ICS 35.240.99
CCS L 67

团 体 标 准

T/CI 557—2024

人工智能驱动的校园欺凌防控管理技术规范

Technical specifications for prevention and control of bullying in schools driven by artificial intelligence

2024－11－01 发布　　2024－11－01 实施

中国国际科技促进会　发 布
中国标准出版社　出 版

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由南方科技大学提出。

本文件由中国国际科技促进会归口。

本文件起草单位：南方科技大学、吉林大学、深圳市铠硕达科技有限公司、深圳市骏嘉科技发展有限公司、北京大学、北京大学深圳研究生院、北京师范大学珠海校区、北京林业大学、深圳树米网络科技有限公司、北京高科中创科学技术中心、深圳浑沌数字化实验室科技有限公司、中国第一汽车股份有限公司研发总院、杭州海康威视数字技术股份有限公司、上海巨众网络科技有限公司、北京旺凡科技有限公司、天津津湖数据有限公司、吉林省卡思特科技有限公司、吉林省中云数讯科技股份有限公司。

本文件主要起草人：宋轩、张悦、谢洪彬、何忠荣、李博翱、张婧雯、刘仝、范子沛、王田、彭玉佳、毕波、陈欣、吴瑞彬、温鹏、张浩然、袁飞、张凌宇、陈瑶、丁少杰、冯树军、樊玉静、范明耀、周时莹、李长龙、孔祥明、高仕宁、张文杰、宋小龙、王德周。

人工智能驱动的校园欺凌防控管理技术规范

1 范围

本文件规定了基于人工智能的校园欺凌防控管理系统架构、人工智能算法构建、应用场景功能构建等要求。

本文件适用于以人工智能技术为驱动的校园防霸凌系统设计、开发和建设。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 5271.34 信息技术 词汇 第34部分:人工智能 神经网络

GB/T 35119 产品生命周期数据管理规范

GB/T 35273 信息安全技术 个人信息安全规范

GB/T 41867 信息技术 人工智能 术语

3 术语和定义

GB/T 5271.34、GB/T 35119、GB/T 41867界定的术语和定义适用于本文件。

4 符号和缩略语

下列符号和缩略语适用于本文件。

TP：真正例（正确预测的欺凌行为）

FN：假负例（错误预测的非欺凌行为）

n：样本数量

y_i：预测值

x_i：真实值

GCN：图卷积网络（Graph Convolutional Network）

LSTM：长短期记忆网络（Long Short-Term Memory）

MAE：平均绝对误差(Mean Absolute Error)

MAPE：平均绝对百分比误差(Mean Absolute Percentage Error)

MFCC：梅尔倒谱系数（Mel-scale Frequency Cepstral Coefficients）

MSE：均方误差(Mean Squared Error)

RMSE：均方根误差(Root Mean Squared Error)

RNN：循环神经网络(Recurrent Neural Network)

Transformer：注意力模型

5 校园欺凌防控管理系统架构体系

人工智能驱动的校园防霸凌技术系统包含数据采样层、数据处理层、算法层和应用场景层，见图1。

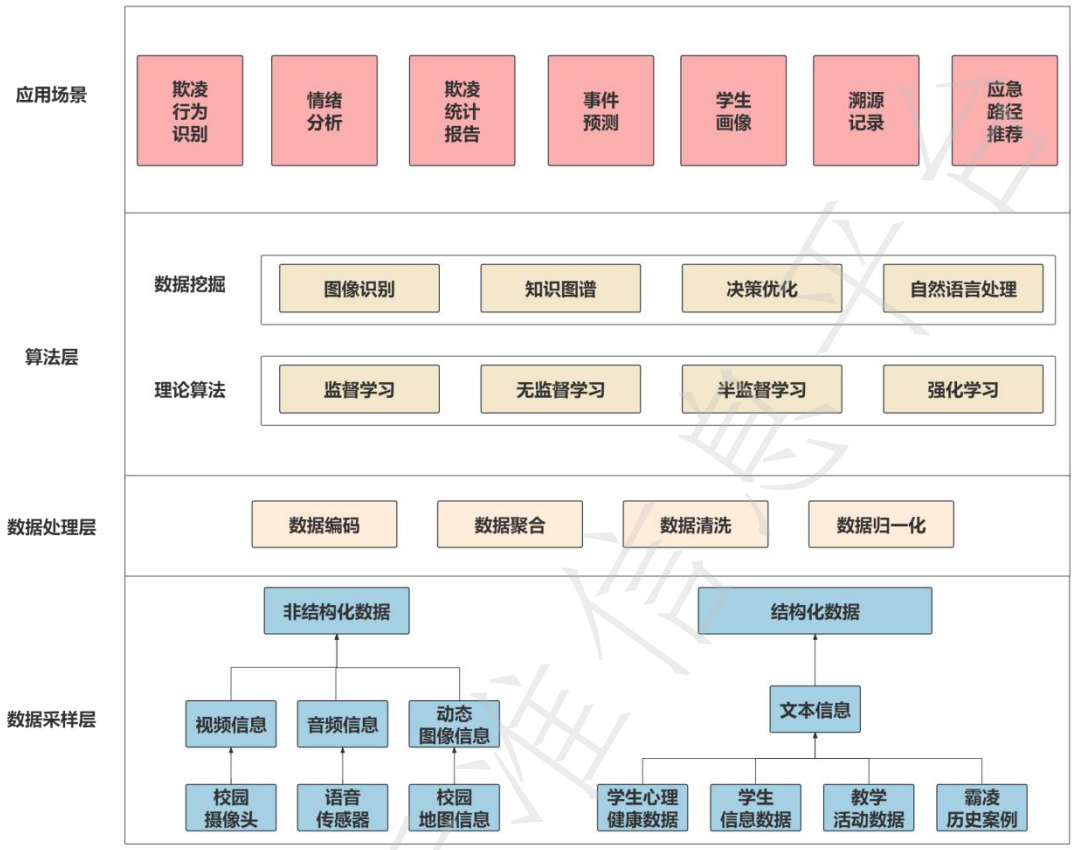

图 1　人工智能驱动的校园欺凌防控管理技术架构体系

其中：

a）　数据采样层通过物理感知设备音视频等非结构化数据，利用问卷、校园数据库收集学生的心理健康数据、教学活动数据等结构化数据；

b）　数据处理层使用数据标注、数据聚合、数据清洗等方法进行数据预处理；

c）　算法层通过图像识别、知识图谱、强化学习、深度学习等人工智能算法实现欺凌行为识别，情绪分析等实际的场景应用。

6　多模态数据集成融合分析要求

6.1　数据来源

采样层应采集下列信息。

a）　公开数据：包含校园内的监控录像、录音、红外传感数据、雷达采集数据等公开数据。

b）　真实校园采集数据：校园内的实时监控录像、录音、红外传感等数据。

c）　学生个人信息：包含学生成绩、年龄、性别、照片、性格趋向，个人关系等信息。

d）　教学信息：包含班级信息，年级信息，教学作息安排，课程成绩等相关教学数据。

e）　校园地理位置信息：包含相关建筑、设备位置、重点监控区域等学校二维和三维的校园地理位置数据。

6.2　数据处理

6.2.1　数据清洗

应采用下列方法进行数据清洗。

a）　信噪比估计算法：用于去除信噪比低的语音识别音频。

b）　去除质量较低或没有有效信息的图像，例如模糊、全黑的图像。

c) 数据去重：利用固定值、均值、中位数的填充方法，并进行分词、停用词去除、词干化等处理去除文本中出现重复或缺失数据。
d) 异常检测：用算法修正因机器故障而产生的传感器和红外数据。

6.2.2 数据标注

针对不同分类的识别任务，应根据数据的不同类型或不同类别进行标注处理，用于算法模型的训练。

注：常见标注类别有文本标注、语音标注、图像标注、3D点云标注等。

6.3 数据分析

6.3.1 特征提取

针对不同类型的数据，应采用下列特征提取技术。

a) 文本数据：使用词袋模型、词频—逆文本频率指数（TF-IDF）、词向量（Word2Vec）、双向编码器表示（BERT）、N 元词袋（bag of n grams）等技术提取文本数据特征。
b) 图像数据：使用尺度不变特征转换（SIFT）、关键点快速创建特征向量（ORB）、方向梯度直方图（HOG）、局部二值模式（LBP）、特征和级联分类器（HAAR）等算法和模型描述图像，并区分图像类别。
c) 音频数据：MFCC、RNN、小波变换（WT）等模型分析并提取线性预测编码（LPC）、感知线性预测编码（PLP）、MFCC、滤波器组（Fbank）、语谱图等音频特征。

6.3.2 特征融合

应采用下列方法进行特征数据的融合。

a) 特征拼接：将不同模态的特征向量端到端拼接在一起，形成一个长向量。
b) 特征相加：将不同模态的特征向量相加，得到一个新的特征向量。
c) 特征相乘：将不同模态的特征向量相乘，得到一个新的特征向量。

7 算法构建要求以及评估指标

7.1 构建要求

7.1.1 图像识别技术

7.1.1.1 主成分分析

对于维度较高的图像数据，采用主成分分析方法进行预处理，应支持下列功能要求：

a) 按目标要求构建校园监控视频数据集，同时将视频数据分解为连续的独立帧，将每帧图像数据转换为数值矩阵形式；
b) 通过计算图像数据的协方差矩阵进行特征值分析，找到能够最大限度表示原始数据分布的主成分；
c) 保留原始图像数据的结构和信息，通过线性变换将高维数据映射到低维空间，使得在保留尽可能多信息的前提下，图像数据的维数得以降低。

7.1.1.2 支持向量机

在欺凌行为识别过程中应满足下列要求：

a) 按目标要求构建带有类别标签的数据集，对训练数据集和测试数据集进行划分；
b) 对模型进行迭代训练，完成训练后更新模型参数，并对训练好的模型进行测试；
c) 通过摄像头捕捉校园内的实时图像，并使用支持向量机模型进行欺凌行为识别；
d) 通过人脸识别，使用支持向量机识别出涉及欺凌事件的学生个体。

7.1.1.3 卷积神经网络

卷积神经网络实现欺凌行为识别和暴力倾向分析的功能，构建过程中应满足下列要求：

a) 按目标要求构建带有类别标签的数据集，对训练数据集和测试数据集进行划分；

b) 对模型进行迭代训练，完成训练后更新模型参数，并对训练好的模型进行测试；
c) 使用卷积神经网络对校园监控视频进行实时的欺凌行为和事件的识别分析；
d) 在表情语义分析方面，使用卷积神经网络分析图像中学生的面部表情等生理信号，准确判别可能受到欺凌或有欺凌倾向的学生。

7.1.1.4 图卷积网络

采用 GCN 实现欺凌事件空间关联关系的分析功能，构建过程中应满足下列要求：
a) 将校园图像中每个像素点视作图中的一个节点，像素之间的相邻关系形成图的边，同时使用 GCN 有效捕获这些像素之间的关系，完成基于空间关系的图像分割；
b) 将欺凌对象识别作为图结构预测任务，使用 GCN 提取对象各部分之间的结构信息，准确识别欺凌对象；
c) 在复杂的校园欺凌防控场景理解任务中，校园内物体之间的关系能被建模为图结构。

7.1.2 语音识别技术

7.1.2.1 Transformer

在校园欺凌防控中，Transformer用于欺凌事件中的语音分析，应满足下列要求：
a) 按目标要求构建训练数据集，并完成对模型的迭代训练和测试；
b) 通过使用 Transformer 分析语音内容中的抱怨、威胁、侮辱等关键词，侦测和识别学生可能发生的欺凌行为；
c) 使用 Transformer 分析涉及多个学生的对话场景，判断复杂的欺凌行为事件。

7.1.2.2 循环神经网络

使用RNN识别复杂的学生欺凌行为，构建过程中应满足下列要求：
a) 按目标要求构建训练数据集，并完成对模型的迭代训练和测试；
b) 使用 RNN 分析学生的语音数据，准确地判别相关的语气变化、情绪状态和语言习惯，识别是否存在欺凌的行为。

7.1.2.3 长短期记忆网络

使用LSTM处理长周期的校园欺凌防控分析任务，构建过程中应满足下列要求：
a) 按目标要求构建训练数据集，并完成对模型的迭代训练和测试；
b) 使用 LSTM 持续监测和分析学生的语音数据，通过长期分析学生的语音模式变化趋势，通过识别他们的情绪变化、语言习惯的改变判断可能发生的欺凌迹象；
c) 通过 LSTM 分析多个学生之间长时间交流的复杂对话，识别复杂场景下的欺凌行为。

7.1.2.4 高斯混合模型

可以使用高斯混合模型（GMM）识别校园欺凌行为，构建过程中应满足下列要求：
a) 按目标要求构建训练数据集，完成对模型的迭代训练和测试；
b) 支持分析学生的语音特征的功能，通过识别异常的语音模式，并判断欺凌事件是否发生。

7.1.3 多模态预训练模型

多模态预训练模型在构建过程中应满足下列要求：
a) 按目标要求构建训练数据集，并完成对模型的迭代训练和测试；
b) 使用自然语言大模型分析学生在社交媒体、校园论坛、聊天记录等文本数据中的言论，发现含有侮辱、歧视等不当言论的文本，识别带有强烈负面情绪（如愤怒、恐惧等）的文本信息；
c) 使用语音大模型分析学生在语音或视频通话中的言论，识别出含有恶意威胁、侮辱的语音信息以及相关的语调、语速等元信息；
d) 使用视频大模型来分析校园海量的监控视频，识别推搡、打击等欺凌行为。

7.2 评估指标

7.2.1 平均绝对误差

MAE 是一种用于评估预测模型或者估计方法的常用指标，它计算的是预测值与实际观测值之间差的绝对值的平均。MAE 的值越小，表示预测模型的准确性越高。MAE 能衡量预测模型的准确性，得到预测结果和实际情况的偏差程度。见公式（1）。

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \quad (1)$$

7.2.2 均方误差

MSE是一种常用的评价指标，它计算的是预测值和实际观测值之间的差值的平方的平均。与MAE相比，MSE能更详细地衡量预测模型的准确性，尤其是对模型是否能够处理异常值的评估。见公式（2）。

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2 \quad (2)$$

7.2.3 均方根误差

RMSE 是 MSE 的平方根。与 MSE 相比，RMSE 更能反映模型预测的平均偏差。在欺凌行为预测中，RMSE 用来衡量预测模型的准确性，尤其是在评估模型预测的一般性能力上。见公式（3）。

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \quad (3)$$

7.2.4 平均绝对百分比误差

MAPE 是一种相对误差指标，计算的是预测值与实际值之间的绝对误差的百分比的平均。在欺凌行为预测中，MAPE 能反映预测误差相对于实际值的大小，展现预测精度的相对性能。见公式（4）。

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - x_i}{y_i}\right| * 100\% \quad (4)$$

7.2.5 准确率

准确率是分类问题中最常用的评估指标，计算的是预测正确的样本数占总样本数的比例。在欺凌行为预测中，准确率能衡量模型在正确预测欺凌行为和非欺凌行为方面的能力。见公式（5）。

$$A = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

式中：

A——准确率；

TN——真负例（正确预测的非欺凌行为）；

FP——假正例（错误预测的欺凌行为）。

7.2.6 召回率

也称为敏感性，即预测为正样本（即欺凌行为）且实际为正样本的比例占所有实际为正样本的比例。在欺凌行为预测中，召回率用于衡量模型在找出所有实际欺凌行为方面的能力。见公式（6）。

$$R = \frac{TP}{TP+FN} \quad (6)$$

式中：

R——召回率。

8 应用场景功能构建要求

8.1 功能要求

8.1.1 知识检索

知识检索系统架构见图 2，应支持下列功能：

a) 从 pdf、doc、csv、txt、png、ppt 等格式的文档中采集数据；

b) 文档数据的向量化编码，并存储在向量数据库中；

c) 在用户提问过程中，根据用户的问题在向量数据库中进行相似答案的召回，将多个召回的相似答案按照大模型的提示词模板进行格式组装；

d) 基于大模型输出检索结果，并注明结果所在文档的出处；

e） 法律法规检索、处理建议推荐、心理辅导和智能多轮对话功能。

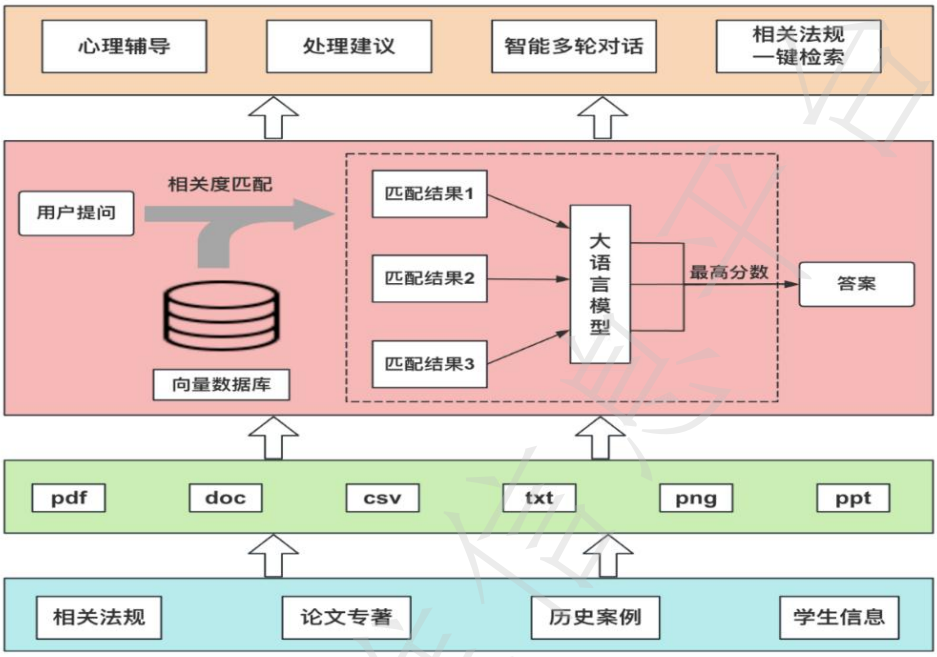

图 2 基于大模型的校园欺凌防控专家知识系统架构

8.1.2 行为识别

应支持下列行为识别功能。

a） 监控图像分析：利用监控摄像头识别和检测异常行为。

b） 红外图像分析：利用红外摄像头识别和检测人体发出的热辐射。

c） 传感波形分析：利用传感器捕捉到的波形信号，识别和检测异常声音和震动。

8.1.3 场景分析

应支持下列场景分析功能。

a） 行为判定：通过语义识别提取对话关键词和主题，判断内容是否与霸凌相关。

b） 时间预测：通过对视频或图像中的光线等因素判断霸凌行为发生的时间，并对霸凌高发时段进行预测。

c） 场所定位：语音和语义识别周围环境音，并分析得出霸凌行为发生场所。

d） 还原现场：通过语音识别或视频图像处理判断参与霸凌的人数及霸凌的方式，结合语音转文字和语意识别技术，通过情绪识别分析文本数据中的情感倾向和情绪变化。

8.1.4 交互预警

交互预警应支持下列模式。

a） 主动报警：当检测到霸凌事件时，自动触发报警机制，并通知学校管理人员或警方进行处理。

b） 自适应预警：通过历史数据和行为分析，自动调整报警敏感程度。

8.1.5 风险识别

风险识别功能应支持通过预测建模、行为模式分析实现，其中：

a） 预测建模：利用历史数据和实时行为分析，构建机器学习模型预测潜在的欺凌行为；

b） 行为模式分析：分析学生行为，使用神经网络、时间序列分析、自然语言识别等人工智能算法识别欺凌行为模式。

8.1.6 欺凌报警

应支持下列欺凌报警功能：

a) 通过运动检测和文本分析实时分析学生行为；
b) 自定义行为阈值，超过阈值自动触发报警；
c) 通过人机交互设备或终端进行报警。

8.2 公共场所场景应用

在公共场所中，校园欺凌防控管理系统应：

a) 通过在公共场所（如操场、食堂、图书馆）安装高分辨率摄像头和麦克风阵列采集学生群体的行为图像和语音信息；
b) 通过图像识别和语音分析算法来判别特定的行为和声音模式，当算法检测到重复的快速异常运动或听到尖叫声时，模型能判断正在发生身体冲突，并触发警报；
c) 监控数据的采集和处理符合 GB/T 35273 的要求，并通过加密和访问控制来确保数据安全。

8.3 隐私场所场景应用

在宿舍、厕所等隐私场所中，校园欺凌防控管理系统应在保护学生隐私的前提下，使用声音传感器和其他非侵入性监测技术来检测异常行为，支持下列功能：

a) 通过基于声音传感器的波形分析来监测和分析学生的情绪状态，判别某个区域内的学生表现出恐惧或焦虑的情绪特征；
b) 监听特定的关键语句和语调，如侮辱性质的言语或威胁性的语调；
c) 利用红外摄像头捕捉到的图像，通过算法进行分析以检测学生的状态和行为；
d) 利用微波雷达和毫米波雷达等技术对学生的异常行为和位置等信息进行检测。