基于Python及人工智能的集贸市场电子计价秤全周期数据处理方法

□廖才学 玉新华 雷海霞 张冉 磨昕玥

广西壮族自治区计量检测研究院

【摘 要】针对集贸市场电子计价秤强制检定工作数据量大后期数据处理繁琐等实际问题，本文提出基于Python及人工智能的全周期数据处理方法。该方法通过启发式算法，优化计量器具台账更新流程，解决列名异构与格式保留难题；结合OpenCV与OCR技术，实现表格图像的预处理与结构化转换；利用Python组件库，完成数据清洗与格式标准化，形成“数据整合→清洗处理→OCR识别→人工校对→可视化分析”的全流程解决方案。研究表明，该方法将台账列名匹配成功率提升至96.3%，手写数据录入效率提高6倍，全周期数据处理时间从传统15小时大幅缩短至2小时内，有效降低人工操作误差与成本，为集贸市场计量检定工作的信息化、智能化转型提供技术支撑。

【关键词】强制检定；电子计价秤；Python；人工智能；全周期数据处理

文献标识码：A 文章编号：1003-1870（2025）09-0042-08

Full-cycle Data Processing Method for Electronic Pricing Scales in Fair Trade Markets Based on Python and Artificial Intelligence

【Abstract】In response to the practical problems of large data volume and cumbersome data post-processing in the mandatory verification of electronic pricing scales in fair trade markets, this paper proposes a full-cycle data processing method based on Python and artificial intelligence. In this method, the heuristic algorithm is used to optimize the updating process of the measuring instrument ledger and solve the problems of column name heterogeneity and format retention; OpenCV and OCR technologies are combined to realize the preprocessing and structural transformation of table images; the Python component library is used to complete data cleaning and format standardization, forming a whole-process solution of "data integration - cleaning and processing - OCR - manual proofreading - visual analysis". Studies have shown that this method increases the success rate of ledger column name matching to 96.3%, increases the efficiency of handwritten data entry by 6 times, and significantly shortens the full-cycle data processing time from the traditional 15 hours to less than 2 hours, effectively reducing manual operation errors and costs, and providing technical support for the informatization and intelligentization transformation of metrological verification in the fair trade market.

【Keywords】mandatory verification; electronic pricing scale; Python； artificial intelligence; full-cycle data processing

引言

电子计价秤作为贸易结算的重要计量工具，广泛运用于集贸市场、超市、连锁零售店等领域。而强制检定作为维护市场秩序、防作弊与量值准确的

技术手段，守护了广大消费者的合法权益。

国内外研究学者主要从“电子计价秤检定技术”“AI在计量领域应用”“计量数据处理”三个方面开展电子计价秤相关研究：

（1）在电子计价秤检定技术方面，赵栋[1]系统分析了电子计价秤软硬件作弊方式及技术原理，提出了电子计价秤“快速检测三步法”，即外观检查→标准砝码测试→作弊码模拟测试的三步检测方法，为执法人员提供电子计价秤防作弊技术参考。王喜阳[2]基于Python编程语言，提出一种电子计价秤全自动检定装置设计方案，基于机器视觉系统实时采集与装置的自动化控制，为高效、准确且自动化地完成电子计价秤的全部检定项目提供全新的技术路径。何开宇[3]结合软、硬件两方面设计了一种防作弊电子计价秤，通过增设高安全IoT计量模组、RFID感应模块以实现作弊监测并保证计量准确。

（2）在AI在计量领域应用方面，顾方[4]基于机器视觉AI智能识别原理，采用ResNet深度学习分类算法、智能信息检索方法及计量器具数据库，研发了可实现器具准确识别的系统、提高信息输出效率的计量信息检索系统。朱东骏[5]设计了拆回计量器具的自动化分拣方法，利用AI识别技术和物流行业智能分拣技术，在传统的逻辑空间引入AI处理技术，有效改善现有分拣方法存在人工效率低及数据价值难挖掘等问题。

（3）在计量数据处理方面，孔令滨[6]基于Python及SQL语言设计开发了可用于实际工作的计量不合格电子计价秤查询系统，实现计量不合格电子计价秤的快捷查询。何琦[7]设计计量检定实验室仪表读数自动识别方案，结合机器学习领域的SSDA算法、BP算法及特征算法定位仪表读数区域，有效提升读数识别的效率和精度。Du Y等人[8]提出SVTR算法，引入Transformer结构以挖掘字符图像的关联信息，并通过引入全局及局部混乱块、提取笔画特征和字符间的相关性，提升字符识别准确率。

基于以上研究分析可知，近年来，国内外计量行业人工智能技术发展迅速，在数据处理、图像OCR识别领域[9–13]展现出极大的潜力。国内外研究学者对电子计价秤检定技术、AI在计量领域应用以及计量数据处理三个方面已有深入研究及成果，但未见结合人工智能及Python的全周期数据处理方法相关研究。

本文结合人工智能工具，基于Python编程语言对强制检定全周期数据处理的方法进行研究。基于数据各类特点，匹配相应的对策建议，为强制检定工作全周期数据处理提供系统性解决方法。

1　数据概况及特点

集贸市场电子计价秤数据量庞大，据不完全统计，2024年6月1日至2025年5月31日期间，广西计量院检定员对南宁市主要城区156家集贸市场电子计价秤开展强制检定工作。集贸市场强制检定后期数据具有如下特点：

（1）数据量大且信息复杂。集贸市场电子计价秤台件数较多，据统计，本周期共检定1.44万台电子计价秤，信息复杂体现在强制检定申报表中，表中字符包括有序号、姓名或摊位号、类别、计量器具名称、规格型号、测量范围、准确度等级、出厂编号、制造单位及备注栏，每一台电子计价秤需要填报10个参数，则1.44万台电子计价秤共计需要统计$1.44\times10=14.4$万个相关参数。

（2）申报表填报格式不统一。集贸市场多为公司对外租赁的经营模式，市场数量多，但主办方不一致，使得市场主办方对于强检申报表的填写格式不重视，在检定员检定返回进行后期数据处理时，会出现无效空格、重复行及乱码等情况，使得检定员出现差错的概率提升20%（出厂编号、摊位姓名填写错误）。

（3）需人为处理手写信息。强制检定时，可能出现集贸市场主办方台账信息不全的情况，集贸市场主办方台账信息登记准确性，可以用申报–检定一致率来表示，本周期集贸市场申报一致率最高为100%，最低为0，此时，该市场所有电子计价秤需要手工补填，检定完成后由检定员补登至电子表格，后期处理时会出现人工输入时编号误填等情况。据统计，申报一致率为0的该集贸市场，共有83台电子计价秤，后期数据录入耗时由0.5小时增加至3小时。

2　Python在集贸市场后期数据处理中应用场景

Python在数据处理领域的应用极为广泛，凭借其丰富的库生态和简洁的语法，成为数据科学、机器

学习、商业分析等领域的首选工具。针对集贸市场数据结构的特点，研究Python 的相应模块库在数据处理中的应用场景如下：

（1）针对集贸市场数据量大且复杂、数据表格格式不一致的情况，首先利用openpyxl 模块库对字体格式不统一、空格等格式问题进行处理，然后运用Pandas 对重复值、空值进行数据清洗。

（2）针对多个Excel 文件的数据合并、格式处理的情况，可先利用openpyxl 模块库保留原始Excel 的格式（如颜色、公式、图表），然后用xlwings 模块库与Excel 进行实时交互，实现图表的动态更新。

（3）针对需要将手写数据导入至结构化电子表格的情况，可首先结合OpenCV 进行表格检测，然后采用EasyOCR、PaddleOCR 或TesseractOCR 等模块库进行图像识别，即图像预处理→表格结构解析→单元格分割→OCR 识别→结构化输出的流程，可实现从图像表格到OCR 识别结果的转换，大幅减少手动录入工作量。

3　实例——集贸市场电子计价秤强制检定全周期数据处理方法

选取本周期已完成强制检定工作的一个集贸市场，对该市场强制检定全周期的数据处理方法进行研究。首先，对电子计价秤强制检定流程进行简述，强制检定流程图如图1 所示。

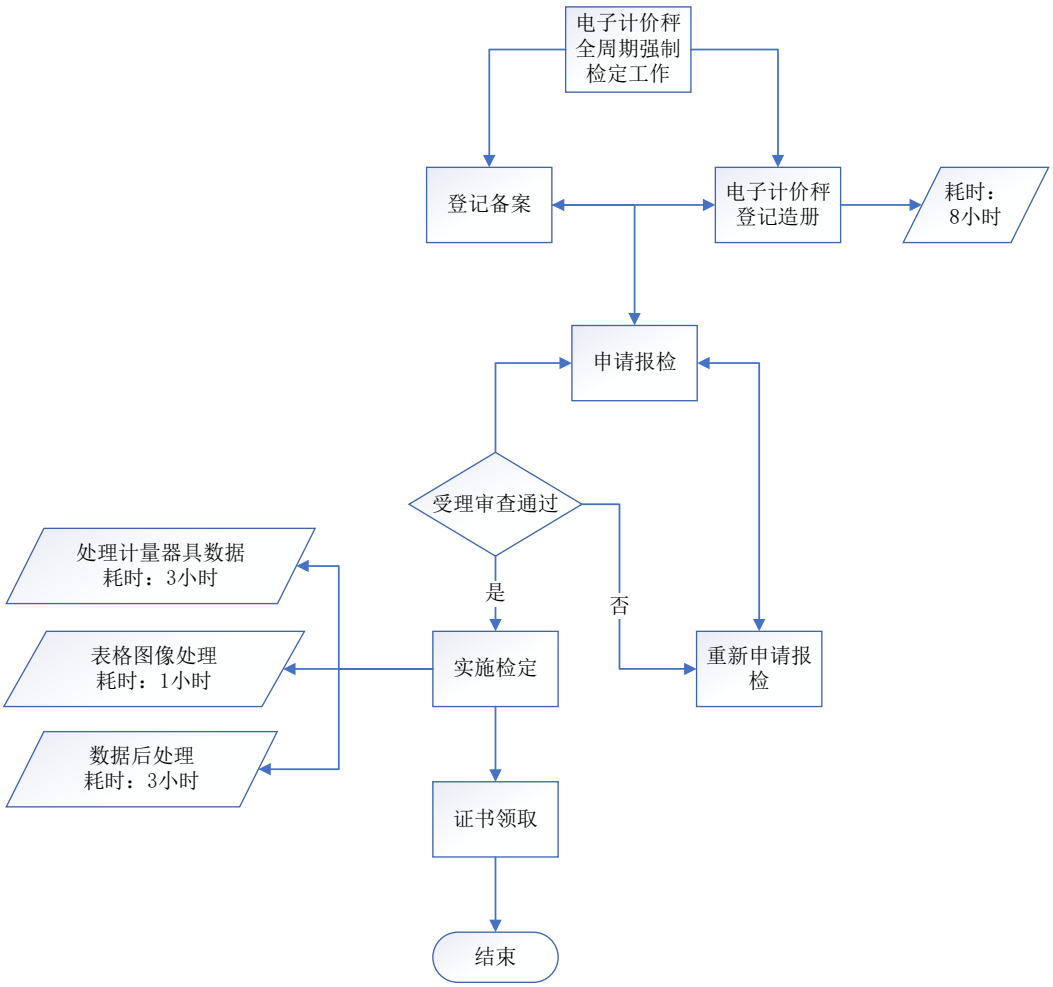

图1　强制检定工作流程图（仅展示数据处理耗时）

由图1 可知，按照传统人工处理集贸市场电子计价秤数据的方式耗时约为15 小时，其中集贸市场主办方登记造册耗时最多为8 小时，检定员处理计量器具数据、表格图像处理、数据后处理耗时分别为3 小

时、1 小时、3 小时。

3.1　基于启发式算法处理计量器具数据的方法

根据《集贸市场计量监督管理办法》[14]，集贸市场主办者应及时根据计量器具新增、减少及更换等情况对计量器具登记表进行更新、备案。集贸市场主办者更新台账时，产生人工处理数据效率低的问题。

针对集贸市场计量器具台账中列名异构问题（如同义不同形：“出厂编号”“设备ID”），结合人工智能领域启发式算法[15]进行研究，提出基于混合相似度计算的动态匹配规则，以Python 为基础形成一个计量器具台账信息表格更新方案：对短列名（如“出厂编号”）增加Levenshtein 权重，对长列名（如“计量器具名称”）增加词向量权重，然后不断动态调整权重，直至基于动态匹配规则的启发式匹配算法技术路径，实现了合并单元格感知的数据更新，即采用人工智能工具中的“模式识别”思想，处理新旧表格列名不一致的问题。

基于动态匹配规则的启发式匹配算法核心在于动态匹配过程，随后进行记录→更新→恢复的三步式格式保留机制：

（1）动态匹配规则

```
def dynamic_lev_score(s1, s2, len_thres=5, ratio_
thres=0.3):
    dist = distance(s1, s2)
    l1, l2 = len(s1), len(s2)
    max_l = max(l1, l2)// 动态计算Levenshtein 相似度，解决长短字符串评分偏差
    denominator = (l1 + l2) if max_l <= len_thres else max_l
    if abs(l1 – l2) / max_l > 0.5:
    denominator *= (1 + abs(l1 – l2) / max_l)
    base_score = 1 – dist / denominator// 短字符串用长度和归一化，长字符串用最大长度
    if dist / max_l > ratio_thres:
    penalty = (dist / max_l – ratio_thres) * 0.5
    base_score = max(0, base_score – penalty)
    return round(base_score, 4)// 距离占比过高时额外惩罚
```

（2）记录过程

```
old_cols = [cell.value if cell.value is not None else f”Unnamed_{i}”
for i, cell in enumerate(old_sheet[1], start=1)]
new_cols = new_df.columns.tolist()// 获取新旧表格的列名，处理空值
```

（3）更新过程

```
for col in new_df.columns:
if col in updated_df.columns:
updated_df[col] = new_df[col].values// 更新列数据
```

（4）恢复过程

```
old_wb = load_workbook(old_file)// 加载旧文件（保留格式）
old_sheet = old_wb.active// 获取活动工作表
for merged_range in original_merged_ranges:
old_sheet.merge_cells(str(merged_range))// 恢复单元格的合并状态
```

表1　动态调整效果验证（数据来源于算法运行结果）

列名对1	列名对2	原lev_score（固定计算）	动态lev_score（调整后）	差异原因
“kg”	“kgs”	1–1/3=0.67	0.8333	短字符串用 “长度和”（3+2=5）作为分母，弱化微小差异影响
“电子秤”	“电子计价秤”	1–2/4=0.5	0.4167	长度差异大（2 倍），增加分母惩罚，降低得分
“精度”	“准确度”	1–2/2=0	0.1667	短字符串编辑距离占比高，但动态惩罚后保留一定得分（更符合语义相似性）
“测量范围”	“测量范”	1–1/4=0.75	0.75	长字符串且编辑距离占比低，无额外惩罚，与原得分一致

如表1 动态调整效果所示，在列名异构场景下，当短列名匹配阈值设置为0.75 时，基于动态匹配规则的启发式算法准确率较传统方法提升17.8%，匹配成功率达到96.3%，高于传统规则匹配的78.5%，将数据录入的时间由8 小时降低至1 小时。

本方案核心在于制定动态匹配规则算法，然后提取原始表格的格式特征（包括合并单元格区域），在临时清除合并状态下更新数据，最后根据保存的特征恢复格式结构。该方案在Python 环境下通过openpyxl 库完成，有效解决了集贸市场计量器具台账更新中的列名异构动态匹配和复杂格式保留两大核心问题。

3.2　基于Python 的表格图像预处理优化方法

检定员赴集贸市场实施强制检定工作时，受限于集贸市场现场检定条件，检定员及集市主办方对电子计价秤信息登记时均为人工填写。

针对手写表格的人工识别效率低的问题，提出基于Python 的表格图像预处理优化方法，结合Python 技术提高效率的核心思路为：图像预处理→表格结构提取→OCR 识别→人工校对辅助的流程，可显著减少纯人工输入的工作量，如下图2 所示。

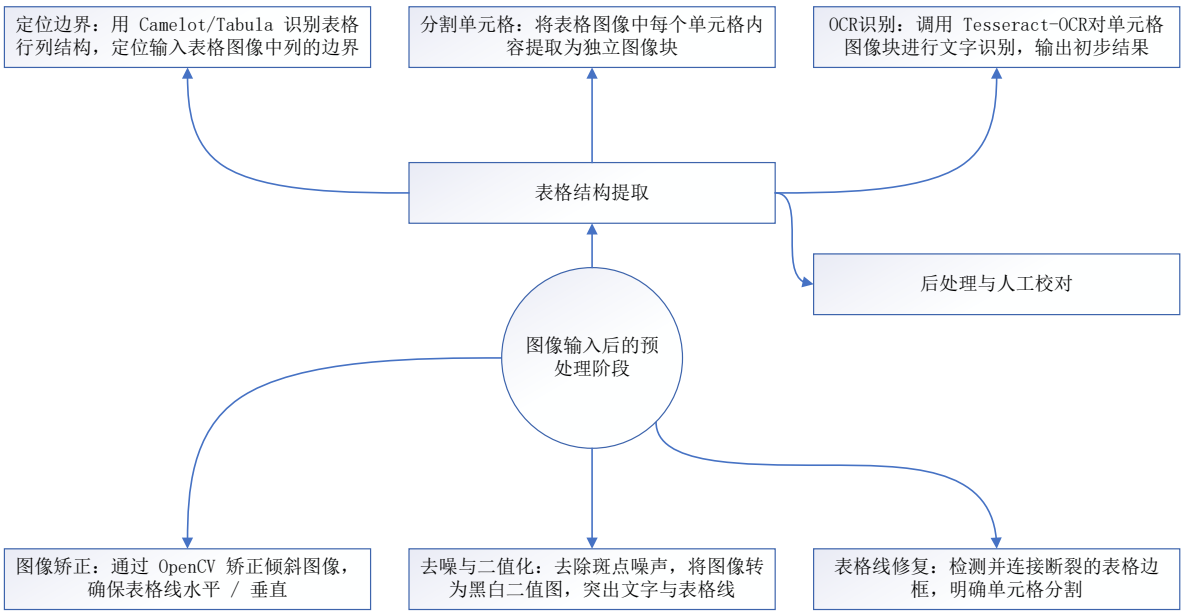

图2　表格图像预处理优化方法

图像输入预处理优化方法的关键点为：

（1）图像矫正与去噪

用cv2.warpAffine 矫正图像倾斜（解决扫描时的角度偏差）。在去噪环节，经过多次对比测试，采用“5×5 高斯核”的cv2.GaussianBlur 函数进行去噪处理，该参数设置能够在有效去除斑点噪声的同时，最大程度保留字符的边缘细节。

如表2 所示，在确定二值化最优阈值环节，对比分析了不同阈值下的字符边缘保留率。结合AI 工具，可知手写表格图像的阈值一般为100~160 之间。通过测试发现，当二值化阈值设为“127”时，字符边缘特征能够得到最佳的突出效果，此时文字为黑、背景为白，方便后续的识别处理。突出例如产品型号“BPS-30”、出厂编号“716R551459”等字符的边缘特征。

表2　阈值优化实验结果（数据来源于实际测试）

二值化阈值	平均边缘保留率	OCR 准确率
110	0.85	87.9%
120	0.89	88.6%
127	0.92	89.3%
130	0.90	88.9%

（2）基于Camelot/Tabula 的表格提取

用Camelot 库（camelot-py）自动检测表格区域，通过stream 模式提取行列结构（适合线条清晰的表格），或lattice 模式识别网格线（处理文档中可能的虚线、实线边框），输出为DataFrame 临时存储。

（3）OCR 识别与文字提取

由于数据表格均为手写文字，需引入深度学习模型（如CNN+LSTM），用Keras 或PyTorch 构建小型识别模型，基于公开手写数据集（如MNIST、IAM）进行微调。为进一步提高数字、字母和常见符号（如规格型号“BPS-30”、出厂编号“182495”）的识别精度。

研究基于MNIST 数据集扩展了200 张计价秤铭牌样本对CNN+LSTM 模型进行微调，经过测试，数字识别准确率从85% 提升至96%。同时，对OCR 识别性能进行了测试，测试了50 份数据表格，使用PaddleOCR 进行识别，其字符识别准确率为85.2%，经过人工校对辅助工具的修正后，准确率提升至93.2%，仅需10 分钟即可完成一个集贸市场数据表格的录入工作，较纯人工录入效率提升了6 倍。

（4）稳定性与鲁棒性测试

为验证算法的抗干扰能力，进行极端场景测试。测试场景包括“含大量乱码的申报表（乱码占比10%）”以及“倾斜45° 的手写表格图像”。测试结果表明，在乱码占比≤15% 时，经过数据清洗，数据的准确率仍保持70% 以上，说明该方法在一定程度上具备一定的抗干扰能力和鲁棒性。

（5）识别结果可视化校对及格式修正

用Pandas 将OCR 识别的结果生成临时Excel，结合Tkinter 或PyQt 制作简单GUI，将扫描图像与识别文本并排显示，制作开发一个人工校对辅助工具，如图3 所示。如图所示，OCR 识别结果为“BPS-3”，根据实际情况，可人工修改规格型号为电子计价秤铭牌的实际型号“BPS-30”，并保存修改，然后可点击“下一个”按钮，可自动定位至下一个需要修改的单元格，修改完成后可点击“导出最终”，生成最终的Excel 表格。

图3　人工校对辅助工具

3.3　基于Python 的数据后处理方法

检定员出具检定证书后，为及时将相关信息反馈至市场监管部门，需对集贸市场数据进行后期处理，即出具相应的检定情况报告，报告涉及数据分析、不合格原因归类的相关问题。

为提高效率，提出基于Python 的数据后处理分析方法，该方法基于“数据整合→清洗处理→深度分析→可视化输出”的流程实施，利用Pandas、NumPy 等库的批量处理能力，结合Matplotlib/Seaborn 实现可视化。以下是具体方法与实现细节：

（1）数据整合：多源表格标准化合并

集贸市场数据保存为互相独立的Excel 文件，本研究通过Python 实现自动化整合，利用os.listdir 遍历存储检定表格文件夹下的所有数据文件（如“青秀区-XX 市场.xlsx”“良庆区-B_XX 市场.xls”），通过文件命名规则（如“市场名_ 日期”）自动提取标

签。

（2）数据处理：清洗与标准化

整合后的数据集可能存在缺失值、异常值或格式错误（如格式混乱、数值单位不统一），需通过Python 批量处理，根据字段类型选择填充方式如铭牌信息缺失的，以“缺失”填充文本型字段final_df[“出厂编号”]=final_df[“出厂编号”].fillna(“缺失”)。

（3）数据分析：多维度挖掘与可视化

基于处理后的标准化数据，通过Python 进行趋势分析、对比分析和关联分析，并用可视化工具呈现结果，处理结果如图4 所示。

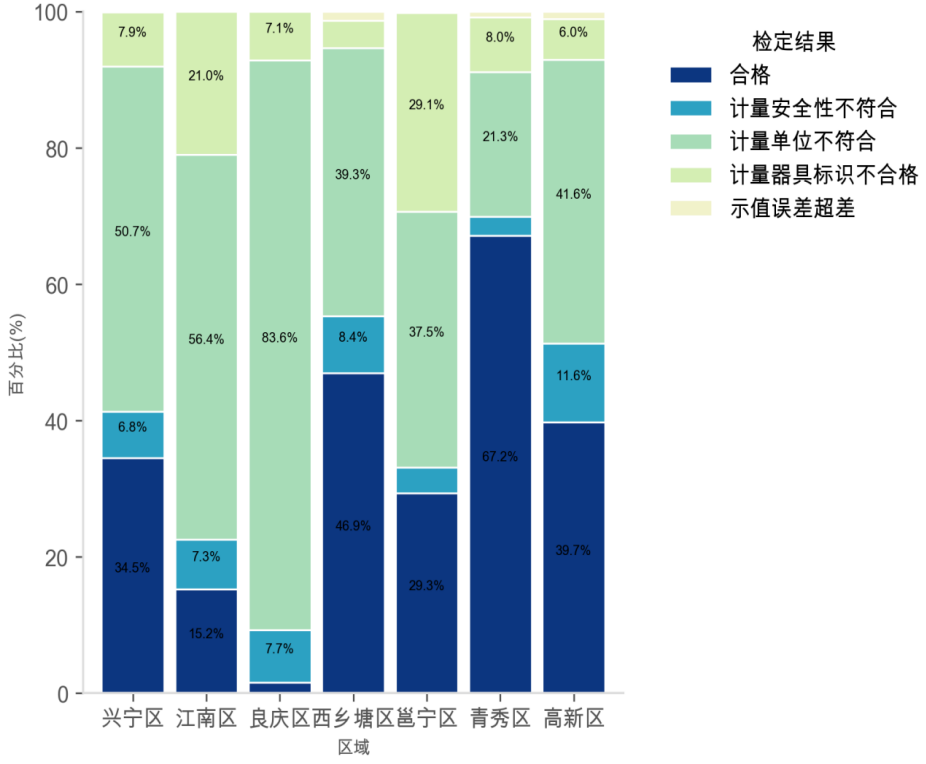

图4　可视化工具导出图（数据来源于南宁市7个城区156家集贸市场检定数据）

（4）结果输出与自动化报告

结合Jupyter Notebook 或ReportLab，将代码、可视化图表和分析结论整合为PDF 报告，支持定期更新数据后一键生成最新报告。经测定，检定员编制一份报告需要2~3 小时，而自动化工具仅需5 分钟，效率显著提升。

4　结语

本文针对集贸市场电子计价秤强制检定全周期数据处理的痛点，构件了基于Python 及人工智能技术的系统性解决方案，主要结论如下：

（1）提升处理效能。Python 生态中的Pandas、openpyxl 等组件库可高效应对数据量大、格式混乱的问题，通过批量清洗重复值、统一格式，解决了传统人工处理的低效问题。结合OpenCV 与OCR 技术的表格图像识别流程，实现了从手写表格到结构化数据的自动化转换，将OCR 识别准确率提升至93.2%，大幅减少人工录入误差。

（2）启发式算法的实用价值。基于动态匹配规则的启发式算法，通过“记录→更新→恢复”机制在保留表格格式（如合并单元格）的同时，实现列名异构场景下的精准匹配，匹配成功率较传统启发式算法提升17.8%，为市场主办方的台账维护提供了高效工具。

（3）全流程闭环处理体系。整合“数据整合→

清洗处理→OCR 识别→人工校对→可视化分析→自动化报告生成”的全周期数据处理流程，将集贸市场计量器具登记造册时间由8 小时缩短至1 小时，数据处理时间从3 小时缩短至0.5 小时，图像识别时间由1 小时缩短至10 分钟，报告生成时间从2~3 小时缩减至5 分钟。有效解决了申报表格式不统一、多源数据统计困难等问题，为集贸市场计量检定工作的信息化升级与智能化管理提供了实践支撑。

本研究对图像表格的识别局限于OCR，而表格图像往往字迹潦草，需较多人工校对，未来深化图像识别算法研究。本研究未涉及现场数据采集优化及深度数据分析功能，功能覆盖有一定局限性。研究学者可在现有分析基础上，结合AI 工具引入大数据预测模型，通过历史数据挖掘计量器具不合格趋势、作弊风险区域等规律，为市场监管部门提供主动预警支持，提升计量监管的前瞻性。

参考文献

[1] 赵栋, 孙兆军. 民生领域电子计价秤作弊方式分析与应对策略研究[J]. 衡器,2025,54(06):37−40.

[2] 王喜阳, 刘文佳, 肖福礼等. 基于电动推杆矩阵的电子计价秤自动检定装置设计[J]. 衡器,2025,54(06):28−33.

[3] 何开宇, 江浩, 李鹏飞. 基于云平台的电子秤防作弊控制模组设计[J]. 自动化与仪表,2025,40(04):153−156+161.

[4] 顾方, 胡良勇, 黄坚等. 基于机器视觉AI 智能识别的计量器具信息检索系统研究与应用[J]. 中国计量,2022,(10):47−48.

[5] 朱东骏, 汪如毅, 江婷等. 基于AI 识别技术的拆回计量器具自动化分拣方法[J]. 机械制造与自动化,2023,52(03):152−155.

[6] 孔令滨, 宋世栋, 李涛等. 计量不合格电子计价秤查询系统的设计与实现[J]. 衡器,2024,53(04):35−38.

[7] 何琦, 张鑫, 许晓平等. 基于SSD 的数显仪表读数检测系统的字符识别研究[J]. 数据通信, 2022, (05): 41−49.

[8] Du Y, Chen Z, Jia C, et al. Svtr: Scene text recognition with a single visual model[J]. arXiv preprint arXiv: 2205. 00159, 2022.

[9] Luo J , Li Z , Wang J ,et al. ChartOCR: Data Extraction from Charts Images via a Deep Hybrid Framework[J].IEEE, 2021.DOI:10.1109/WACV48630.2021.00196.

[10] 高良才, 李一博, 都林等. 表格识别技术研究进展[J]. 中国图象图形学报,2022,27(06):1898−1917.

[11] 盖鑫, 黄进, 王丹琳等. 基于结构理解的手绘草图表格识别[J]. 计算机辅助设计与图形学学报,2024,36(12):2051−2068.

[12] Zhelezniakov D , Zaytsev V , Radyvonenko O .Online Handwritten Mathematical Expression Recognition and Applications: A Survey[J].IEEE Access, 2021, PP(99):1−1.

[13] Lin W H, Sun Z, Ma C X, et al. TSRFormer: table structure recognition with transformer [C] // Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM Press, 2022: 6473 − 6482.

[14] 《集贸市场计量监督管理办法》解读[J]. 中国计量,2025,(01):14−16.

[15] 武优西, 吴信东, 江贺等. 一种求解MPMGOOC问题的启发式算法[J]. 计算机学报,2011,34(08):1452−1462.

作者简介

廖才学，广西壮族自治区计量检测研究院衡器所副所长，中级工程师，长期从事衡器领域强制检定工作，负责统筹南宁市各集贸市场电子计价秤检定工作。研究方向：计量检测技术。