AI 辅助的档案开放审核风险评估模型

赵　越

河南港投航程城建集团有限公司　河南　郑州　451100

摘　要：档案开放审核是保障档案资源共享与信息安全的核心环节，人工审核面临效率低、风险识别不精准、主观偏差大等问题。本文基于相关法规与 AI 技术原理，梳理核心风险类型，分析 AI 辅助评估的可行性与优势，构建风险评估模型，明确核心架构、指标体系与实现路径，结合实践案例验证有效性并提出优化策略，为提升档案开放审核风险防控能力、实现智能化转型提供理论与实践指引，助力平衡开放共享与信息安全的关系。

关键词：AI 技术；档案开放审核；风险评估；风险指标；模型构建

引言：随着档案数字化转型深入与档案开放共享需求提升，档案开放审核工作的重要性日益凸显。根据《国家档案馆档案开放办法》，档案开放需严格审核，既要保障满 25 年档案的及时开放，也要防范涉密、隐私泄露等风险。AI 技术的飞速发展，尤其是自然语言处理、深度学习等技术的突破，为档案开放审核风险评估提供了新路径。本文聚焦 AI 辅助的档案开放审核风险评估模型，梳理风险类型、构建评估体系、明确实现路径，破解人工审核困境，提升风险防控精准度与效率，推动档案开放审核工作高质量发展。

1 档案开放审核的核心风险类型

结合档案开放审核实践与相关法规要求，档案开放审核的核心风险主要分为四类，贯穿审核全流程。一是涉密风险，即档案中包含的国家秘密、工作秘密未被识别，开放后危及国家安全与社会稳定，这是审核工作的首要风险，涵盖保密期限未届满、解密条件未达成的涉密内容。二是隐私泄露风险，档案中涉及的个人信息、商业秘密等被违规开放，损害第三方合法权益，如个人身份信息、企业核心技术资料等。三是合规风险，审核流程未遵循相关法规标准，如擅自开放归属权限不属于本馆的档案、未按规定延期开放应限制利用的档案，面临监管处罚。四是操作风险，审核过程中因人员失误、流程不规范、技术故障等，导致风险识别不及时、审核结果偏差，如破损档案未修复即开放造成的二次损害[1]。

2 AI核心技术赋能档案开放审核风险评估的路径解析

2.1 文本智能处理与语义理解技术

档案开放审核的首要任务是精准解读文本内容，文本智能处理技术为此提供基础支撑。光学字符识别技术将纸质档案图像转化为可编辑文本，针对手写体、繁体字、印章遮挡等复杂情况，需采用专门训练的 OCR 引擎并通过多模型投票机制提升识别准确率。完成文本化转换后，中文分词工具对档案文本进行语义单元切分，结合档案领域自定义词典解决专业术语的分词歧义。在语义理解层面，预训练语言模型通过双向编码机制捕捉词语的上下文依赖关系，能够区分不同语境下的风险差异。相比单纯依赖敏感词匹配的传统方式，语义理解技术可有效降低一词多义或同义替换导致的误判率，使模型能够识别“以代号称谓的涉密项目”“隐晦表述的敏感事件”等深层次风险，显著提升审核精准度。

2.2 风险模式学习与智能分类技术

风险评估的核心任务是判断档案的风险类别与等级，依赖机器学习与深度学习构建风险模式识别能力。监督学习算法以已标注的档案样本为训练素材，自动学习风险特征与判定规则的映射关系。支持向量机适用于小样本场景下的风险二分类任务；决策树的可视化输出便于审核人员理解判断依据。对于缺乏标注的历史档案，无监督聚类算法可根据文本特征将档案自动归入不同风险簇群，辅助发现潜在风险模式。深度神经网络拓展了特征提取的深度与广度，卷积神经网络擅长捕捉局部语义模式，Transformer 架构通过自注意力机制建立全局语义关联，能够识别跨段落的风险线索。多模态扩展技术将图像识别与语音转写能力引入评估体系，同步分析扫描件中的敏感图标和会议录音中的禁语表述[2]。

2.3 模型可信增强与动态适配技术

为确保 AI 风险评估结果被信任和采纳，需要部署模型可信增强与动态适配技术。可解释性技术解决深度学习模型“黑箱”操作的透明度问题，当模型判定

某份档案为高风险时，同步输出决策依据报告，利用注意力权重可视化技术标注对判断贡献最大的敏感词句，并生成各风险指标的贡献度排序，提升人工复核的针对性和效率。隐私计算技术为跨机构协同审核提供安全保障，联邦学习框架允许各档案馆在不共享原始档案数据的前提下联合训练模型，既扩大训练样本规模，又避免敏感数据外泄。法规动态适配技术建立法律条文与算法规则的转换通道，当法规修订时，规则引擎自动解析新增条款，将其转化为模型可执行的判定规则，同步更新风险识别阈值和敏感词库，确保模型始终运行在合规边界之内。

3 AI 辅助的档案开放审核风险评估模型构建

3.1 模型构建原则与目标

模型构建遵循“合规性、精准性、可扩展性、实用性”四大原则，严格贴合《国家档案馆档案开放办法》等法规要求，确保风险评估符合法律规范；聚焦风险识别的精准性，降低误判、漏判率；采用模块化设计，支持风险指标、算法模型的灵活调整，适配不同类型档案与审核需求；兼顾技术可行性与业务适配性，便于档案管理机构落地应用。模型核心目标是：通过 AI 技术自动识别档案开放审核中的各类风险，量化风险等级，生成风险评估报告，为人工审核提供辅助，提升风险防控效率与精准度，平衡档案开放共享与信息安全，降低审核成本与合规风险。

3.2 风险评估指标体系设计

结合档案开放审核风险类型与 AI 技术特性，构建多层次风险评估指标体系，分为一级指标、二级指标、三级指标三个层级，共 16 个分析维度，全面覆盖各类风险。一级指标包括 4 类核心风险：涉密风险、隐私泄露风险、合规风险、操作风险。二级指标是对一级指标的细化，如涉密风险分为国家秘密、工作秘密、军事秘密 3 个二级指标；隐私泄露风险分为个人信息、商业秘密 2 个二级指标。三级指标是具体的风险识别点，如国家秘密对应保密期限、涉密载体、涉密内容 3 个三级指标，个人信息对应身份信息、联系方式、隐私事件 3 个三级指标。同时，采用德尔菲法确定各指标权重，通过专家打分与算法优化，确保指标体系的科学性与合理性。

3.3 模型总体架构设计

AI 辅助的档案开放审核风险评估模型采用分层解耦架构，整体分为四层，从下至上依次为数据资源层、技术支撑层、模型核心层、应用服务层，各层协同运行，确保模型高效稳定。数据资源层负责存储档案数据（多模态）、风险指标数据、法规规则数据、训练样本数据等，采用分布式存储架构，支持海量数据的高效访问与管理，同时构建档案知识库，包含敏感词库、开放词库等，为风险评估提供支撑。技术支撑层封装自然语言处理、机器学习、隐私计算等核心技术，以 API 接口形式为模型核心层提供技术服务。模型核心层是核心模块，包括风险指标提取、风险等级判定、模型优化三个子模块，实现风险识别与评估。应用服务层为用户提供交互界面，实现档案上传、风险评估、报告查看、人工复核等功能。

3.4 模型实现流程

模型实现遵循“数据输入—预处理—风险识别—等级判定—报告生成—人工复核”的闭环流程。第一步，数据输入，收集各类待审核档案数据，包括文本、图像、音视频等，完成数据汇总与整理。第二步，数据预处理，对档案数据进行格式标准化、噪声去除、文本转换等操作，确保数据质量，为风险评估奠定基础。第三步，风险识别，通过 NLP 技术提取档案中的风险特征与关键词，结合风险指标体系，识别各类风险点。第四步，等级判定，利用训练好的机器学习模型，结合指标权重，量化风险等级（分为低、中、高三个等级），高风险档案直接判定为不可开放，中低风险档案进入人工复核[3]。第五步，报告生成，自动生成风险评估报告，明确风险类型、风险点、风险等级及防控建议。第六步，人工复核，管理员对模型评估结果进行校验、修正，形成最终审核意见，确保审核结果准确合规。

4 模型训练、优化与实践验证

4.1 模型训练与参数优化

模型训练以真实档案数据为基础，选取不同类型、不同领域的档案样本 2.1 万份，由专业档案审核人员进行人工标注，明确风险类型、风险等级与风险点，构建训练数据集与测试数据集。选用 BERT+Dense、TextCNN 等算法模型，依托 TensorFlow 框架进行模型训练，调整学习率、迭代次数、窗口大小等参数，采用交叉验证方法提升模型泛化能力。针对算法偏见问题，采用对抗性训练，补充各类别档案样本，确保模型对不同群体、不同类型档案的审核公正性。通过多轮训练与优化，模型风险识别准确率提升至 91%以上，与人工审核一致率达到较高水平，满足实际审核需求。

4.2 实践验证

选取某省级档案馆作为实践试点，将构建的模型应用于档案开放审核工作，选取 3000 份待审核档案进行实

践验证，涵盖文书档案、图像档案、音视频档案等多种类型。实践结果显示，模型可快速完成批量档案的风险评估，处理效率较人工审核提升 3 倍以上；风险识别准确率达 91.5%，其中涉密风险、隐私泄露风险识别准确率分别为 93.2%、92.7%，显著高于人工审核的 78%；模型生成的风险评估报告为人工复核提供了明确指引，大幅缩短了复核时间，降低了人工工作量与主观偏差。同时，模型可适配不同类型档案的审核需求，与现有档案管理系统平滑对接，验证了模型的实用性与可行性。

4.3 模型现存问题与优化方向

实践验证中发现模型存在三点不足：一是多模态档案（如手写档案、模糊图像）的风险识别准确率有待提升，部分手写文本 OCR 识别误差导致风险漏判；二是模型对新型风险（如新型涉密内容、新型隐私信息）的适配性不足，识别滞后；三是模型可解释性仍需强化，部分复杂风险的判定逻辑难以直观呈现。针对以上问题，优化方向主要包括：优化 OCR 识别算法，提升手写文本、模糊图像的识别准确率；建立风险特征动态更新机制，及时纳入新型风险点与法规条款；完善可解释性模块，采用 LIME 技术生成更详细的决策依据，提升模型透明度。

5 模型应用保障措施

5.1 法规与制度保障

完善档案开放审核相关法规与管理制度，明确 AI 辅助风险评估模型的应用规范，界定模型应用的权责边界，确保模型应用合规有序。结合《档案法实施条例》等法规，制定 AI 辅助审核的操作流程，明确人工复核的范围与标准，对高风险档案、复杂档案实行强制人工复核，避免模型误判导致的安全风险。同时，建立模型应用监督机制，定期对模型应用效果进行评估，及时发现并整改问题，确保模型应用符合法规要求与审核工作需求。

5.2 技术与人才保障

加强技术支撑，搭建稳定的技术平台，配备国产化硬件设备与算力资源，保障模型高效运行；建立模型定期更新与维护机制，及时优化算法参数、更新风险词库与法规规则，提升模型适配性。同时，加强人才培养，培育兼具档案管理与 AI 技术知识的复合型人才，负责模型的训练、维护、优化与应用指导；开展档案审核人员 AI 技术培训，提升其对模型评估结果的解读与复核能力，确保模型与人工审核高效协同[4]。

5.3 数据与安全保障

加强档案数据管理，建立档案数据采集、存储、处理的标准化流程，确保数据的完整性、准确性与安全性；采用加密技术对敏感档案数据进行保护，防范数据泄露风险。同时，建立模型安全保障体系，加强对模型训练数据、算法参数的安全防护，防止模型被篡改、滥用；定期开展安全检测与风险评估，及时排查技术安全隐患，确保模型应用过程中的信息安全，兼顾档案开放共享与数据安全保护。

结束语

AI 辅助的档案开放审核风险评估模型，有效破解了传统人工审核效率低下、风险识别不精准的困境，为档案开放审核工作提供了高效、科学的解决方案。本文基于档案开放审核法规与 AI 技术原理，梳理了核心风险类型，分析了 AI 技术的应用机理，构建了包含四层架构、多维度指标体系的风险评估模型，通过实践验证了模型的实用性与可行性，并提出了优化方向与保障措施。未来，需结合档案管理实践与 AI 技术发展，持续优化模型性能，完善应用保障体系，推动 AI 技术与档案开放审核深度融合，提升风险防控能力，实现档案开放共享与信息安全的协同发展。

参考文献

[1] 李淑军, 刘彦麟, 曾英. 基于开源与通用大模型整合的多模态可解释 AI 辅助档案开放审核技术研究 [J]. 云南档案,2025(3):56-59.

[2] 周友泉, 连波, 曹军." 浙里数字档案" 重大应用场景实践——" 档案 AI 辅助开放审核" 组件的性能与应用 [J]. 浙江档案,2022(11):22-24.

[3] 郝召红. 基层档案馆人工智能辅助开放审核的困境与突破路径探析 [J]. 山东档案,2026(1):13-14.

[4] 王萃. 档案管理中 AI 大模型驱动的知识图谱构建与应用研究 [J]. 办公室业务,2025(20):85-87.