基于可信 AI 的船舶健康监测模型可解释性分析

吴永华
浙江交通职业技术学院 海运学院
DOI:10.32629/pe.v4i1.19043

[摘　要] 船舶健康监测作为航运安全保障、运维成本降低的核心环节,可信AI技术的融入为监测精度提升给予重要支撑,然而模型“黑箱”状况严重限制了其在船舶领域的规模化落地。本文从可信AI与船舶健康监测的核心关联开始,剖析可解释性对船舶健康监测模型的关键意义,解析当前模型可解释性欠缺的现实难题,探究适配船舶场景的可解释性达成路径,为可信AI在船舶健康监测领域的实用化发展提供参考。
[关键词] 可信AI；船舶健康监测；模型可解释性；航运安全；运维优化
中图分类号：F560.1　文献标识码：A

Explainability analysis of ship health monitoring model based on trusted AI

Yonghua Wu
Marine department ,Zhejiang Institute of Communications

[Abstract] As a core component of shipping safety assurance and operational and maintenance cost reduction, ship health monitoring relies heavily on the integration of trusted AI technology to enhance monitoring accuracy. However, the "black box" nature of models severely limits its large-scale implementation in the shipping industry. This paper starts from the core correlation between trusted AI and ship health monitoring, analyzes the crucial significance of interpretability for ship health monitoring models, explores the practical challenges of current models' lack of interpretability, and investigates the path to achieving interpretability suitable for ship scenarios. It provides a reference for the practical development of trusted AI in the field of ship health monitoring.
[Key words] Trustworthy AI; Ship Health Monitoring; Model Interpretability; Shipping Safety; Operation and Maintenance Optimization

引言

伴随航运业向智能化、绿色化转变,船舶健康监测已从传统的人工巡查检验、定期维护保养,逐步升级成为基于AI技术的实时监测、预警与诊断模式。可信AI凭借其可靠性、安全性与可控性特征,能够精确捕获船舶动力装置、船体结构、航行系统等关键部件的运行异常信号,向船员提供决策支撑,有效避免设备故障引发的航行风险,降低停运损失。但当下主流的AI监测模型,特别是深度学习模型,大多存在“黑箱”属性,其预测与诊断结果缺少清晰的逻辑支撑,船员难以判断结果的合理性与可信度,一旦模型出现错误判断,可能造成运维决策失误,反而埋下安全隐患[1]。所以,增强可信AI船舶健康监测模型的可解释性,打破“黑箱”障碍,让模型结果“能够阐述清楚、可以看得明白”,成为推动AI技术在船舶健康监测领域深入应用的关键所在。本文立足于船舶运维的实际场景,聚焦模型可解释性的核心问题,探寻兼具专业性与实用性的可解释性实现方案,为提升船舶健康监测的智能化水平与可信性提供思路。

1 可信AI与船舶健康监测的核心关联

1.1可信AI的核心特质匹配船舶监测需求

可信AI并非单一技术概念,而是包含可靠性、安全性、可解释性、公平性等多维度的技术体系,其核心特质与船舶健康监测的场景需求高度吻合。船舶航行环境复杂多样且不断变化,海洋风浪、极端气候、航行负荷波动等因素都会对设备运行状态产生影响,对监测模型的可靠性提出极高要求——模型需在复杂干扰之下稳定捕获异常信号,防止漏判、误判情况发生。同时,船舶设备价值高昂、故障后果严重,监测模型的安全性至关重要,需保证模型运行过程可控制、结果可追溯,不会因模型漏洞引发二次风险。而可解释性作为可信AI的核心构成部分,是连接模型预测结果与实际运维决策的纽带,能够让船员理解模型判断的依据,进而安心、合理地运用模型结果开展运维工作,这亦是可信AI有别于传统AI技术,更匹配船舶健康监测场景的关键优势[2]。

文章类型：论文|刊号（ISSN）：2972-4112(P) / 2972-4120(O)

1.2可解释性是船舶健康监测模型实用化的必要前提

船舶健康监测最终目标是服务于实际运维决策,而非单纯追求预测精度。船员作为运维决策的主体,将专业经验与模型结果相结合,才能达到最优的故障处置效果。若模型仅仅输出“设备异常”“存在故障风险”等模糊结果,却不能阐释异常位置所在、故障原因构成、风险等级的判断依据等事项,船员非但难以迅速定位问题、采取具有针对性措施,还可能对模型结果产生质疑,甚至放弃使用AI工具,回归传统巡检模式。与之相对,具备良好可解释性的模型,能够清晰地展现“哪些运行参数呈现异常状态、异常参数以何种方式影响设备状态、故障概率通过何种途径推算得出”等关键信息,与船员的专业经验形成互补格局,协助船员迅速锁定故障节点位置、制定维修方案规划,大幅提升运维效率。除此之外,于船舶故障溯源流程、责任认定等场景环境当中,可解释性模型可提供完整的决策链路追溯内容,明晰异常信号捕捉动作、分析过程、判断结论的全部流程,为后续故障复盘工作、运维优化事项提供可靠依据,此亦为模型实用化的重要保障[3]。

2 船舶健康监测AI模型可解释性不足的现实困境

2.1模型复杂度与可解释性的固有矛盾

为提升监测精度水准,当前船舶健康监测领域多采用深度学习技术、集成学习方法等复杂AI模型类型,该类模型借助多层神经网络架构、多算法融合方式实现对复杂运行信号的精准拟合操作,但其结构的复杂性亦引致“黑箱”问题。譬如,基于卷积神经网络(CNN)的船舶发动机振动信号分析模型,能够精准识别微小振动异常,但模型内部的卷积运算、池化运算过程难以用工程语言解释,无法明确哪些频率成分、特征向量是判断故障的关键；基于梯度提升树(XGBoost)的船体结构应力监测模型,预测精度较高,但多个弱分类器的融合逻辑复杂,难以追溯单个特征对预测结果的影响。这种“精度居于优先地位、解释处于滞后状态”之现实状况,致使复杂模型虽能够满足监测精度方面之需求,却难以适配船舶运维的实际决策场景,形成“模型具备可用性但却不敢投入使用”的困境[4]。

2.2船舶场景特性加剧可解释性实现难度

船舶健康监测场景的特殊性,进一步加大了模型可解释性的实现难度。一方面,船舶设备运行参数具有维度多、关联性强的特点,一艘船舶的动力系统就包含转速、油温、油压、排气温度等数十项监测参数,这些参数相互影响,处于动态变化,模型在分析过程中难以割裂单个参数的作用,致使解释逻辑混乱,难以梳理。另一方面,船舶运行环境存在强干扰、非稳态特性,风浪冲击、负载波动等外部因素会使监测参数出现瞬时异常,模型需区分“真实故障信号”与“环境干扰信号”,但此类划分的内在机理难以直观展现,船舶工作人员难以明确模型过滤干扰信号的依据。除此之外,不同类别的船舶、不同服役时长的设备运行特征存在较大差别,模型需要具备一定的泛化性能,而泛化性能的增强往往要以舍弃部分可解释性为代价,导致模型在不同船舶场景中的解释逻辑不一致,进一步削弱了船舶工作人员对模型的信赖程度[5]。

2.3可解释性技术与船舶运维需求脱节

当前主流的人工智能模型可解释性技术,大多来源于计算机、人工智能领域,其解释形式与船舶运维的实际需求存在脱节情况。比如,部分可解释性方法通过输出特征重要性排序来解释模型结果,但仅仅告知“某一参数对结果影响最为显著”,却无法说明该参数异常达到何种程度会引发故障、与其他参数的协同影响状况如何；部分办法通过可视化模型内部构造来呈现解释逻辑,但可视化结果过于抽象,需要具备专业的人工智能知识才能够理解,而船舶工作人员大多具备船舶工程、航海技术背景,难以解读复杂的模型构造与特征映射联系。另外,现有的可解释性技术大多聚焦于“事后解释”,也就是在模型输出结果之后再追溯解释依据,而船舶健康监测更加需要“事前预警+事中分析+事后追溯”的全流程可解释,现有的技术难以满足这一全链路需求,导致解释结果对运维决策的支撑作用较为有限。

3 基于可信AI的船舶健康监测模型可解释性实现路径

3.1采用“轻量化模型+可解释模块”融合设计

针对模型复杂度与可解释性之间的矛盾,可采用“以轻量化模型为核心、以可解释模块为补充”的融合设计思路,在保障监测精准度的同时,提升模型的可解释性。一方面,结合船舶监测场景的核心需求,简化模型构造,优先选用逻辑清晰、易于解释的基础模型,如决策树、逻辑回归等,针对关键设备的核心监测参数构建基础模型,确保模型核心逻辑能够直观理解。另一方面,针对复杂场景的精准度需求,在基础模型当中嵌入轻量化可解释模块,而非采用复杂的深度学习模型。例如,在船舶发动机故障监测工作中,以决策树模型为核心,捕捉转速、油温等关键参数与故障类型的对应联系,同时嵌入特征交互解释模块,明确不同参数协同作用对故障判断的影响,如“油温超过85℃且转速波动大于5%时,发动机轴承故障风险提升60%”,用直白的工程语言呈现解释逻辑,适配船舶工作人员的理解能力。这种设计方式既避免了复杂模型的“黑箱”问题,又能够通过可解释模块补充模型的分析维度,兼顾精准度与可解释性。

3.2构造适配船舶场景的可解释性评价架构

可解释性地达成需联结船舶场景的现实需求,搭建针对性的评价架构,而非照抄通用领域的评价规范。评价架构应聚焦“船员可领会、决策可支撑”两大核心宗旨,从解释明晰度、逻辑连贯性、运维实效性三个维度施行评价。解释明晰度层面,规定解释成果运用船舶工程领域的专业术语,规避抽象的AI理念,能够明确异常参量、故障缘由、风险层级的对应联系；在逻辑连贯性层面,规定模型在不同运行工况、不同船舶场景里,对同类故障的解释逻辑维持一致,规避出现“同类故障异样解释”的状况；运维实效性层面,规定解释成果能够直接引导运维操作,明确“需要查验哪些组件、采取何种应急办法、维修优先次序怎样”,为船员供应具体的决策依据。同时,可引入船员参与评价,借助问卷调查、实操测验等途径,搜集船员对解释成果的理

解程度、信任度及应用成效反馈,持续优化可解释性设计,保证解释成果贴合实际运维需求。

3.3嵌入领域知识增强全流程可解释性

船舶健康监测具备鲜明的领域属性,嵌入船舶工程领域知识,是提升模型可解释性的关键途径。在模型搭建阶段,依据船员的运维经验、设备说明书、故障案例等领域知识,筛选核心监测参量、界定参量阈值范畴、梳理故障类型与特征参量的对应关联,将领域知识嵌入模型的特征选取、结构设计环节,使模型的分析逻辑与船舶设备的运行原理、故障机制相契合,从根源提升模型的可解释性。在模型运行阶段,结合实时航行环境、设备运行工况等场景信息,对模型成果进行补充阐释,例如,当模型检测到油压异常时,同步表明"当前海况等级为5级,可能存在风浪干扰引发的瞬时异常,建议持续监测30分钟后再判别",协助船员辨别故障信号与环境干扰。在模型成果追溯阶段,依据领域知识搭建故障溯源链路,明确异常信号从产生到被模型捕获、分析、判定的全流程,结合历史故障案例,说明同类故障的处置经验,使解释成果更具说服力与参考价值。

4 结论

可信AI为船舶健康监测供应了高效、精准的技术支撑,而可解释性作为打破模型"黑箱"、构建船员信任的核心环节,直接决定AI技术在船舶领域的实用化水准。当前船舶健康监测AI模型面临复杂度与可解释性矛盾、场景属性加剧实现难度、技术与需求脱钩等困境,需通过"轻量化模型+可解释模块"融合设计、搭建场景化可解释性评价架构、嵌入领域知识增强全流程可解释性等路径,提升模型的解释能力。在实际应用中,需结合动力系统、船体结构等不同监测场景的需求,优化解释逻辑与呈现方式,同时通过持续迭代优化与船员培训,让模型结果真正服务于运维决策。

受浙江省教育厅一般科研项目资助(项目编号；Y202559357)；船舶动力设备健康状态监测评估系统设计的关键技术研究。

[参考文献]

[1]杨泽文,郭力,袁昱超.大模型驱动的船舶结构健康监测系统关键技术研究综述及展望[J].中国舰船研究,2025,20(6):3-18.

[2]赵永超,曲志林,杨军.船舶柴油机进排气系统健康监测工艺研究[J].现代制造技术与装备,2025,61(08):111-114.

[3]周兵,袁伟.工程在线健康监测系统及其在船闸系统中的应用前景[J].水电与抽水蓄能,2025,11(04):108-111+120.

[4]黄峰.无损检测技术在船舶结构健康监测中的应用[J].船舶物资与市场,2025,33(06):124-126.

[5]张威.船舶撞击力作用下码头工程多尺度健康监测与现代技术融合的设想[J].中国水运,2025,(11):27-29.

作者简介：

吴永华(1981--),男,汉族,浙江天台人,研究生,讲师,从事的研究方向：从事船舶工程技术、轮机工程技术等领域的研究。