0.21]	-0.04 [-0.47]	-0.03 [-0.35]	0.04 [0.44]	0.02 [0.19]	0.01 [0.14]	0.17 [1.79]	0.16 [1.53]	0.26 [2.12]	0.25 [1.99]	0.05 [0.31]	1.02
Random	-0.07 [-0.68]	-0.10 [-1.35]	-0.10 [-1.49]	-0.03 [-0.49]	-0.03 [-0.50]	-0.00 [-0.03]	0.08 [1.51]	0.09 [1.45]	0.16 [2.08]	0.13 [1.49]	0.20 [1.45]	0.17 [1.37]	0.75

表中汇报了各股票池基于 K 线图识别的卷积神经网络预测收益率的表现, 与表2类似。Panel A 汇报了分组准则为机构持仓占比的结果, Panel B 汇报了分组准则为基金持仓占比的结果, Panel C 汇报了分组准则为换手率的结果, Panel D 汇报了分组准则为收盘价的结果。

### 7. 历史波动率 (*Historical Volatility*)

$$HisVol_{i,t} = \text{Std}_{s=t-\tau+1}^t (Ret_{i,s})$$

### 8. 真实震荡幅度 (*Real Oscillate Magnitude*)

$$RealOsMag_{i,t} = \frac{1}{\tau} \sum_{s=t-\tau+1}^t high_{i,s} - low_{i,s}$$

### 9. 最大回撤 (*Maximal Drawback*)

$$MaxBack_{i,t} = \frac{\max_{s=t-\tau+1}^t close_{i,s} - \min_{s=t-\tau+1}^t close_{i,s}}{close_{t-\tau+1}}$$

### 10. 交易量加权平均收盘价 (*Moving Average Price of Volume*)

$$MAVol_{i,t} = \frac{\sum_{s=t-\tau+1}^t close_{i,s} \times volume_{i,s}}{\sum_{s=t-\tau+1}^t volume_{i,s}}$$

以上指标表达式中，时间间隔  $\tau = 60$ ,  $open$ 、 $high$ 、 $low$ 、 $close$  分别为开盘价、高价、低价、以及收盘价， $volume$ 、 $ret$  为交易量和收益率。每个技术指标因子的周度收益率如表14所示。

## 附录 E 颜色特征指标描述性统计

正文章节5.1中的颜色特征指标在每个股票池中的描述性统计如表15所示。在不同股票池中，颜色特征指标都具有相似的值，并且在时间序列上有很强的稳健性。表16给出了每个颜色特征指标因子的周度收益率。

## 附录 F 光滑梯度类别激活映射 ++ 模型

正文章节5.2中采用的光滑梯度类别激活映射 ++ 模型技术是一种通过梯度反映神经网络对像素图不同部分敏感性的方法。对于任意类别  $c$ ，神经网络的输出为  $Y^c$ ；输入的特征图通道数为  $K$ ，例如一个 RGB 彩色图， $K = 3$ ；每一张特征图在通道  $k$ ，坐标为  $i, j$  的位置处的值为  $A_{i,j}^k$ ，定义输出  $Y^c$  对于输入  $A_{i,j}^k$  的三阶导数为

$$D_1^k = \frac{\partial Y^c}{\partial A_{i,j}^k}, D_2^k = \frac{\partial^2 Y^c}{\partial (A_{i,j}^k)^2}, D_3^k = \frac{\partial^3 Y^c}{\partial (A_{i,j}^k)^3}. \quad (16)$$

(Smilkov et al., 2017) 发现，在计算灵敏图时，在原图的基础上添加白噪声可以对灵敏图起到降噪作用。因此在计算任意一个点的偏导数时，首先需要将这个点的值  $A_{i,j}^k$  添加一个微小扰动  $\varepsilon$  并重复  $n$

表 14: 基本技术指标因子收益率 (%)

		<i>cum</i>	<i>cum</i>	<i>Mom</i>	<i>HL</i>	<i>RSI</i>	<i>Vol</i>	<i>Will-</i>	<i>OC</i>	<i>His</i>	<i>Real</i>	<i>Max</i>	<i>MA</i>
		5	60	<i>O<sub>s</sub></i>	<i>Ratio</i>	<i>Ind</i>	<i>iams</i>	<i>Diff</i>	<i>Vol</i>	<i>OsMag</i>	<i>Back</i>	<i>Max</i>	<i>Vol</i>
Q1		-0.21 [-1.11]	-0.27 [-1.21]	-0.32* [-1.72]	-0.06 [-0.29]	-0.21 [-1.00]	-0.34* [-1.98]	0.06 [0.27]	-0.15 [-0.77]	-0.15 [-0.68]	0.06 [0.25]	-0.07 [-0.31]	0.06 [0.25]
Q2		-0.29* [-1.70]	-0.09 [-0.42]	-0.20 [-1.05]	0.03 [0.13]	-0.04 [-0.19]	-0.26* [-1.86]	0.02 [0.10]	0.08 [0.41]	-0.11 [-0.46]	0.17 [0.85]	0.03 [0.14]	0.18 [0.91]
Q3		-0.31* [-1.74]	-0.11 [-0.55]	-0.14 [-0.84]	0.12 [0.62]	-0.11 [-0.59]	-0.05 [-0.30]	0.15 [0.81]	-0.08 [-0.45]	0.06 [0.26]	0.06 [0.00]	0.10 [0.51]	-0.03 [-0.16]
Q4		-0.12 [-0.71]	-0.06 [-0.25]	-0.07 [-0.38]	0.11 [0.56]	-0.02 [-0.09]	-0.02 [-0.16]	-0.11 [-0.57]	0.05 [0.26]	0.13 [0.64]	0.18 [1.02]	0.11 [0.59]	0.09 [0.51]
Q5		-0.00 [-0.03]	-0.05 [-0.25]	-0.11 [-0.66]	0.21 [1.02]	-0.01 [-0.05]	-0.10 [-0.72]	-0.08 [-0.47]	-0.09 [-0.49]	0.18 [0.87]	0.07 [0.42]	0.25 [1.28]	-0.00 [-0.03]
Random		-0.17 [-0.93]	-0.21 [-0.92]	-0.29 [-1.58]	0.17 [0.81]	-0.17 [-0.82]	-0.24 [-1.60]	0.04 [0.20]	-0.05 [-0.26]	0.18 [0.87]	0.12 [0.61]	0.20 [0.94]	0.16 [0.81]

表中汇报了十二个基本技术指标的投資组合的周度收益率。首先根据每个技术指标排序后的前 1/3 与后 1/3 的股票根据流通市值加权平均得到多头资产组合和空头资产组合，然后作差得到多空组合的周度收益率。

表 15: 颜色特征指标描述性统计 (%)

	<i>ColorInfoRatio</i>	<i>RedRatio</i>	<i>GreenRatio</i>	<i>RedAvg</i>	<i>GreenAvg</i>
Q1	7.78	4.18	3.60	24.52	23.21
Q2	7.78	4.18	3.60	24.51	23.20
Q3	7.77	4.17	3.60	24.51	23.18
Q4	7.77	4.17	3.61	24.46	23.10
Q5	7.75	4.14	3.60	24.20	22.86
Random	7.77	4.17	3.60	24.45	23.12

表中汇报了五个颜色特征指标的均值。*ColorInfoRatio* 为 K 线图区域占全图的比例; *RedRatio* 为红色区域占全图的比例; *RedRatio* 为绿色区域占全图的比例; *RedAvg* 以及 *GreenAvg* 的定义为

$$\begin{aligned} RedAvg_{i,t} &= \frac{\sum_{h=1}^H \sum_{w=1}^W \mathbf{1}_{R_{i,t,h,w}=255} \times \left(1 - \frac{G_{i,t,h,w}}{255}\right)}{\sum_{h=1}^H \sum_{w=1}^W \mathbf{1}_{R_{i,t,h,w}=255}}, \\ GreenAvg_{i,t} &= \frac{\sum_{h=1}^H \sum_{w=1}^W \mathbf{1}_{G_{i,t,h,w}=255} \times \left(1 - \frac{R_{i,t,h,w}}{255}\right)}{\sum_{h=1}^H \sum_{w=1}^W \mathbf{1}_{G_{i,t,h,w}=255}} \end{aligned} \quad (15)$$

次求均值，作为这一点的真实值，其中， $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ 。本文采用 Smilkov et al. (2017) 中的参数取值， $n = 4, \sigma = 0.15$ 。于是通道  $k$  的特征图在位置  $i, j$  处的权重为

$$\alpha_{i,j}^{kc} = \frac{\frac{1}{n} \sum_1^n D_1^k}{\frac{2}{n} \sum_1^n D_2^k + \sum_a \sum_b A_{a,b}^k \frac{1}{n} \sum_1^n D_3^k}. \quad (17)$$

通道  $k$  的权重为其各个位置一阶导数经过 *ReLU* 激活层后的加权平均，即

$$W_k^c = \sum_i \sum_j \alpha_{i,j}^{kc} ReLU \left( \frac{1}{n} \sum_1^n D_1^k \right) \quad (18)$$

最终将每个通道加权，得到注意力热力图

$$L_{Grad-CAM}^c = ReLU \left( \sum_k W_k^c A^k \right). \quad (19)$$

图6给出了 SmoothGradCAM++ 输出的热力图。

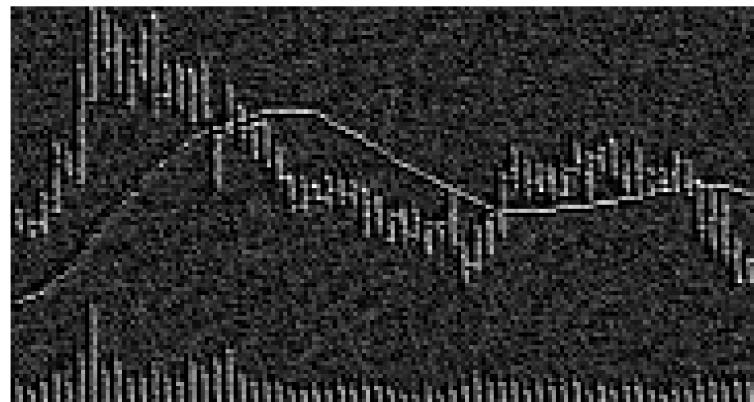
正文章节 5.2 中，使用注意力指标构造的因子收益率如表 17 所示。

正文章节 5.2 中，考虑边界的注意力因子载荷如表 18 所示。

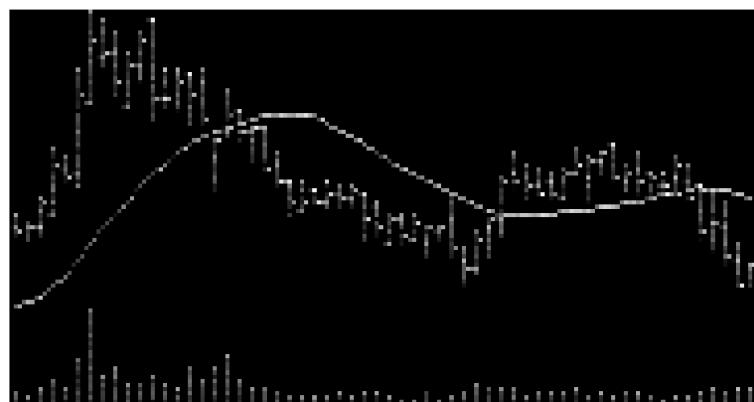
表 16: 颜色特征指标因子收益率 (%)

	<i>ColorInfoRatio</i>	<i>RedRatio</i>	<i>GreenRatio</i>	<i>RedAvg</i>	<i>GreenAvg</i>
Q1	0.08 [0.44]	-0.07 [-0.40]	0.17 [0.98]	-0.17 [-0.83]	-0.05 [-0.23]
Q2	0.12 [0.84]	0.09 [0.59]	0.17 [1.06]	-0.19 [-0.93]	-0.11 [-0.50]
Q3	-0.02 [-0.11]	0.06 [0.36]	0.05 [0.30]	0.08 [0.40]	0.07 [0.32]
Q4	0.13 [0.95]	0.19 [1.18]	0.07 [0.48]	0.07 [0.37]	0.05 [0.26]
Q5	0.06 [0.40]	-0.06 [-0.37]	0.14 [0.90]	0.07 [0.38]	0.09 [0.44]
Random	0.08 [0.51]	-0.04 [-0.23]	0.22 [1.41]	0.18 [0.93]	0.11 [0.51]

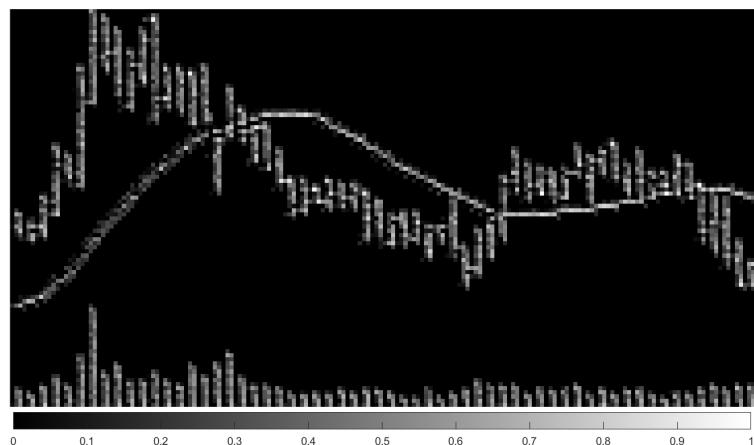
表中汇报了五个颜色特征指标的投资组合的周度收益率。首先根据每个颜色特征指标排序后的前 1/3 与后 1/3 的股票根据流通市值加权平均得到多头资产组合和空头资产组合，然后作差得到多空组合的周度收益率。



(a) 原始输出



(b) 去噪后



(c) 包含边缘

图 6: SmoothGradCAM++ 注意力热力图

图为 SmoothGradCAM++ 算法给出的神经网络注意力分布热力图，用于构建资产组合的十个模型平均后的结果，越亮的地方代表着越高的注意力。图 (a) 为原始输出，图 (b) 为去掉噪音后的输出，图 (c) 在 (b) 的基础上，包含了 K 线图边缘上的注意力信息。

表 17: 注意力因子收益率 (%)

	Price+ Volume	Price	Volume	Open	Close	Open + Close	High + Low
Q1	0.06 [0.51]	0.04 [0.31]	0.05 [0.40]	0.02 [0.20]	0.05 [0.36]	0.14 [1.09]	-0.00 [-0.04]
Q2	0.17 [1.59]	0.20* [1.77]	0.03 [0.33]	0.25** [2.24]	-0.06 [-0.48]	0.13 [1.13]	0.32*** [2.88]
Q3	0.02 [0.23]	0.11 [1.03]	-0.09 [-0.86]	0.16 [1.64]	0.15 [1.36]	0.15 [1.33]	0.05 [0.43]
Q4	0.02 [0.15]	0.03 [0.26]	-0.01 [-0.08]	0.12 [1.11]	-0.15 [-1.36]	-0.02 [-0.20]	0.02 [0.23]
Q5	0.27*** [2.86]	0.29*** [2.95]	0.24** [2.43]	0.34*** [3.26]	0.15 [1.51]	0.26** [2.54]	0.25*** [2.75]
Random	0.02 [0.19]	0.07 [0.61]	-0.05 [-0.39]	0.07 [0.58]	0.21** [2.04]	0.19* [1.69]	0.04 [0.34]

表中汇报了七个注意力指标的投资组合的周度收益率。首先根据每个注意力指标排序后的前 1/3 与后 1/3 的股票根据流通市值加权平均得到多头资产组合和空头资产组合，然后作差得到多空组合的周度收益率。

表 18: 考虑边界的注意力分布因子载荷

	H - L	Price				Price				Open				Open + Close		
		+ Volume		and Volume		and Close		Open		High		Open + Low				
		Price (%)	Volume (%)	Price (%)	Volume (%)	$R^2$	$\alpha$ (%)	Open	Close	$R^2$	$\alpha$ (%)	Open	High	$R^2$		
Q1	0.48*** [3.38]	0.54*** [3.90]	0.33*** [2.84]	0.14 [0.97]	0.55*** [3.97]	0.20 [0.92]	0.19 [3.99]	0.17 [2.37]	0.55*** [0.67]	0.24** [4.07]	0.09 [1.04]	0.14 [1.04]	0.55*** [2.87]	0.13 [0.13]	0.30*** [0.19]	
Q2	0.51*** [3.42]	0.56*** [3.49]	0.27** [2.40]	0.11 [3.39]	0.55*** [1.00]	0.16 [0.94]	0.14 [3.09]	0.11 [3.01]	0.50*** [-0.36]	0.33*** [3.36]	-0.04 [2.30]	0.12 [2.30]	0.55*** [-0.52]	0.32** [0.36]	-0.07 [0.36]	0.11 [0.11]
Q3	0.08 [0.51]	0.09 [0.58]	0.10 [0.77]	0.06 [0.48]	0.07 [2.98]	0.40*** [-1.56]	-0.26 [0.44]	0.09 [0.98]	0.07 [0.44]	0.12 [0.44]	0.08 [0.44]	0.07 [0.44]	0.07 [0.44]	-0.02 [0.44]	0.36*** [3.36]	0.12 [0.12]
Q4	0.31** [2.56]	0.31*** [3.07]	0.08 [0.83]	0.07 [3.00]	0.30*** [1.35]	0.17 [-0.52]	-0.07 [3.00]	0.08 [0.65]	0.31*** [1.11]	0.04 [1.11]	0.10 [1.11]	0.08 [1.11]	0.30*** [1.11]	0.12 [1.11]	0.03 [1.11]	0.08 [0.31]
Q5	0.59*** [4.61]	0.46*** [4.19]	0.45*** [4.38]	0.22 [4.26]	0.46*** [1.48]	0.33 [0.64]	0.12 [4.26]	0.22 [2.04]	0.47*** [2.10]	0.25*** [3.64]	0.25*** [3.15]	0.22 [3.15]	0.41*** [3.15]	0.33*** [3.15]	0.26** [2.52]	0.27 [0.27]
Random	0.20 [1.45]	0.22 [1.37]	0.02 [0.96]	0.22 [1.31]	0.15 [0.89]	0.02 [0.09]	0.02 [1.33]	0.02 [-0.10]	0.22 [0.88]	-0.01 [1.30]	0.08 [1.30]	0.01 [1.30]	0.22 [1.30]	0.18 [1.30]	-0.02 [1.30]	0.03 [1.30]

46

表中汇报了每个股票池中，累积神经网络的资产组合在考虑边界的注意力因子上的因子载荷，回归方程为：

$$Ret_{i,t} = \alpha_i + \sum_{j=1}^n \beta_{i,j} \times Attention_{i,j,t} + \sum_{j=1}^m \gamma_{i,j,t} \times FF5_{i,j,t} + \varepsilon_{i,t},$$

其中  $Attention_{i,j,t}$  为第  $i$  个股票池、 $j$  类型、在  $t$  时刻的注意力因子时间序列。控制变量  $FF5_{i,j,t}$  为 Fama French 五因子。

## 附录 G 图型技术指标构造

正文章节5.3中采用的 Andrew Lo 图型技术指标的定义方式如下：基于 K 线图的最后 5 个极值点  $E_1, \dots, E_5$ ,

### 1. HS 和 IHS

$$HS \equiv \begin{cases} E_1 \in Maximum \\ E_3 > E_1, E_5 \\ (1 - 1.5\%) \frac{E_1 + E_5}{2} \leq E_1, E_5 \leq (1 + 1.5\%) \frac{E_1 + E_5}{2} \\ (1 - 1.5\%) \frac{E_2 + E_4}{2} \leq E_2, E_4 \leq (1 + 1.5\%) \frac{E_2 + E_4}{2} \end{cases} \quad (20)$$

$$IHS \equiv \begin{cases} E_1 \in Minimum \\ E_3 < E_1, E_5 \\ (1 - 1.5\%) \frac{E_1 + E_5}{2} \leq E_1, E_5 \leq (1 + 1.5\%) \frac{E_1 + E_5}{2} \\ (1 - 1.5\%) \frac{E_2 + E_4}{2} \leq E_2, E_4 \leq (1 + 1.5\%) \frac{E_2 + E_4}{2} \end{cases}$$

### 2. BTOP 和 BBOT

$$BTOP \equiv \begin{cases} E_1 \in Maximum \\ E_1 < E_3 < E_5 \\ E_2 > E_4 \end{cases} \quad BBOT \equiv \begin{cases} E_1 \in Minimum \\ E_1 > E_3 > E_5 \\ E_2 < E_4 \end{cases} \quad (21)$$

### 3. TTOP 和 TBOT

$$TTOP \equiv \begin{cases} E_1 \in Maximum \\ E_1 > E_3 > E_5 \\ E_2 < E_4 \end{cases} \quad TBOT \equiv \begin{cases} E_1 \in Minimum \\ E_1 < E_3 < E_5 \\ E_2 > E_4 \end{cases} \quad (22)$$

#### 4. RTOP 和 RBOT

$$RTOP \equiv \begin{cases} E_1 \in Maximum \\ (1 - 0.75\%) \frac{E_1 + E_3 + E_5}{3} \leq E_1, E_3, E_5 \leq (1 + 0.75\%) \frac{E_1 + E_3 + E_5}{2} \\ (1 - 0.75\%) \frac{E_2 + E_4}{2} \leq E_2, E_4 \leq (1 + 0.75\%) \frac{E_2 + E_4}{2} \\ \min(E_1, E_3, E_5) > \max(E_2, E_4) \end{cases} \quad (23)$$

$$RBOT \equiv \begin{cases} E_1 \in Minimum \\ (1 - 0.75\%) \frac{E_1 + E_3 + E_5}{3} \leq E_1, E_3, E_5 \leq (1 + 0.75\%) \frac{E_1 + E_3 + E_5}{2} \\ (1 - 0.75\%) \frac{E_2 + E_4}{2} \leq E_2, E_4 \leq (1 + 0.75\%) \frac{E_2 + E_4}{2} \\ \min(E_2, E_4) > \max(E_1, E_3, E_5) \end{cases}$$

此处介绍了正文章节5.3中采用的高斯平滑方法。第  $t$  天的收盘价  $P_t$  的核估计为

$$\hat{m}_\sigma(t) = \frac{\sum_{\tau=t-r}^{\tau=t+r} K_\sigma(t-\tau) P_\tau}{\sum_{\tau=t-r}^{\tau=t+r} K_\sigma(t-\tau)}, \quad (24)$$

$$K_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}.$$

其中  $r$  为核半径,  $\sigma$  为核标准差。使用核估计结果  $\hat{m}_\sigma(t)$  判断极值点前, 需要确定  $\sigma$  时。Lo et al 使用截面验证 (cross-sectional validation) 方法, 最优化界面误差函数

$$CV(\sigma) = \sum_{t=1}^T (P_t - \hat{m}_\sigma(t))^2, \quad (25)$$

$$\hat{m}_\sigma(t) = \frac{\sum_{\tau \neq t} K_\sigma(t-\tau) P_\tau}{\sum_{\tau \neq t} K_\sigma(t-\tau)}.$$

公式25中的  $\hat{m}_\sigma(t)$  为公式24中的  $\hat{m}_\sigma(t)$  去掉  $P_t$  后的结果。 $\sigma^*$  为最小化  $CV(\sigma)$  后的方差,  $\sigma^* \in \arg \min_\sigma CV(\sigma)$ , 用于高斯平滑的方差  $\sigma = 0.3 \times \sigma^*$ . 本文并没有采用这种优化方式, 因为在  $3 \times \sigma$  范围内截断得到的核估计的结果并没有明显差别。因此, 本文使用 5 个交易日作为核半径, 在  $3 \times \sigma$  范围外截断, 方差  $\sigma = 5/3$ 。图7给出了高斯核平滑法和经验模态分解法下, 对于图型技术指标的判断结果。

正文章节5.3中采用的高斯平滑方法和经验模态分解方法得到图型技术指标数量及其占比如表19所示。多头具有某个图型技术指标的股票, 并空头不具备任何图型技术指标的股票, 得到的资产组合即为这种图形技术指标的因子值。表20给出了八种图型技术指标在两种平滑方法下的因子收益率。表21给出了高斯平滑方法得到的因子时间序列与经验模态分解得到的因子时间序列的相关性, 方括号中为 p 值。

正文章节5.3中, 采用经验模态分解方法得到的因子载荷如表22所示。

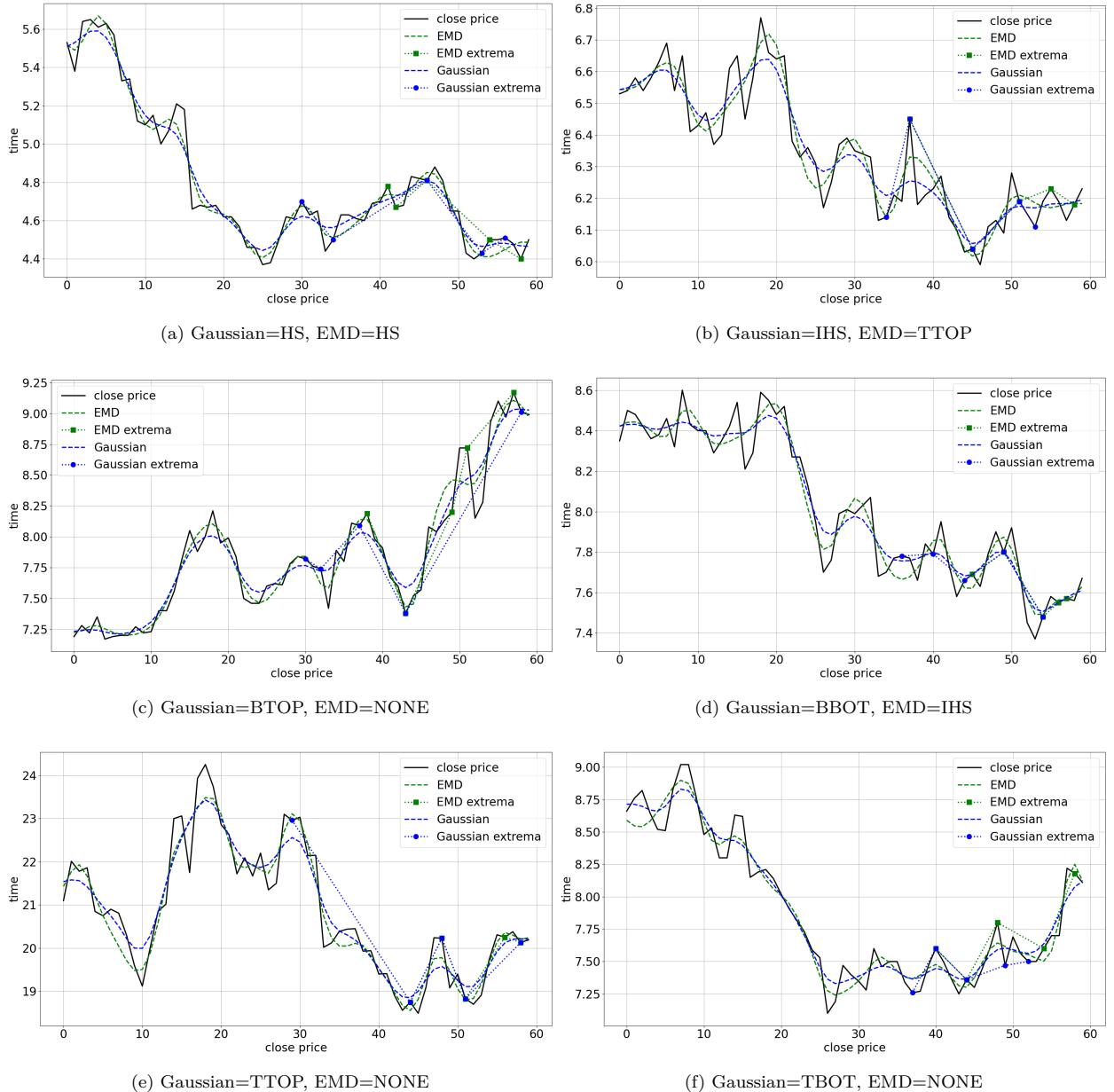


图 7: 收盘价平滑处理后技术八类指标判断图

表 19: 图型技术指标描述性统计

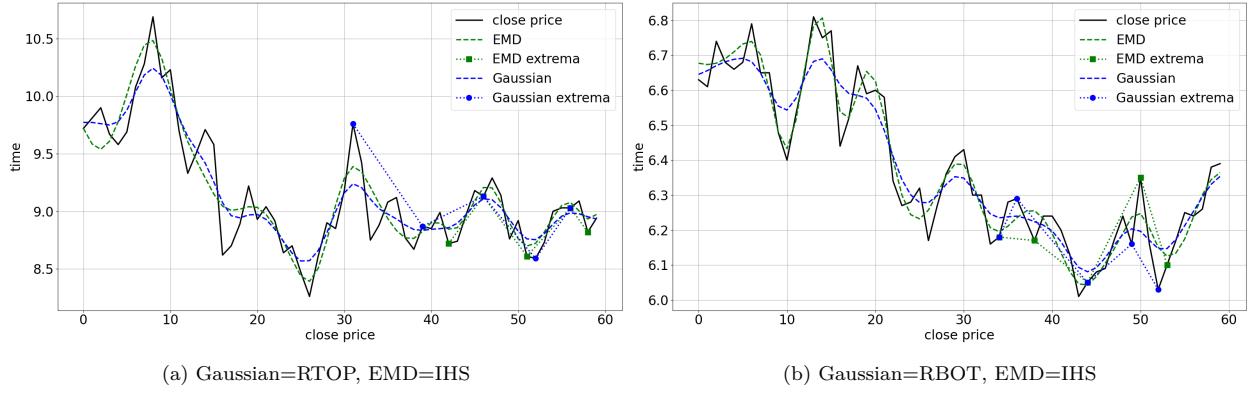
	NONE	HS	IHS	BTOP	BBOT	TTOP	TBOT	RTOP	RBOT	sum
Panel A: Gaussian Smoothing										
Q1	324059 [63.36]	34882 [6.82]	42875 [8.38]	10746 [2.10]	10460 [2.05]	11872 [2.32]	11390 [2.23]	30731 [6.01]	34420 [6.73]	511435 [100.00]
Q2	323471 [63.25]	35179 [6.88]	42658 [8.34]	10764 [2.10]	10350 [2.02]	11962 [2.34]	11373 [2.22]	31398 [6.14]	34280 [6.70]	511435 [100.00]
Q3	322474 [63.05]	36593 [7.15]	41793 [8.17]	11223 [2.19]	10282 [2.01]	12330 [2.41]	10606 [2.07]	32525 [6.36]	33609 [6.57]	511435 [100.00]
Q4	322345 [63.03]	39042 [7.63]	38960 [7.62]	11542 [2.26]	9929 [1.94]	13425 [2.62]	9662 [1.89]	34450 [6.74]	32080 [6.27]	511435 [100.00]
Q5	318511 [62.05]	44838 [8.73]	36054 [7.02]	11303 [2.20]	9147 [1.78]	14994 [2.92]	8307 [1.62]	39888 [7.77]	30286 [5.90]	513328 [100.00]
Random	322061 [62.97]	37995 [7.43]	40383 [7.90]	10961 [2.14]	9993 [1.95]	13111 [2.56]	10198 [1.99]	33881 [6.62]	32852 [6.42]	511435 [100.00]
Panel B: Empirical Mode Decomposition										
Q1	320460 [62.66]	32782 [6.41]	40805 [7.98]	10938 [2.14]	11761 [2.30]	10047 [1.96]	10226 [2.00]	22544 [4.41]	51872 [10.14]	511435 [100.00]
Q2	320663 [62.70]	32996 [6.45]	40195 [7.86]	11454 [2.24]	11368 [2.22]	10202 [1.99]	9942 [1.94]	23136 [4.52]	51479 [10.07]	511435 [100.00]
Q3	320994 [62.76]	34735 [6.79]	38612 [7.55]	12089 [2.36]	11068 [2.16]	10423 [2.04]	9385 [1.84]	24329 [4.76]	49800 [9.74]	511435 [100.00]
Q4	322400 [63.04]	37223 [7.28]	36343 [7.11]	12973 [2.54]	10057 [1.97]	11098 [2.17]	8606 [1.68]	26106 [5.10]	46629 [9.12]	511435 [100.00]
Q5	320492 [62.43]	42878 [8.35]	33929 [6.61]	13035 [2.54]	9183 [1.79]	12946 [2.52]	7795 [1.52]	29537 [5.75]	43533 [8.48]	513328 [100.00]
Random	320548 [62.68]	35965 [7.03]	38043 [7.44]	12048 [2.36]	10728 [2.10]	10913 [2.13]	9092 [1.78]	25593 [5.00]	48505 [9.48]	511435 [100.00]

表中汇报了两种平滑方法在各个股票池中的技术指标分布情况。Panel A 为高斯平滑方法；Panel B 为经验模态分级。方括号中为该种图型技术指标在所有样本中的占比。

表 20: 图型技术指标因子收益率 (%)

	HS	IHS	BTOP	BBOT	TTOP	TBOT	RTOP	RBOT
Panel A: Gaussian Smoothing								
Q1	0.20 [1.03]	-0.15 [-0.71]	0.11 [0.44]	0.50 [1.42]	0.06 [0.27]	-0.48 [-1.57]	0.28 [1.45]	-0.21 [-0.98]
Q2	-0.24 [-1.27]	-0.08 [-0.38]	-0.17 [-0.70]	0.07 [0.27]	-0.02 [-0.09]	-0.22 [-0.79]	-0.06 [-0.34]	-0.36 [-1.50]
Q3	0.12 [0.68]	-0.28 [-1.49]	0.17 [0.66]	0.21 [0.72]	-0.23 [-0.88]	-0.04 [-0.11]	0.00 [0.02]	-0.23 [-1.29]
Q4	-0.13 [-0.86]	-0.57** [-2.44]	-0.03 [-0.14]	-0.16 [-0.52]	-0.35 [-1.56]	-0.83*** [-2.63]	-0.42** [-2.29]	-0.24 [-0.97]
Q5	0.03 [0.19]	-0.03 [-0.17]	0.40 [1.54]	-0.13 [-0.45]	-0.12 [-0.55]	-0.83** [-2.01]	-0.02 [-0.10]	0.25 [0.94]
Random	0.01 [0.03]	-0.36 [-1.42]	-0.00 [-0.01]	0.32 [0.89]	-0.43* [-1.70]	-0.25 [-0.73]	0.02 [0.12]	-0.09 [-0.35]
Panel B: Empirical Mode Decomposition								
Q1	0.32* [1.91]	0.06 [0.31]	0.38 [1.47]	0.30 [1.19]	0.14 [0.55]	0.23 [0.77]	0.30 [1.38]	0.02 [0.09]
Q2	0.05 [0.30]	-0.12 [-0.67]	0.55** [2.12]	0.05 [0.21]	-0.28 [-1.03]	-0.53** [-2.21]	0.47** [2.39]	-0.30 [-1.59]
Q3	0.14 [0.89]	-0.23 [-1.21]	0.00 [0.01]	-0.22 [-1.00]	-0.14 [-0.68]	-0.58* [-1.73]	-0.27 [-1.40]	-0.27 [-1.46]
Q4	-0.13 [-0.92]	-0.16 [-0.75]	-0.03 [-0.12]	-0.40 [-1.43]	-0.33 [-1.48]	-0.19 [-0.73]	-0.11 [-0.63]	-0.30* [-1.75]
Q5	-0.11 [-0.86]	-0.08 [-0.31]	0.06 [0.22]	0.20 [0.77]	0.07 [0.36]	-0.34 [-1.12]	0.13 [0.64]	0.01 [0.04]
Random	0.06 [0.35]	-0.02 [-0.14]	-0.06 [-0.27]	-0.01 [-0.03]	-0.24 [-1.06]	0.47 [1.52]	-0.02 [-0.08]	-0.29 [-1.61]

表中汇报了由高斯平滑方法和经验模态分解方法构造出的两种因子的收益率。



续图7：收盘价平滑处理后技术指标判断图

图中绘出了按照高斯平滑法分类的8种技术指标，并于经验模态分解得到的技术指标进行对比。HS、IHS、BTOP、BBOT、TTOP、TBOT、RTOP、RBOT 定义与公式20-23相同。NONE 表示图中不含有任何一种技术指标。图中分别为：收盘价、高斯平滑曲线、高斯平滑曲线极值点、经验模态分解曲线、经验模态分解曲线极值点。

表 21: 高斯平滑与经验模态分解因子相关性

	HS	IHS	BTOP	BBOT	TTOP	TBOT	RTOP	RBOT
Q1	0.12 [0.11]	0.18 [0.01]	0.07 [0.32]	0.12 [0.11]	0.23 [0.00]	0.27 [0.00]	0.23 [0.00]	0.37 [0.00]
	0.39 [0.00]	0.20 [0.01]	0.23 [0.00]	0.11 [0.13]	0.20 [0.01]	0.20 [0.00]	0.41 [0.00]	0.11 [0.14]
Q3	0.42 [0.00]	0.26 [0.00]	0.28 [0.00]	0.17 [0.02]	0.40 [0.00]	0.20 [0.00]	0.39 [0.00]	0.25 [0.00]
	0.23 [0.00]	0.22 [0.00]	0.12 [0.11]	0.38 [0.00]	0.13 [0.07]	0.45 [0.00]	0.42 [0.00]	0.15 [0.04]
Q5	0.41 [0.00]	0.34 [0.00]	0.26 [0.00]	0.10 [0.19]	0.14 [0.06]	0.32 [0.00]	0.35 [0.00]	0.05 [0.49]
	0.25 [0.00]	0.38 [0.00]	0.21 [0.00]	0.18 [0.01]	0.17 [0.02]	0.17 [0.02]	0.27 [0.00]	0.25 [0.00]
Random								

表中汇报了由高斯平滑方法和经验模态分解方法构造出的两种因子时间序列的相关性。方括号中为 p 值。

表 22: 经验模态分解图型技术指标因子投影

	H - L	$\alpha(\%)$	HS	IHS	BTOP	BBOT	TTOP	TBOT	RTOP	RBOT	$R^2$
Q1	0.48*** [3.38]	0.45*** [3.14]	0.07 [1.07]	-0.10 [-1.65]	-0.02 [-0.55]	0.02 [0.39]	0.02 [0.56]	0.01 [0.15]	0.05 [1.06]	-0.13* [-1.82]	0.05
Q2	0.51*** [3.42]	0.46*** [2.71]	0.04 [0.48]	0.11* [1.74]	-0.07 [-1.19]	0.03 [0.58]	-0.03 [-0.85]	-0.03 [-0.68]	0.15*** [3.23]	-0.02 [-0.40]	0.09
Q3	0.08 [0.51]	0.07 [0.50]	-0.08 [-1.12]	0.03 [0.52]	0.01 [0.26]	-0.06 [-1.40]	-0.03 [-0.57]	-0.03 [-1.15]	0.07 [1.37]	-0.03 [-0.37]	0.03
Q4	0.31** [2.56]	0.32*** [3.63]	-0.11 [-1.54]	-0.11** [-2.22]	0.04 [0.97]	-0.01 [-0.21]	0.08** [2.01]	-0.02 [-0.56]	-0.01 [-0.13]	0.07 [1.23]	0.08
Q5	0.59*** [4.61]	0.58*** [4.84]	0.11 [1.37]	-0.11*** [-3.33]	-0.04 [-0.94]	-0.01 [-0.15]	0.17*** [3.89]	-0.01 [-0.18]	0.01 [0.18]	-0.03 [-0.72]	0.13
Random	0.20 [1.45]	0.21 [1.49]	-0.18** [-2.41]	0.09 [1.25]	0.07 [1.57]	0.04 [1.25]	0.10* [1.75]	-0.01 [-0.25]	0.16*** [2.61]	-0.11* [-1.81]	0.15

表中汇报了各个股票池神经网络资产组合在图型技术指标因子上因子载荷。Panel A 为高斯平滑方法，Panel B 为经验模态分解方法。