

一种时序信号分类算法在纳米孔道离子电流信号识别中的应用^{*}

倪雪^{†,a} 辛凯莉^{†,b} 胡正利^{*,b} 蒋翠玲^{*,a}

万永菁^a 应佚伦^b 龙亿涛^b

(^a 华东理工大学信息科学与工程学院 上海 200237)

(^b 南京大学化学化工学院 分子传感与成像中心 南京 210023)

摘要 纳米孔道单分子分析技术通常利用阻断电流的时域特征进行识别。但对于结构、分子量等相似的物质,由于其时域特性交叠,采用传统的纳米孔道识别方法难以准确分辨。为了充分挖掘具有差异性的深层特征,提升纳米孔道离子流信号识别准确率,提出了一种时序信号分类算法。通过有重叠的滑动窗口对原始信号进行分帧,并利用连续小波变换对逐帧信号进行处理,可准确获取单分子事件的时频域浅层特征信息。在此基础上,利用多分支层间特征融合网络处理获取深层特征。采用可信统计预测策略对子信号的分类概率统计,该算法对单氨基酸差异多肽的纳米孔道离子电流信号的识别准确率高达 99.00%,可显著提高纳米孔道对分子量相似甚至相同的单分子的传感能力。

关键词 纳米孔道分析; 时序信号; 深层特征提取; 层间特征融合; 可信统计预测

A Time-Series Signal Classification Algorithm and Its Application to Nanopore Ionic Current Signal Identification^{*}

Ni, Xue^{†,a} Xin, Kaili^{†,b} Hu, Zhengli^{*,b} Jiang, Cuiling^{*,a}

Wan, Yongjing^a Ying, Yi-Lun^b Long, Yi-Tao^b

(^a School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237)

(^b School of Chemistry and Chemical Engineering, Molecular Sensing and Imaging Center (MSIC), Nanjing University, Nanjing 210023)

Abstract Nanopore-based single molecular analysis technique usually uses time-domain features such as time-current scatter plots of blocking currents for event recognition. However, as the time-domain features overlap with each other, the substances with extremely similar molecular structures are difficult to be accurately discriminated using traditional nanopore recognition methods. The differences in the deep feature representations need fully explored to obtain credible recognition results, thus improving the recognition accuracy of nanopore ionic current signals. Here, a time-series signal classification algorithm is proposed in this paper: firstly, the original signal is framed with overlapping sliding windows to generate sub-signals and extract their shallow feature information; then a time-series signal classification network based on Emphasized Channel Attention, Propagation and Aggregation in time delay neural network (ECAPA-TDNN) is proposed to develop a multi-branch inter-layer feature fusion model for deep feature extraction, where the multi-branch multi-level attention module of this model (RepVGG-SE-Res2Block, RSR-Block) obtains multi-scale features by constructing a feature pyramid structure within each residual block, reduces the inference speed based on structural reparameterization techniques while ensuring the model performance, and introduces Adaptively Spatial Feature Fusion (ASFF) to fuse the features of different layers in the network; finally, a credible statistical prediction strategy is used to obtain reliable classification results by counting the classification probabilities of sub-signals. The experimental results show that for the peptide sequences *N'*-DDFFIFFDD-*C'* (DF_I) and *N'*-DDFFLFFDD-*C'* (DF_L) containing only the different amino acids I (isoleucine) and L (leucine), which are isomers of each other, the algorithm achieves a recognition accuracy of 99.00%, obviously improving the sensing capability of nanopores for single molecules with similar or even identical molecular weights.

Keywords nanopore analysis; time-series signal; deep feature extraction; inter-layer feature fusion; credible statistical predictions

^{*} Dedicated to the 90th anniversary of *Acta Chimica Sinica*.

^{*} E-mail: zhenglihu@nju.edu.cn; cuilingjiang@ecust.edu.cn

Received April 3, 2023; published June 25, 2023.

Supporting information for this article is available free of charge via the Internet at <http://sioc-journal.cn>.

Project supported by the Ministry of Science and Technology Key R&D Program of China (No. 2022YFA1304604) and the National Natural Science Foundation of China (Nos. 22106066, 22027806).

^{*} 庆祝《化学学报》创刊 90 周年。

项目受科技部重点研发计划(No. 2022YFA1304604)和国家自然科学基金(Nos. 22106066, 22027806)资助。

[†] 共同第一作者(These authors contributed equally to this work).

1 引言

纳米孔道技术作为一种高灵敏的单分子分析技术,在组学研究^[1-4]、临床诊断^[5-8]、环境监测^[9-10]等多个研究领域均存在广泛应用. 其中,生物纳米孔道因具有精准可调的传感界面以及高度限域的电化学环境^[11-12],近年来在 DNA^[13]、蛋白质^[14]以及代谢小分子^[15-16]的测量研究过程中展现了出色的检测分辨能力. 典型的生物纳米孔道装置由嵌入膜中的纳米级孔道和孔两侧的电解质溶液组成. 其基本检测原理是通过在嵌入孔道的脂质双层两侧施加电压,驱动单个分子穿过纳米孔道产生阻断电流信号,对比信号的阻断电流幅值和持续时间等特征的差异,可表征包括大小、电荷、构象和相互作用等在内的多种分子特征^[17]. 然而,仅利用上述常规参数往往难以将阻断事件准确归属于待测对象,尤其是对于组成或结构相似的分子. 以纳米孔道多肽检测为例,由于氨基酸种类复杂,且体积或分子量(其中亮氨酸 L 与异亮氨酸 I 分子量相同均为 113.16)差异较小,引起的电流往往非常相似,以致区分仅含有单氨基酸差异的多肽序列仍具有较大挑战^[18-19].

因此,为进一步提高纳米孔道分辨检测能力,开发可行的算法或多维参数以揭示电流阻断信号的更多特征,并完善对单个捕获分子的描述. 目前已开发了阻断电流标准偏差^[20]、半峰宽(Full width at half maximum, FWHM)^[21]、偏度(Skewness, skew)^[22]等多种时序参数用于疾病相关多肽、蛋白质以及糖类等分子的分辨. 伴随着对于生物纳米孔道传感机制的深入理解,单分子电流信号频域维度信息的发掘为提升纳米孔道分子识别能力提供了一个新的解决思路^[23-24]. 通过分析频率的幅值、平均能量等参数与温度、pH、电压等条件的依赖关系,已建立了孔道内离子指纹图谱的谱学分析方法,证明了上述频域参数与纳米孔道内离子分子相互作用网络特性密切相关^[25].

然而,上述分析方法均基于所提取浅层特征的本征含义,识别准确性受到特征数量的限制,难以进一步提升识别准确率. 随着深度学习的发展,采用神经网络在初步提取时序信号的数字形态学特征、时频特征、声学特征^[26]等浅层特征信息的基础上,结合深度神经网络进一步挖掘更高维度的差异性信息,也称为深层特征,能够有效提升分类识别的准确率.

为充分挖掘单分子事件的频域特征信息,以提升纳米孔离子流信号识别准确率,获得可信的识别结果,本文提出了一种基于深度学习网络的时序信号分类算法. 该算法设计了具备结构重参数化性能的多分支多层次注意力模块(RepVGG-SE-Res2Block, RSR-Block)用于获取多尺度的感受野,引入 ASFF 模块融合经过 RSR-Block 模块得到的特征,增强网络对不同层特征的重要性感知,从而提升了神经网络的性能以获取更具差异性的深层特征信息,同时制定了可信统计预测策略,获得

了更好也更具解释性的分类识别效果. 实验结果表明,该算法对仅含单同分异构氨基酸差异多肽序列识别准确率可达 99.00%,显著提高纳米孔道对分子量相似甚至相同的单分子的传感能力.

2 算法模型构建

2.1 数据获取

本文数据获取自突变型 Aerolysin 生物纳米孔道检测 DF_L 与 DF_I 多肽,序列分别为 N'-DDFFLFFDD-C' 和 N'-DDFFIFFDD-C'(详见实验部分与支持信息 S1).

2.2 整体算法架构

本文提出了基于层间特征融合网络的纳米孔道离子流信号识别算法(整体架构见图 1d),主要分为浅层特征提取、深层特征提取和可信统计预测 3 个阶段.

(1)浅层特征提取阶段. 该阶段通过子信号分帧方法将原始信号分为 N 段,即 $x_i(i=1,2,\dots,N)$,对每个子信号采用连续小波变换(Continuous Wavelet Transform, CWT)提取时频特征 CWT_i .

(2)深层特征提取阶段. 将浅层特征提取阶段获得的子信号 CWT 特征作为输入,采用本文提出的多分支层间特征融合网络以提取深层特征(该网络结构见图 2),最终输出类别的概率值 p^i .

(3)可信统计预测阶段. 将深层特征提取阶段输出的概率值作为输入,通过联合狄利克雷分布将其转换为置信质量 $b_k^i(k=1,2,\dots,K)$ 以及不确定性质量 u^i ,其中 K 为类别的总数量. 不确定性质量越小,说明当前的预测可信度越高. 根据不确定性质量与投票策略提出可信统计预测策略进行综合判断,最小化风险决策,得出可信的总体预测标签.

2.3 浅层特征提取

浅层特征提取阶段主要包含使用滑动窗口进行有重叠的子信号分帧与 CWT 时频特征提取两个步骤. 其中,子信号分帧方法通过对整段阻塞离子电流信号进行处理,可获得逐帧子信号,在这个过程中需选取参数:子信号长度 l 与相邻窗口的重叠部分占比 r (具体过程见于支持信息 S3). 时频分析方法包括短时傅里叶变换、连续小波变换和希尔伯特-黄变换等. 然而,短时傅里叶变换受固定的时间和频率分辨率限制,而希尔伯特-黄变换则依赖于模态分解方法,容易产生模态混叠和端点效应,且其提取的特征矩阵稀疏,所包含的信息较为有限. 因此,本文选择采用连续小波变换作为浅层特征提取的方法. CWT 时频特征提取方法利用了 CWT 的多尺度分析能力,能为相似分析物引起的离子电流信号提供准确且丰富的时频信息,通过对逐帧子信号进行处理,提取了信号的局部频域特性.

CWT 的原理是通过小波基函数 $\psi_{\tau,s}(t)$ 对原始时域信号 $x(t)$ 进行内积运算,然后分解为不同尺度下的小波

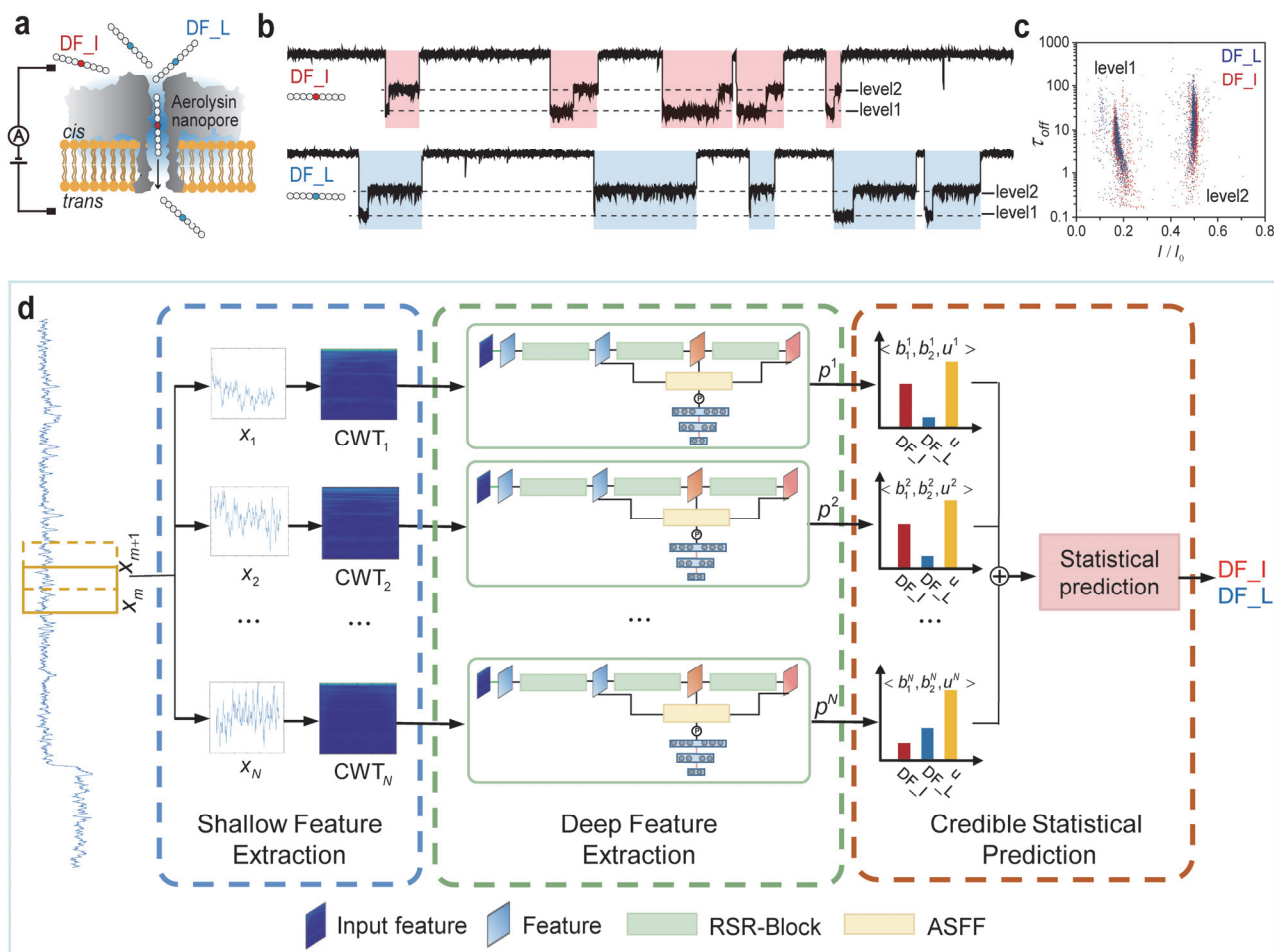


图1 (a)突变型 Aerolysin 纳米孔道检测多肽的原理示意图. 多肽 DF_L 序列为 N -DDFFLFFDD-C', 多肽 DF_I 序列为 N -DDFFIFFDD-C'. 多肽 DF_L 与 DF_I 的原始离子电流信号图(b)和时间-电流散点图(c). I/I_0 表示单个多肽分子进入纳米孔道产生的残余电流程度, 其中 I_0 为单个 Aerolysin 孔道产生的开孔电流, I 为分子进入孔道后的残余电流, τ_{off} 为单个多肽分子在孔道中的停留时间. (d)基于多分支层间特征融合网络的纳米孔道离子电流识别算法流程图

Figure 1 (a) Mutant aerolysin nanopore detection of single peptides. The sequences of DF_L and DF_I peptides are N -DDFFLFFDD-C' and N -DDFFIFFDD-C', respectively. (b) Ionic current signals, and (c) time-current scatter plot of different peptide substances. Herein, I/I_0 indicates the degree of residual current generated by a single peptide molecule entering the nanopore channel, where I_0 is the opening current generated by a single Aerolysin pore, I is the residual current after the molecule enters the pore, and τ_{off} represents the residence time of a single peptide molecule in the pore. (d) Flowchart of a nanopore ionic current signal recognition algorithm based on multi-branch interlayer feature fusion networks

变换系数 $\lambda_{\tau,s}(t)$:

$$\lambda_{\tau,s}(t) = \int_{-\infty}^{\infty} x(t) \psi_{\tau,s}^*(t) dt \quad (1)$$

$$\psi_{\tau,s}(t) = \frac{1}{\sqrt{|s|}} \psi\left(\frac{t-\tau}{s}\right), \quad s, \tau \in R, s \neq 0 \quad (2)$$

式中, s 为分解尺度, τ 为平移参数, t 为分解时间, $\psi_{\tau,s}^*(t)$ 为小波基函数 $\psi_{\tau,s}(t)$ 的共轭运算.

本文采用具有强大分辨多尺度功能以及良好的时、频域局部化优势的复 Morlet 小波(详见支持信息 S3)作为小波基函数来实现 CWT^[27], 复 Morlet 小波的定义如下:

$$\psi = \pi^{-\frac{1}{4}} \left(e^{i2\pi f_0 t} - e^{-\frac{(2\pi f_0)^2}{2}} \right) e^{\frac{t^2}{2}} \quad (3)$$

式中 f_0 为 Morlet 频率常数($f_0=0.849$).

根据分解尺度与频率的对应关系, 连续小波变换后的时频图上同时反映了离子流信号的频率成分随时间变化的关系(详见支持信息 S3). 通过观察频率轴的连续结构与离散点, 分析时间轴上的持续区间与不连续点, 并细致检查能量分布, 可区分信号频率和噪声频率及其对应时间区间. 此外, 还可以通过可视化分析时频图的形态、频谱轮廓和能量分布来进一步区分. 本文中获取的 CWT 特征的尺寸大小为 $C \times T$, 其中 C 表示特征的维度, T 表示子信号的长度, 为了获得较高的频率分辨率将 C 取为 256.

2.4 深层特征提取

深层特征提取阶段提出了多分支层间特征融合网络, 旨在进一步挖掘单分子事件的动态频域特征信息,

以获取有助于区分相似分析物的特征表达. 该网络以加强通道注意力和传播聚合的时序信号分类网络^[26]作为骨架结构, 构建了具有结构重参数化功能的 RSR-Blocks 模块, 采用 ASFF 模块对三层 RSR-Block 模块输出的中间层特征进行深度融合(骨架网络的详细信息见支持信息 S4, 本节所提出的网络结构见图 2).

该网络输入为第 v 个子信号的 CWT 特征, 首先通过卷积核大小为 5 的一维卷积将输入尺寸的通道数升至 512, 即此时的特征尺寸大小为 $512 \times T$; 随后通过三个具有不同大小膨胀卷积($d=2, 3, 4$)的 RSR-Blocks 块结构, 构造多分支多层次注意力模块以获得多尺度特征, 并显式建模通道间的相互依赖关系, 通过每个 Block 块后特征尺寸始终保持为 $512 \times T$; 将三层 Block 结构的输出特征作为 ASFF 模块的输入, 通过学习各层的权重系数 α, β, γ 得到更具判别性的深层特征矩阵 y ; 参照骨架网络的参数设置, 通过注意力统计池化层将融合后的特征降维至 3072 维, 通过全连接层获得 192 维的嵌入级特征, 最后经过一个全连接层获得类别的概率值 p^v .

2.4.1 RSR-Block

结构重参数化技术能够在提升模型性能的同时减少推理时间, 基于骨架网络中的 SE-Res2Block, 本文设计了多分支多层次注意力模块 RSR-Block, 采用具有结构重参数化的 RepVGG 模块代替 Res2Block 中的单路 3×3 小卷积, 在不影响模型性能的前提下极大提升了推理效率(其网络结构见图 2).

SE-Res2Block 采用多尺度膨胀卷积有效地丰富感受野并捕捉了时序信号特征的长期变化(相关理论知识见支持信息 S4). RepVGG 模块在训练阶段, 采用了具有不同卷积核大小以及残差连接的多分支拓扑结构, 解决了随模型深度产生的退化问题和梯度消失问题, 使模型在训练阶段更易于收敛, 模型精度更高; 在推理阶段, 通过参数融合的方法将多个支路的卷积大小都转换为 3×3 卷积从而合并为单路结构, 获得运算效率的提升.

2.4.2 ASFF 模块

神经网络中低层特征与原始输入更为相关, 包含更多细节信息, 但是由于经过的卷积更少, 其语义性更

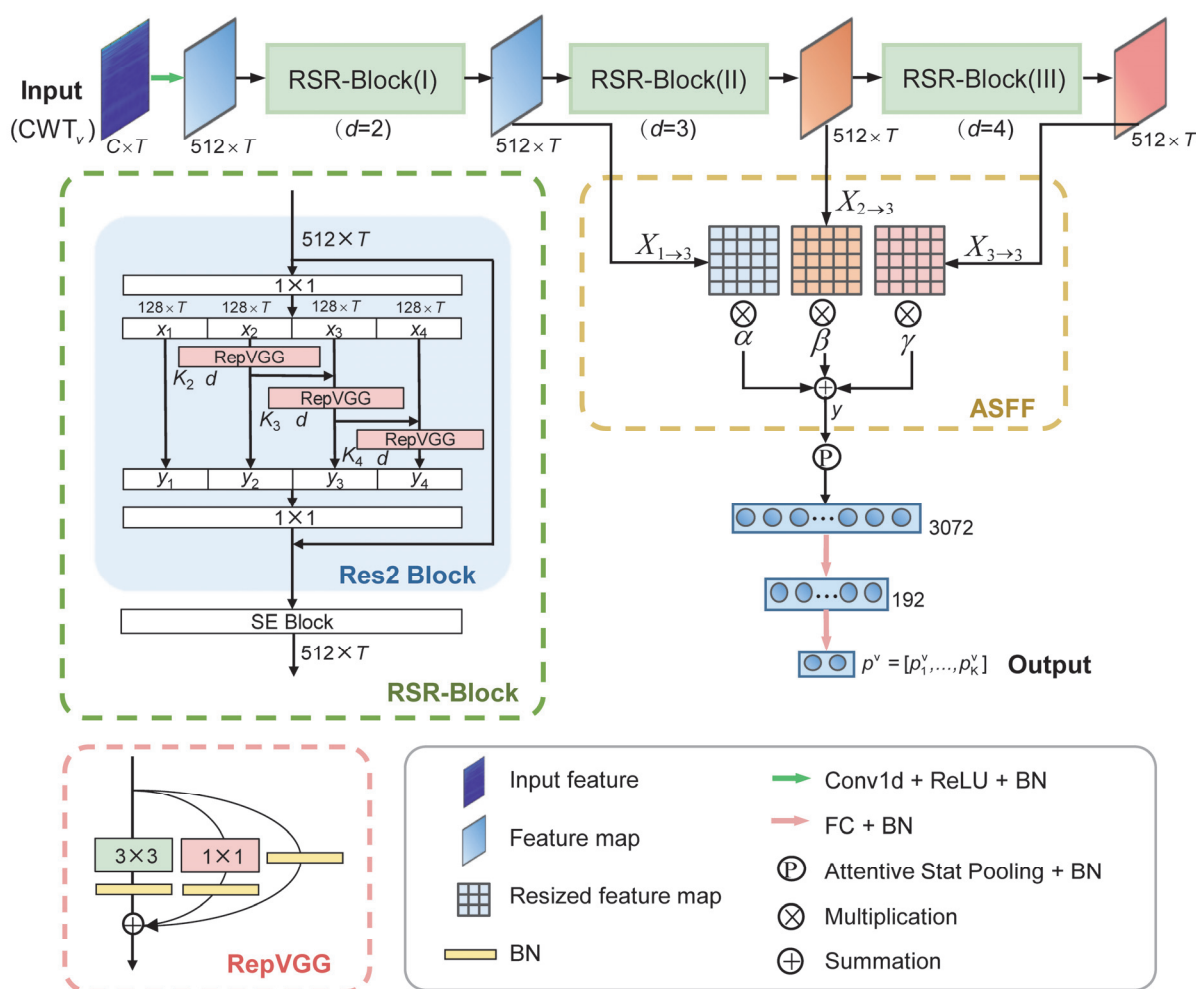


图 2 多分支层间特征融合网络结构图

Figure 2 Structural diagram of multi-branch interlayer feature fusion network

低、噪声更多, 高层特征具有更强的语义信息, 但是对细节的感知不足, 充分融合高低层的特征信息能够改善模型的性能, 获得更好的分类效果.

本文提出将 RSR-Blocks 的输出特征采用 ASFF 模块进行融合(ASFF 模块的结构示意图见图 2). 首先将不同层的特征进行恒等缩放至统一尺寸, 如 $X_{1 \rightarrow 3}$ 表示 RSR-Block (I)层的特征经过缩放后变换为 RSR-Block (III)层的特征尺寸大小, 将各层特征分别与可学习的权重系数 α, β, γ 相乘, 最终相加得到层间融合的特征 y .

$$y = \alpha \cdot X_{1 \rightarrow 3} + \beta \cdot X_{2 \rightarrow 3} + \gamma \cdot X_{3 \rightarrow 3} \quad (4)$$

$$s.t. \alpha + \beta + \gamma = 1 \quad \alpha, \beta, \gamma \in [0, 1]$$

2.5 可信统计预测

针对多个子信号获得的结果, 可以通过投票策略提高模型的泛化能力, 减少模型的错误率. 常用的投票策略主要分为三类: 多数取胜策略, 相对多数取胜策略以及加权投票策略. 然而以上投票策略具有随机性或者需建立在先验知识的基础上, 本文采用了概率值联合狄利克雷分布得出的不确定性质量作为多数取胜策略的一个投票标准, 获得更为可信的预测结果, 可信统计预测分为以下三个步骤:

(1) 经过本文模型后, 得到第 v 个子信号的概率值 $p^v = [p_1^v, \dots, p_K^v]$ ($k = 1, 2, \dots, K$), 其中 K 为类别的总数量, 将其与狄利克雷分布 $\alpha^v = [\alpha_1^v, \dots, \alpha_K^v]$ 相关联, 计算得出各类别的置信概率 b_k^v 以及不确定性质量 u^v , 不确定性质量的大小与模型可信程度呈负相关, 计算公式如式(5), (6)所示:

$$b_k^v = \frac{p_k^v}{S^v} = \frac{\alpha_k^v - 1}{S^v} \quad (5)$$

$$u^v = \frac{K}{S^v} \quad (6)$$

式中 S^v 为狄利克雷强度.

$$S^v = \sum_{i=1}^K (p_i^v + 1) = \sum_{i=1}^K \alpha_i^v \quad (7)$$

(2) 根据置信概率的大小判断子信号所属类别, 统计各类所包含的子信号数量, 以投票的方式进行数量大小的比较, 将数量多者作为此条数据的预测类别.

(3) 若经步骤(2)统计后归属于两类待测物分子子信号数量相同, 则计算各类别的平均不确定性质量, 选取值较小的一类作为此条数据的预测类别.

经过以上步骤, 可以获取整段阻塞离子电流信号的预测信息, 由于综合判断过程联合了投票策略以及不确定性质量, 因此最终的结果决策风险较小, 并具有一定的可解释性.

3 实验结果及分析

3.1 纳米孔道相似分子检测实验

为探索优化纳米孔道频谱深度学习算法, 以提升纳

米孔道对于相似分子的检测能力, 在此, 本文采用序列分别为 N' -DDFFIFFDD- C' (DF_I)和 N' -DDFFLFFDD- C' (DF_L)的仅含单氨基酸差异的模型多肽进行纳米孔道单分子检测. 鉴于所选多肽仅含的差异氨基酸 I(异亮氨酸)、L(亮氨酸)互为同分异构体, 分子量与体积均相似, 在此考虑利用不同氨基酸与孔壁相互作用差异, 协助完成分子的特异性识别. 在原有的 Aerolysin 纳米孔道灵敏区域, 即 R1 区域(核心位点为 R220)与 R2 区域(核心位点为 K238)之间构建了第三灵敏域^[28-29], 将 226 位点的天门冬酰胺 N 突变为体积更大的谷氨酰胺 Q, 270 位点电中性小体积的甘氨酸 G 突变为正电性大体积的赖氨酸 K, 获得突变型 N226Q/G270K Aerolysin 纳米孔道(见支持信息, 图 S4). 通过增强孔道与多肽分子的相互作用, 延长多肽分子在孔道中的停留时间, 以期放大单分子信号的频谱差异. 通过施加 +80 mV 电压, 两种待测多肽分子被驱动进入突变型 N226Q/G270K Aerolysin 生物纳米孔道, 产生了具有相似形态的多台阶信号. 本文选取信号数量占比最高且电流分布较集中的双台阶信号用作后续分析. 其由残余电流量大、时间短的第一台阶与残余电流量小、时间长的第二台阶组成(图 1b). 这或许是由于新构建的第三灵敏域与原有传感区域位置相近, 可共同作用于多肽, 从而显著阻断孔道内的离子流; 随后在快速离开该共同作用域后, 所构建的第三灵敏域产生了较大的能垒, 使得分子难以顺利通过而显著延长孔道过孔时间产生第二台阶. 进一步地, 处理获取多肽分子特征信号的残余电流量 I/I_0 (其中 I_0 为孔道的开孔电流, I 为分子进入孔道后的残余电流)与阻断时间 τ_{off} . 从二维散点图(图 1c)中可发现 DF_I 与 DF_L 两种多肽的特征信号分布基本重合, 其电流台阶残余电流量水平 I/I_0 差异低于 0.01. 然而, 据统计两者的阻断时间仍存在差异(见支持信息, 图 S5), 这表明两者在孔道中的相互作用不同, 但因环境噪声等影响难以利用常规参数对其进行分辨. 因此, 本文以该实验数据集为主要检测对象, 采用连续小波变换电流数据进行时频转换获取单分子时频图, 深入挖掘电流“噪声”中的深层特征, 并利用优化深度学习模型对其进行分辨.

3.2 子信号分帧参数选取实验

选取合适的子信号分帧参数有利于减少模型分类识别的误差, 实验通过分别改变子信号长度 l 与重叠部分占比 r 的大小, 并对本文所提出的网络进行重新训练, 探究参数 l, r 的设置对模型的影响. 训练得到了不同模型在测试集上的验证结果(见表 1).

当子信号长度不变时, 重叠部分占比越大, 获取的训练样本数量越多, 准确率越高; 当维持重叠部分占比不变时, 子信号越长越接近于原始信号, 获取的信息更多, 从而获得的准确率越高. 实验表明, 为获得更高的准确率, 应当设置较大的子信号长度及较大的重叠部分占比, 本文最终在子信号分帧时参数设置为窗口宽度为

表1 子信号分帧参数对实验结果的影响

Table 1 Effect of sub-signal segmentation parameters on experimental results

l	r	ACC/%
256	1/4	89.00
256	1/2	90.18
256	3/4	91.61
192	3/4	90.89
128	3/4	90.17

256 点, 重叠部分占比为 3/4.

3.3 消融实验

3.3.1 本文算法的有效性

基于子信号分帧参数选取实验的最优参数, 并将识别准确率用作判断分类网络找到合适权重的标准. 通过训练网络模型, 获得了本文算法的最优分类效果. 经过浅层特征提取与深层特征提取的两个阶段后, 本文模型对子信号的识别准确率最高可达 91.61%, ROC 曲线下的面积 Area Under Curve (AUC) 值为 0.971, 其中 AUC 值越大表明当前算法的分类准确度越高(ROC 曲线及 AUC 值的相关内容见于支持信息 S5)^[30]. 在此基础上, 利用可信统计预测的方法, 对子信号的实验结果进行统计, 得到整条信号的标签预测信息, 识别准确率达到 99.00% (实验结果的混淆矩阵见于支持信息图 S3). 实验结果表明, 本文提出的算法很好地实现对极其相似的纳米孔道离子流数据的准确识别.

3.3.2 结构重参数化的有效性

为验证 RepVGG 模块的结构重参数化技术对推理时间的影响, 比较了本文模型在采用结构重参数化方法前后进行 100 次推理所耗费的时间. 未采用结构重参数化方法时, 推理时间为 5.055 s, 结构重参数化后, 推理时间降至 4.633 s, 速度相比提升 8.5%. 将结构重参数化后的模型继续与 Block 中直接采用 3×3 卷积的模型进行比较, 其推理时间仅多出 0.032 s. 实验结果表明, 结构重参数化能有效平衡模型性能与推理时间的关系.

3.3.3 提升网络性能的有效性

为了充分验证本文模型各模块的有效性, 进行模块消融实验(实验结果见表 2, 表 2 中各评价指标相关内容见支持信息 S5). 对于子信号, 采用 RepVGG 模块以及

ASFF 模块显著地提升了模型的性能, 其中 ASFF 模块对于模型性能的提升优于 RepVGG 模块, 在准确率上相差 0.2 个百分点, 同时加入两个模块对于模型的性能提升达到最优. 对于子信号的实验结果统计表明, 本文模型获得较好的效果, 分类准确率达到 99%, F1 分数达到 98.99%.

本文模型中的 RepVGG 模块通过多分支训练方式扩展了模型宽度, ASFF 模块则自适应融合了不同层的特征信息, 从而提取到更具区分度的深层特征. 消融实验的 ROC 曲线(图 3)表明, 改进的 RSR-Block 与 ASFF 模块的特征融合策略使多项指标获得了提升, 验证了本文模型的有效性.

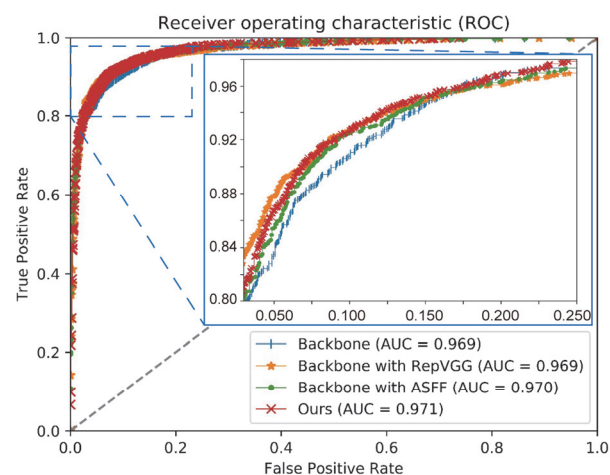


图3 消融实验 ROC 曲线图

Figure 3 ROC curves of ablation experiment

3.3.4 深层特征提取的有效性

为了进一步分析各个模块在深层特征提取阶段的作用, 采用 t-SNE (t-distributed Stochastic Neighborhood Embedding)^[31]进行可视化分析, 它的主要思想是对高维数据进行非线性降维, 该方法能将结果转换至满足分析条件的二维或多维, 实现对样本数据的准确分析. 经过 t-SNE 变换后, 数据如果在低维空间中表现出同类相互聚合、异类相互排斥的特点, 则表明数据之间具有可分性(中间各层所提取到特征数据的可视化分布示意图见图 4).

实验表明, 对于任一数据集, 输入的 CWT 特征均相互交叠, 极难分辨; 经过每一个 RSR-Block 后, 不同

表2 消融实验结果

Table 2 Results of ablation experiments

Models	Subsignal results				Statistical results			
	Recall/%	Precision/%	F1/%	ACC/%	Recall/%	Precision/%	F1/%	ACC/%
Backbone	90.18	90.37	90.25	90.29	98.00	98.00	98.00	98.00
Backbone with RepVGG	91.09	91.27	91.16	91.19	98.00	98.00	98.00	98.00
Backbone with ASFF	91.39	91.36	91.37	91.39	99.00	99.02	98.99	99.00
This work	91.61	91.58	91.59	91.61	99.00	99.02	98.99	99.00

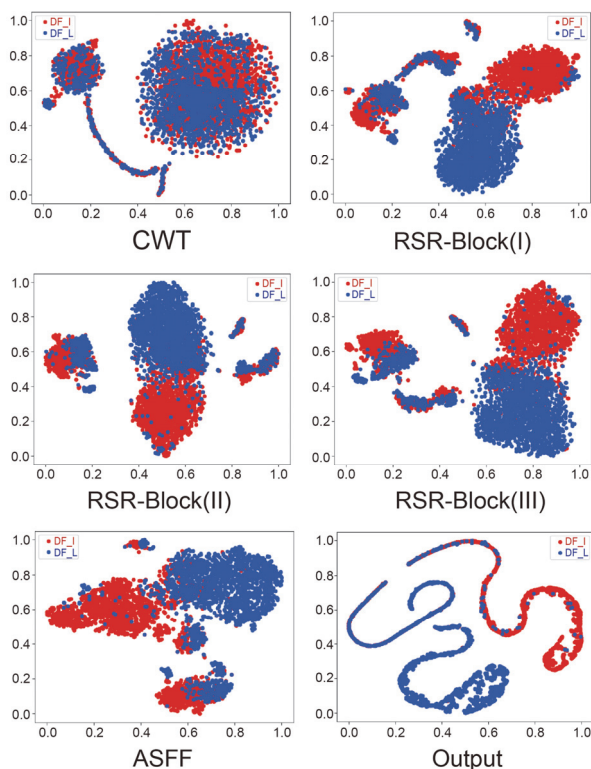


图4 t-SNE 特征分布可视化

Figure 4 Visualization of t-SNE feature distribution

类之间的距离不断扩大, 特征的区分度越来越明显; 使用 ASFF 模块融合层间特征后, 相同类之间不断聚合, 距离缩小, 特征的分辨性进一步提高; 最终在输出层获得较好的分辨效果. 实验结果表明本文所加入的模块能有效地提取到更具分辨能力的特征信息.

3.4 混合数据实验

为进一步验证本文算法的检测效果, 将纯品样本按照一定的比例进行混合, 对于不同比例的混合数据, 预测的平均准确率达到 99.15%, 平均 F1 分数达到 96.24% (见支持信息 S6). 该预测结果表明, 本文算法能够有效识别相似分析物的离子电流信号.

4 结论

本文提出了一种时序信号分类算法, 通过有重叠部分的滑动窗口对原始信号分帧产生子信号并提取浅层时频特征, 搭建了多分支层间特征融合网络, 该网络构建了 RSR-Block 保证在提升检测准确率的同时降低对推理速度的影响, 用 ASFF 模块自适应融合层间特征, 最后通过联合投票策略以及主观逻辑的不确定性质量进行统计预测, 获得可信的预测结果.

实验表明, 该算法对单氨基酸差异同分异构多肽的纳米孔道离子流信号的识别准确率达到 99.00%, F1 分数达到 98.99%. 采用该算法对不同比例的混合数据进行预测, 平均准确率达到 99.15%, 平均 F1 分数达到

96.24%. 本文提出的算法有效提高了单分子离子电流信号识别准确率, 克服了纳米孔道难以准确识别结构相似分析物的难题. 未来该算法可拓展应用于更多场景, 比如辨别多肽中氨基酸修饰或突变状态, 迭代优化后的算法有望在多肽的氨基酸序列探测方面发挥作用.

5 实验部分

纳米孔道检测采用自制的聚四氟乙烯测试器件, 该器件由顺式和反式室组成, 中间夹有 25 μm 厚的聚四氟乙烯片将两个腔室分开. 本文利用自制的电火花发生器在聚四氟乙烯片材上制造了直径为 40 μm 的微孔. 根据 Montal-Mueller 方法, 用十六烷溶液对聚四氟乙烯片孔周围区域进行预处理. 在检测池顺式室和反式室加入 500 μL 的 1 mol/L tris-KCl 电解质溶液, 接着向两个腔室分别加入 1.5~2.0 μL 磷脂溶液, 等待 10 min, 形成单层脂质膜. 随后, 使用注射器装置将气液界面在微孔上下移动, 在此过程中形成脂质双分子层膜. 接下来, 将活化的 N226Q/G270K 突变型 Aerolysin 分子单体加入顺式室, 外加 +200 mV 的偏置电压, 以辅助 Aerolysin 蛋白纳米孔道在脂质双分子层中的自组装. 电压通过一对 Ag/AgCl 电极施加, 并指定顺式腔室为接地端. 在形成单个 Aerolysin 纳米孔道后, 将分析物添加到顺式室中. 在此施加 +80 mV 偏置电压, 驱动分子进入孔道, 引起瞬态电流阻断.

纳米孔道数据由 Axon1440B 与 Axon 200B 设备进行数据采集与信号放大. 所有纳米孔道实验数据都是在 5 kHz 的低通滤波和 100 kHz 的采样频率条件下获得的, 单分子信号 I/I_0 以及 τ_{off} 特征使用实验室开发的软件 “PyNanoLab” (<https://www.pynanolab.com>) 提取. 本工作的纳米孔道实验均在 $(24 \pm 2)^\circ\text{C}$ 进行^[32], 实验中使用缓冲液含 1 mol/L KCl, 10 mmol/L Tris, 待测物 DF_L (序列为 N-DDFFLFFDD-C') 与 DF_I (N-DDFFIFFDD-C') 浓度为 5 $\mu\text{mol/L}$.

致谢 感谢刘少创博士在数据处理软件上的技术支持, 研究生付锡欣、陈天宇、万思远为算法测试提供帮助, 及研究生钟诚兵、张琳琳在仪器运行方面的协助.

References

- [1] Afshar Bakshloo, M.; Kasianowicz, J. J.; Pastoriza-Gallego, M.; Mathé, J.; Daniel, R.; Piguet, F.; Oukhaled, A. *J. Am. Chem. Soc.* **2022**, *144*, 2716.
- [2] Derrington, I. M.; Butler, T. Z.; Collins, M. D.; Manrao, E.; Pavlenok, M.; Niederweis, M.; Gundlach, J. H. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 16060.
- [3] Hu, Z.-L.; Huo, M.-Z.; Ying, Y.-L.; Long, Y.-T. *Angew. Chem., Int. Ed.* **2021**, *60*, 14738.
- [4] Pratanwanich, P. N.; Yao, F.; Chen, Y.; Koh, C. W.; Wan, Y. K.; Hendra, C.; Poon, P.; Goh, Y. T.; Yap, P. M.; Chooi, J. Y. *Nat. Biotechnol.* **2021**, *39*, 1394.
- [5] Goenka, S. D.; Gorzynski, J. E.; Shafin, K.; Fisk, D. G.; Pesout, T.; Jensen, T. D.; Monlong, J.; Chang, P. C.; Baid, G.; Bernstein, J. A. *Nat. Biotechnol.* **2022**, *40*, 1035.

- [6] Sheng, Y.; Zhou, K.; Liu, L.; Wu, H.-C. *Angew. Chem., Int. Ed.* **2022**, *61*, e202200866.
- [7] Ju, Y.; Pu, M.; Sun, K.; Song, G.; Geng, J. *Chem. - Asian J.* **2022**, *17*, e202200774.
- [8] Zhang, Z.; Li, T.; Sheng, Y.; Liu, L.; Wu, H.-C. *Small* **2019**, *15*, 1804078.
- [9] Tsutsui, M.; Yokota, K.; Yoshida, T.; Hotehama, C.; Kowada, H.; Esaki, Y.; Taniguchi, M.; Washio, T.; Kawai, T. *ACS Sens.* **2019**, *4*, 748.
- [10] Tanimoto, I. M. F.; Cressiot, B.; Greive, S. J.; Le Pioufle, B.; Bacri, L.; Pelta, J. *Nano Res.* **2022**, *15*, 9906.
- [11] Wu, X. Y.; Ying, Y. L.; Long, Y. T. *Chem. J. Chin. Univ.* **2019**, *40*, 1825 (in Chinese). (武雪原, 应佚伦, 龙亿涛, 高等学校化学学报, **2019**, *40*, 1825.)
- [12] Wang, L. Y.; Wang, H. W.; Lu, S. M.; Long, Y. T. *Chin. J. Anal. Chem.* **2021**, *49*, 1166 (in Chinese). (王凌远, 王浩炜, 芦思珉, 龙亿涛, 分析化学, **2021**, *49*, 1166.)
- [13] Lu, Y.; Hu, Z. L.; Ying, Y. L.; Long, Y. T. *J. Instrum. Anal.* **2019**, *38*, 129 (in Chinese). (路瑶, 胡正利, 应佚伦, 龙亿涛, 分析测试学报, **2019**, *38*, 129.)
- [14] Chen, Z.; Wang, Z.; Xu, Y.; Zhang, X.; Tian, B.; Bai, J. *Chem. Sci.* **2021**, *12*, 15750.
- [15] Zernia, S.; van der Heide, N. J.; Galenkamp, N. S.; Gouridis, G.; Maglia, G. *ACS Nano* **2020**, *14*, 2296.
- [16] Niu, H. Y.; Hu, Z. L.; Ying, Y. L.; Long, Y. T. *Acta Chim. Sinica* **2019**, *77*, 989 (in Chinese). (牛红艳, 胡正利, 应佚伦, 龙亿涛, 化学学报, **2019**, *77*, 989.)
- [17] Li, M. Y.; Ying, Y. L.; Long, Y. T. *Acta Chim. Sinica* **2019**, *77*, 984 (in Chinese). (李孟寅, 应佚伦, 龙亿涛, 化学学报, **2019**, *77*, 984.)
- [18] Huo, M.-Z.; Li, M.-Y.; Ying, Y.-L.; Long, Y.-T. *Anal. Chem.* **2021**, *93*, 11364.
- [19] Ouldali, H.; Sarthak, K.; Ensslen, T.; Piguet, F.; Manivet, P.; Pelta, J.; Behrends, J. C.; Aksimentiev, A.; Oukhaled, A. *Nat. Biotechnol.* **2020**, *38*, 176.
- [20] Xin, K.-L.; Hu, Z.-L.; Liu, S.-C.; Li, X.; Li, J.-G.; Niu, H.; Ying, Y.-L.; Long, Y.-T. *Angew. Chem., Int. Ed.* **2022**, *61*, e202209970.
- [21] Hu, R.; Rodrigues, J. V.; Waduge, P.; Yamazaki, H.; Cressiot, B.; Chishti, Y.; Makowski, L.; Yu, D.; Shakhnovich, E.; Zhao, Q. *ACS Nano* **2018**, *12*, 4494.
- [22] Liu, Y.; Zhang, S.; Wang, Y.; Wang, L.; Cao, Z.; Sun, W.; Fan, P.; Zhang, P.; Chen, H. Y.; Huang, S. *J. Am. Chem. Soc.* **2022**, *144*, 13717.
- [23] Li, S. J.; Li, X.; Wan, Y.-J.; Ying, Y.-L.; Yu, R.-J.; Long, Y.-T. *Chem. - Asian J.* **2023**, *18*, e202201144.
- [24] Liu, S.-C.; Li, M.-X.; Li, M.-Y.; Wang, Y.-Q.; Ying, Y.-L.; Wan, Y.-J.; Long, Y.-T. *Faraday Discuss.* **2018**, *210*, 87.
- [25] Li, X.; Ying, Y.-L.; Fu, X.-X.; Wan, Y.-J.; Long, Y.-T. *Angew. Chem., Int. Ed.* **2021**, *60*, 24582.
- [26] Desplanques, B.; Thienpondt, J.; Demuyne, K. In *21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2020)*, Shanghai, **2020**, pp. 3830~3834.
- [27] Morlet, J.; Arens, G.; Fourgeau, E.; Giard, D. *Geophysics* **1982**, *47*, 222.
- [28] Niu, H.; Li, M.-Y.; Ying, Y.-L.; Long, Y.-T. *Chem. Sci.* **2022**, *13*, 2456.
- [29] Cao, C.; Li, M.-Y.; Cirauqui, N.; Wang, Y.-Q.; Dal Peraro, M.; Tian, H.; Long, Y.-T. *Nat. Commun.* **2018**, *9*, 2823.
- [30] Fawcett, T. *Pattern Recognit. Lett.* **2006**, *27*, 861.
- [31] Van der Maaten, L.; Hinton, G. *J. Mach. Learn. Res.* **2008**, *9*, 2579.
- [32] Yang, C. Y.; Gu, Z.; Hu, Z. L.; Ying, Y. L.; Long, Y. T. *J. Electrochem.* **2019**, *25*, 312 (in Chinese). (杨铖宇, 顾震, 胡正利, 应佚伦, 龙亿涛, 电化学, **2019**, *25*, 312.)

(Cheng, F.)