Hans Journal of Biomedicine 生物医学, 2025, 15(3), 611-625
Published Online May 2025 in Hans. https://www.hanspub.org/journal/hjbm
https://doi.org/10.12677/hjbm.2025.153070

Hans汉斯

TCNLDA：基于自编码器和时序卷积网络预测lncRNA-疾病的关联

孟令宇，谭建军*

北京工业大学化学与生命科学学院，北京

收稿日期：2025年4月14日；录用日期：2025年5月22日；发布日期：2025年5月30日

摘　要

研究表明长非编码RNA (long non-coding RNA, lncRNA)影响着许多疾病的生物学过程，例如疾病的发生、传播、治愈等。因此，预测潜在lncRNA-疾病关联(lncRNA-disease associations, LDAs)对疾病的诊疗和治疗有着重要意义。本文提出一种新的深度学习方法预测LDAs，称为TCNLDA。首先，分别构建了lncRNA的功能相似性矩阵、高斯相似性矩阵和序列相似性矩阵，以及疾病的语义相似性矩阵和高斯相似性矩阵，并将其进行矩阵融合处理。然后，构建lncRNA-疾病对，并对其使用自编码器(Autoencoder, AE)进行特征提取。最后，将提取好的特征输入到时序卷积网络(Temporal Convolutional Network, TCN)中进行训练输出预测得分。两个数据集中，TCNLDA与其他模型进行了比较，结果显示TCNLDA优于其他LDAs预测方法。消融实验验证了TCNLDA中各部分的不可缺少性。案例研究进一步表明，TCNLDA在预测新型LDAs方面有着很好的实用性。

关键词

lncRNA-疾病关联，自编码器，时序卷积网络，lncRNA相似性，疾病相似性

TCNLDA: Prediction of lncRNA-Disease Associations Based on Autoencoder and Temporal Convolutional Network

Lingyu Meng, Jianjun Tan*

College of Chemistry and Life Science, Beijing University of Technology, Beijing

Received: Apr. 14th, 2025; accepted: May 22nd, 2025; published: May 30th, 2025

*通讯作者。

文章引用: 孟令宇, 谭建军. TCNLDA: 基于自编码器和时序卷积网络预测 lncRNA-疾病的关联[J]. 生物医学, 2025, 15(3): 611-625. DOI: 10.12677/hjbm.2025.153070

Abstract

Studies have shown that long non-coding RNA (lncRNA) influences the biological processes of many diseases, such as disease onset, spread, and cure. Therefore, predicting potential lncRNA-disease associations (LDAs) is important for disease diagnosis and treatment. In this paper, we propose a new deep learning method to predict LDAs, called TCNLDA. Firstly, the functional similarity matrix, Gaussian similarity matrix and sequence similarity matrix of lncRNAs, and the semantic similarity matrix and Gaussian similarity matrix of diseases are constructed and processed for matrix fusion, respectively. Then, the lncRNA-disease pairs are constructed and feature extraction is performed on them using autoencoder (AE). Finally, the extracted features were input into Temporal Convolutional Network (TCN) for training to output the prediction scores. In both datasets, TCNLDA was compared with other models, and the results showed that TCNLDA outperformed other LDAs prediction methods. Ablation experiments verified the indispensability of the components in TCNLDA. The case study further shows that TCNLDA has good utility in predicting novel LDAs.

Keywords

lncRNA-Disease Associations, Autoencoder, Temporal Convolutional Network, lncRNA Similarity, Disease Similarity

1. 引言

长非编码 RNA (long non-coding RNA, lncRNA)是一种长度超过 200 个核苷酸，且缺乏编码蛋白质能力的 RNA [1]。越来越多研究表明，lncRNA 参与很多生物学过程，例如细胞的增殖[2]、分化和凋亡[3]、基因表达与转录[4]、与蛋白质的相互作用[5]以及免疫应答[6]等。正是 lncRNA 在生命活动中的重要调控作用，使其与很多疾病的发生、传播和治愈有关。然而，生物实验的方法有耗时长和成本高的缺陷，这很大程度上限制了相关领域的研究进展。基于生物实验所积累的实验数据，研究者们提出使用计算方法来预测 lncRNA-疾病关联(lncRNA-disease associations, LDAs)，这些方法大致可以分为三类：基于生物网络的方法、基于矩阵分解(补全)的方法和基于机器学习的方法。

基于生物网络的计算方法是通过整合 lncRNA 和疾病的各种相似度和关联信息构建异构网络，并使用随机漫步和各种传播算法构建预测模型，揭示 lncRNA 和疾病的潜在关联[7]。例如，RWRlncD [8]通过在 lncRNA 功能相似网络上进行重启随机游走来检测潜在 LDAs。LncRDNetFlow [9]利用 lncRNA 相似网络、蛋白质相互作用网络和疾病相似网络之间的关系，并采用流传播算法推断 LDAs。然而，基于生物网络的计算方法一般不适用于没有已知关联的 lncRNA 和疾病。

基于矩阵分解(补全)的方法通过将矩阵分解为低秩矩阵，矩阵分解提供了一个降维和矩阵补全的框架[10]，可以应用于 LDAs 预测。LDCMFC [11]使用相关熵的协同矩阵分解来识别 LDAs。GMCLDA [12]采用基于几何矩阵互补的计算方法来推断潜在 LDAs。然而，基于矩阵分解(补全)的方法仅仅使用简单的线性建模，无法捕捉疾病与 lncRNA 之间的复杂关系。

在 LDAs 预测领域，机器学习方法是近些年最热门的方法之一[13] [14]。机器学习方法是将 lncRNA

和疾病的各种特征进行学习训练，最终预测潜在 LDAs。机器学习被分为两大类：传统机器学习和深度学习。LRLSLDA [15]发现了一种基于拉普拉斯正则化最小二乘法的半监督传统机器学习框架，其被用以预测 LDAs。随机森林(RF)是一种典型的传统机器学习模型，RFLDA [16]开发出一种结合 RF 和特征选择的 LDAs 预测框架。与传统机器学习方法相比，深度学习算法能够更好地适应复杂的数据分布和特征之间的非线性关系。CNNLDA [17]是一种基于注意力机制的双卷积神经网络，用以预测与疾病相关的潜在 lncRNA。GCRFLDA [18]提出了一种基于具有条件随机场的图卷积矩阵补全的 LDAs 预测方法。

尽管上述的基于机器学习的方法已经取得了不错的成绩，但仍有些改进的空间。传统卷积神经网络(CNN)在图像领域的成功启发了研究者将其扩展到时间序列领域，但传统 CNN 缺乏对时序因果关系的建模能力。因此 2018 年 Bai 等人提出了时序卷积网络(Temporal Convolutional Network, TCN)，TCN 通过因果卷积、扩张卷积和残差连接，解决了传统 RNN 的并行性差和长期依赖问题，同时继承了 CNN 的高效计算优势，成为时间序列建模的强有力工具。TCN 结合了 CNN 的并行处理能力和循环神经网络(RNN)的长期依赖建模能力，成为一种专门用于处理序列数据的深度学习模型。

因此，本文提出了一种基于自编码器和时序卷积网络的 LDAs 预测模型 TCNLDA。具体来说，首先使用了三种 lncRNA 相似性和两种疾病相似性中的信息数据，并使用矩阵融合的方法将其融合为 lncRNA 相似性矩阵和疾病相似性矩阵。其次，TCNLDA 将构建好的 lncRNA-疾病对利用自编码器进行数据降维，以获取特征的低维表示。最后，TCNLDA 采用时序卷积网络框架学习特征并完成最终的预测。在两个数据集上，使用多个评价指标对预测 LDAs 的结果进行可视化，相比于其他几个模型，TCNLDA 拥有更优越的性能。案例研究进一步说明 TCNLDA 是一个有前景的预测模型。

2. 材料与方法

2.1. 数据集

在本研究中使用两个数据集对 TCNLDA 进行了评估：

数据集 1 来自于 Li 的研究[19]的基准数据集。它包含 861 个 lncRNA，253 种疾病，495 个 miRNA，4517 个 LDAs 来自 Lnc2Cancer v2.0 [20]和 LncRNADisease [21]，831 个 lncRNA-miRNA 关联对来自 starBase v2.0 [22]，11,486 个 miRNA-疾病关联对来自 HMDD v2.0 [23]。

数据集 2 是我们自己集成的数据集。它包含 707 个 lncRNA，269 种疾病，252 个 miRNA，8101 个 LDAs 来自 Lnc2cancer v3.0 [24]和 LncRNADisease v2.0 [25]，来自 starBase v2.0 的 1942 个 lncRNA-miRNA 关联，来自 HMDD v3.2 [26]的 9825 个 miRNA-疾病关联。

随着近些年各个版本的公共数据库更新，数据集 2 拥有更加全面的 LDAs。同时，为了验证模型的泛化能力以及避免过拟合，将保证数据集 2 中数据集 1 的 lncRNA 和 LDAs 的数据重叠率低于 20%。两个数据集中所有已知的关联对被作为阳性样本，其余的作为阴性样本。

2.2. 模型架构

如图 1 所示，本文介绍了一个新的框架，名为 TCNLDA，将其用于预测 lncRNA-疾病的关联。首先，分别从三个方面构建 lncRNA 相似性矩阵和从两个方面构建疾病相似性矩阵，并将它们进行了矩阵融合处理。将融合后的 lnRNA 相似性矩阵和疾病相似性矩阵结合 lncRNA-miRNA 相互作用矩阵、miRNA-疾病关联矩阵共同构建为 lncRNA-疾病对。然后，使用自编码器进行特征的提取，用以获得 lncRNA-疾病的低维特征表示，并将数据转化为序列特征矩阵。最后，使用时序卷积网络进行特征学习，并输出分类预测结果。

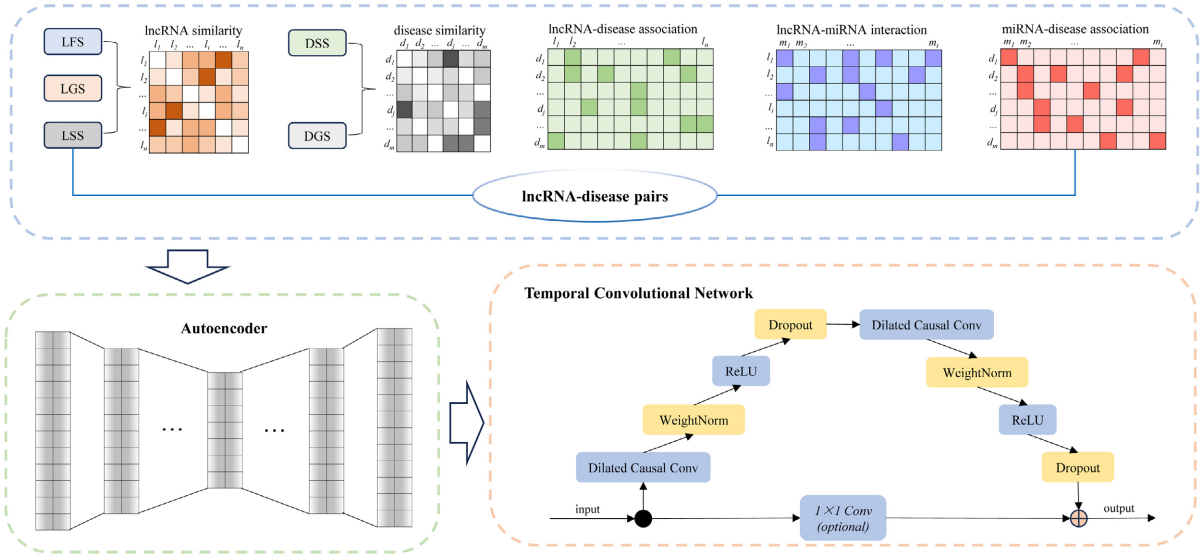

Figure 1. The flowchart of TCNLDA
图 1. TCNLDA 的流程图

2.3. 构建 lncRNA 疾病对

2.3.1. 疾病相似性

Wang 等[27]的方法被采用计算疾病语义相似性(disease semantic similarity, DSS)。DSS 的构建分为四部分：1) 从国家医学图书馆(https://www.nlm.nih.gov/)下载了疾病的医学主题标题(MeSH)。基于获得的 MeSH 信息，为每个疾病构建有向无环图(DAGs)。2) 构建每个疾病 d_i 对 $D(d_i)$ 语义贡献，如公式 1。DAGs 被用于计算疾病的语义相似性。疾病 d_i 可以被描述为 $\mathrm{DAG}(d_i)=(d_i, D(d_i))$，其中 $D(d_i)$ 是 d_i 及其所有祖先节点的节点集。3) 计算疾病 d_i 的最终贡献，如公式 2。4) 计算 d_i 和 d_j 的 DSS，如公式 3。DSS 计算方法如下：

$$\begin{cases} \mathrm{DS}_{d_i}(t)=1 & \text{if } t=d_i \\ \mathrm{DS}_{d_i}(t)=\max\left\{\theta\times\mathrm{DS}_{d_i}(t')\middle| t'\in D(d_i)\right\} & \text{otherwise} \end{cases} \tag{1}$$

这里 θ 代表疾病节点间边缘的语义贡献衰减因子。θ 取值在 $0<\theta<1$，这里 θ 为 0.5。

$$\mathrm{DF}(d_i)=\sum_{t\in D(d_i)}\mathrm{DS}_{d_i}(t) \tag{2}$$

$$\mathrm{DSS}(d_i,d_j)=\frac{\sum_{t\in D(d_i)\cap D(d_j)}\left(\mathrm{DS}_{d_i}(t)+\mathrm{DS}_{d_j}(t)\right)}{\mathrm{DF}(d_i)+\mathrm{DF}(d_j)} \tag{3}$$

基于邻接矩阵 LD 计算了疾病高斯相互作用普核相似性(Gaussian interaction profile kernel similarity for diseases, DGS)。假设疾病有 m 个，则 DGS 为：

$$\mathrm{DGS}(d_i,d_j)=\exp\left(-\left\|\mathrm{LD}(:,i)-\mathrm{LD}(:,j)\right\|^2\xi_d\right) \tag{4}$$

$$\xi_d=m\Big/\left(\sum_{i=1}^{m}\left\|\mathrm{LD}(:,i)\right\|^2\right) \tag{5}$$

2.3.2. LncRNA 相似性

根据计算出的疾病语义相似性和 LDAs 方法计算 lncRNA 功能相似性(lncRNA functional similarity, LFS)。

假设与 lncRNA l_1 和 l_2 相关的疾病分别有 m 和 n 种，这些疾病分别表示为 $d_{1i}\left(1\le i\le m\right)$ 和 $d_{2j}\left(1\le j\le n\right)$。然后，$l_1$ 和 l_2 的 LFS 可以计算为：

$$\mathrm{LFS}\left(l_1,l_2\right)=\frac{\sum_{i=1}^{n}\max_{1\le j\le m}\left(\mathrm{DSS}\left(d_{1i},d_{2j}\right)\right)+\sum_{j=1}^{m}\max_{1\le i\le n}\left(\mathrm{DSS}\left(d_{2j},d_{1i}\right)\right)}{n+m} \tag{6}$$

与 DGS 类似，基于邻接矩阵 LD 计算了 lncRNA 高斯相互作用普核相似性(Gaussian interaction profile kernel similarity for lncRNAs, LGS)。假设 lncRNA 有 n 个，则 LGS 为：

$$\mathrm{LGS}\left(l_i,l_j\right)=\exp\left(-\left\|\mathrm{LD}\left(i,:\right)-\mathrm{LD}\left(j,:\right)\right\|^2\xi_l\right) \tag{7}$$

$$\xi_l=n\Big/\left(\sum_{i=1}^{n}\left\|\mathrm{LD}\left(i,:\right)\right\|^2\right) \tag{8}$$

Liang 等[28]利用 lncRNA 序列计算了 lncRNA 序列相似性(lncRNA sequence similarity, LSS)。我们从国家医学图书馆中下载了 lncRNA 的序列信息。假设 l_1 和 l_2 的序列长度分别为 S_1 和 S_2，l_1 和 l_2 的莱文斯坦距离为 $\mathrm{Dis}\left(l_1,l_2\right)$。则 LSS 为：

$$\mathrm{LSS}\left(l_1,l_2\right)=1-\frac{\mathrm{Dis}\left(l_1,l_2\right)}{S_1+S_2} \tag{9}$$

2.3.3. 疾病和 lncRNA 相似矩阵融合

Lu 等[29]基于 DSS 和 DGS 进行矩阵融合，得到疾病相似性矩阵(DS)，如公式 10。我们进一步基于 LFS、LGS 和 LSS 进行矩阵融合，得到 lncRNA 相似性矩阵(LS)，如公式 11。

$$\mathrm{DS}\left(d_i,d_j\right)=\frac{1}{2}\left[\mathrm{DSS}\left(d_i,d_j\right)+\mathrm{DGS}\left(d_i,d_j\right)\right] \tag{10}$$

$$\mathrm{LS}\left(l_i,l_j\right)=\max\left\{\mathrm{LFS}\left(l_i,l_j\right),\frac{1}{2}\left[\mathrm{LSS}\left(l_i,l_j\right)+\mathrm{LGS}\left(l_i,l_j\right)\right]\right\} \tag{11}$$

2.3.4. lncRNA-疾病对的构建

本文模仿 Xuan 等[30]构建嵌入矩阵的方法构建了 lncRNA-疾病对。在这里使用 l_i 和 d_j 为例说明 lncRNA-疾病对的构建过程，如图 2 所示。首先，如果 l_i 和 d_j 与数据库中的某个 lncRNA 同时具有相似性和相关性，则它们之间存在关联的可能性将会很高。lncRNA-疾病对 $P_{i,j}$ 第一部分的第一行是 l_i 与其他 lncRNA 的相似性，第二行是 d_j 与 lncRNA 的相关性。其次，如果 l_i 和 d_j 与数据库中的某个疾病同时具有相关性和相似性，则它们之间存在关联的可能性将会很高。$P_{i,j}$ 第二部分的第一行是 l_i 与疾病的相关性，

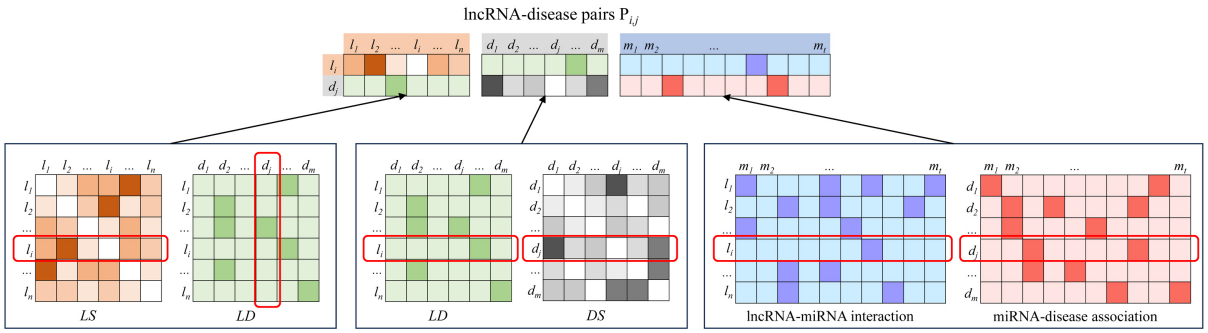

Figure 2. Construction of l_i-d_j pairs
图 2. l_i-d_j 的构建

第二行是 d_j 与其他疾病的相似性。最后，如果 l_i 和 d_j 与数据库中的某个 miRNA 同时相互作用和关联，则它们之间存在关联的可能性将会很高。$P_{i,j}$ 第三部分的第一行是 l_i 与 miRNA 的相互作用，第二行是 d_j 与 miRNA 的关联。至此，整合 lncRNA 相似性、疾病相似性、lncRNA-miRNA 相互作用和 miRNA-疾病关联，构建了节点 l_i - d_j 的 lncRNA-疾病对 $P_{i,j}$。用同样的方法可以构建其他 lncRNA-疾病对。

2.4. 自编码器

自编码器[31]是一种强大的神经网络架构，用于数据降维和特征提取。多层式自编码神经网络是一个由多个自编码器层组成的神经网络，其前一层自编码器的输出作为其后一层自编码器的输入。本文模型使用自编码器对构建好的 lncRNA-疾病对进行数据降维，提取数据的低维表示。

多层自编码器的基本思想是尝试将输入数据通过多个编码器映射到一个低维的隐藏表示，然后再通过与编码器对称的多个解码器将隐藏表示重构为原始数据。这个过程可以理解为一个数据的压缩和解压缩过程，其中隐藏表示被认为是数据的有价值特征。多层自编码器的编码器部分的目标是将输入数据 y 压缩为一个较小的隐藏表示。本文使用的编码器结构是多层神经网络，其中逐渐减少神经元的数量，使得网络逐渐捕捉到数据的主要特征。解码器部分的目标是将隐藏表示解码为重构数据 $\hat{y}$。解码器部分的结构与编码器相似，但神经元数量逐渐增加，最终生成与输入数据相匹配的输出。使用 MSELoss 作为损失函数，即：

$$\text{Loss}_1 = \frac{1}{n}\sum\left(\tilde{y}_i - y_i\right)^2 \tag{12}$$

通过多轮学习，得到 lncRNA-疾病对降维后的特征。

2.5. 时序卷积网络

时序卷积网络是一种专门用于处理序列数据的深度学习模型。它结合了卷积神经网络的并行处理能力和循环神经网络的长期依赖建模能力，成为处理序列特征任务中的理想工具。如图 1 所示，TCN 主要包括因果卷积、扩张卷积和残差连接三部分，可以让所有特征步的卷积操作可同步进行，显著提升训练速度，而且通过因果卷积和填充，输入与输出序列长度一致，非常适合序列标注与分类任务。TCN 在 LDAs 预测中的作用机制核心在于将静态生物数据转化为时序或序列化表示，将 lncRNA 的碱基序列视为“伪时间序列”，并通过其特有的因果卷积、扩张卷积和残差连接，捕捉局部与全局特征。

2.5.1. 因果卷积

因果卷积(Causal Convolution)的作用主要是确保模型不会违反序列顺序。本文中因果卷积的输出只依赖于当前 LDAs 及其之前的输入，而不依赖于后面输入的 LDAs。在标准的卷积操作中，每个输出值都基于其周围的输入值，包括未来的时间点。但在因果卷积中，权重仅应用于当前和过去的输入值，确保了信息流的方向性，避免了信息泄露到当前输出中。为了实现这一点，通常会在卷积核的右侧填充零，即因果填充，这样只有当前和过去的信息被用于计算输出。

对于输入序列 $X = \left[x_0, x_1, \cdots, x_{T-1}\right]$ 和卷积核 $W = \left[w_0, w_1, \cdots, w_{k-1}\right]$，因果卷积的输入 y_t 表示为：

$$y_t = \sum_{i=0}^{k-1} w_i \cdot x_{t-(k-1)+i} \tag{13}$$

其中 k 为卷积核大小。

2.5.2. 扩张卷积

扩张卷积(Dilated Convolution)的使用主要是为了增加感受野而不增加参数数量。扩张卷积，也被称

为空洞卷积，是一种在卷积核之间插入空隙的卷积形式，即跳过某些输入单元。这种技术允许模型在不增加参数数量的情况下捕获更大的感受野，从而更好地理解输入数据中的上下文信息。通过引入扩张率(dilation rate) d，扩大卷积核的感受野，捕捉长期依赖，其决定了卷积核中元素之间的间距。扩张卷积的输出 y_t 为：

$$y_t = \sum_{i=0}^{k-1} w_i \cdot x_{t-d \cdot i} \tag{14}$$

其中扩张率 d 按层级指数增长(如 d = 1, 2, 4, 8)。当 d = 1 时，退化为普通因果卷积。

TCN 的总感受野(Receptive Field, RF)为：

$$\mathrm{RF} = 1 + (k-1) \cdot \sum_{l=0}^{L-1} d_l \tag{15}$$

这里 L 为扩张卷积层数，d_l 为第 l 层的扩张率。

2.5.3. 残差连接

残差连接是残差网络(ResNets)的关键组成部分，它的主要目的是解决深层神经网络训练中的梯度消失或爆炸问题，以及提高网络的训练效率和性能。在残差连接中，网络的某一层的输出直接加到几层之后的另一层上，形成所谓的“跳跃连接”。本文 TCNLDA 中使用残差连接来缓解梯度消失问题并促进更深层网络的训练。具体来说，假设有一个输入 x，经过几层后得到 $F(x)$，那么最终的输出不是 $F(x)$而是 $x + F(x)$，也就是输入 + 输出。这种结构允许梯度在反向传播时可以直接流回更早的层，减少了梯度消失的问题，并且使得网络能够有效地训练更深的架构。残差块的输出可以表示为：

$$\mathrm{Output} = \mathrm{RuLU}\left(F(X) + X\right) \tag{16}$$

TCN 的基本结构包括多个残差块，每个残差块的内部流程包括：

$$H_1 = \mathrm{ReLU}\left(\mathrm{WeightNorm}\left(\mathrm{CausalDilatedConv}(X)\right)\right) \tag{17}$$

$$H_2 = \mathrm{WeightNorm}\left(\mathrm{CausalDilatedConv}(H_1)\right) \tag{18}$$

$$\mathrm{Output} = \mathrm{ReLU}\left(H_2 + X\right) \tag{19}$$

最后，对于输入序列 X，经过多个残差块后，得到输出 Y，取最后一个时间步的输出并通过 Softmax 进行分类：

$$P(y=c) = \frac{\exp\left(Y_{T-1,c}\right)}{\sum_{c'=1}^{C} \exp\left(Y_{T-1,c'}\right)} \tag{20}$$

3. 结果

3.1. 交叉验证和评估指标

在实验中，5 折交叉验证(5-CV)被用在评估 TCNLDA 的性能。我们将已知 LDAs 所得到的 lncRNA-疾病对作为阳性样本，未知的作为阴性样本。由于样本中负样本远远大于正样本，这会影响最后的预测概率，欠采样被用于平衡正负样本。我们将正样本分成五个子集，其中四个与等大小随机选择的阴性样本进行训练，其余一个与剩余的阴性样本中随机抽取等数量的阴性样本进行测试。

为了准确评估 TCNLDA 的性能，之前的很多研究采用 ROC 曲线和 PR 曲线，以及曲线下的面积(AUC 和 AUPR 值)作为评价指标。因此本文绘制了 ROC 曲线和 PR 曲线，并计算了 AUC 和 AUPR 值。除此之

外，还使用了其他六个评价指标进一步评估模型的预测 LDAs 的能力，它们分别是准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1 值(F1-score)、kappa 系数和马修斯相关系数(Matthews correlation coefficient, MCC)。这六个评价指标的计算公式如下：

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{21}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{22}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{23}$$

$$\text{F1-score} = \frac{2\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{24}$$

$$\text{kappa} = \frac{\text{Accuracy} - p}{1 - p} \tag{25}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{26}$$

其中 TP 和 TN 分别表示正确预测阳性和阴性样本的数量，FP 和 FN 分别表示错误预测阳性和阴性样本的数量。

3.2. 参数选择

在这里展示了在数据集 1 的实验中的六个重要超参数的调整结果，它们分别是：自编码器的层数(表示为 *m*)，时序卷积网络中扩张卷积的扩张率(表示为 *d*)，批次大小(表示为 *bs*)和学习率(表示为 *lr*)。*m* 从 3，5，7 中选择(当 *m* = 3 时，自编码器只有输入层、输出层和重构层，没有中间层)，*d* 在 2，4，8 中选择，*bs* 在 32，64，128 中选择，*lr* 在 0.002，0.001，0.0005 中选择。如表 1 所示，当 *m*、*d*、*bs* 和 *lr* 分别设置为 5，4，128，0.001 时，TCNLDA 在 5-CV 中可以获得本数据集的最佳 AUC 和 AUPR。

Table 1. Performance of TCNLDA using different values of *m*, *d*, *bs* and *lr* on Dataset 1
表 1. TCNLDA 在数据集 1 上使用 *m*，*d*，*bs* 和 *lr* 的不同值时的性能

参数	参数值	AUC	AUPR
m	3	0.9323	0.9425
	5	**0.9762**	**0.9773**
	7	0.9147	0.9022
d	2	0.9665	0.9723
	4	**0.9762**	**0.9773**
	8	0.9548	0.9572
bs	32	0.9732	0.9769
	64	**0.9762**	**0.9773**
	128	0.9643	0.9721
lr	0.002	0.9540	0.9327
	0.001	**0.9762**	**0.9773**
	0.0005	0.9637	0.9705

3.3. 与其他模型比较

为了证明 TCNLDA 的优越性能，在两个数据集中将其与以下四种方法进行了比较：SIMCLDA (2018) [32]、IPCARF (2021) [33]、VGAELDA (2021) [34]、gGATLDA (2022) [35]。其中，SIMCLDA 和 IPCARF 使用了传统机器学习方法，gGATLDA 和 VGAELDA 使用的是深度学习预测方法。VGAELDA 与本文方法的中间介质不同，其使用基因作为中间介质。为了比较的公平性，将上述模型在本文所用的两个数据集上进行训练和测试，并采用 5-CV 的方法输出预测结果。同时，训练和测试中所使用的超参数皆为原模型的默认参数。

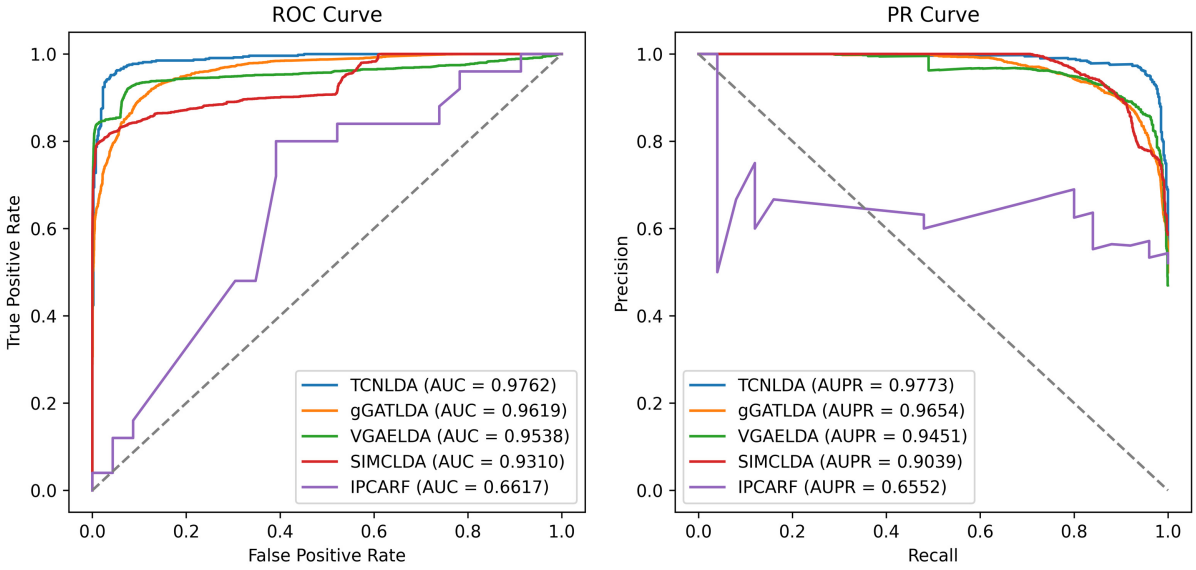

Figure 3. The ROC and PR curves of TCNLDA compared to other benchmark models on Dataset 1
图 3. 在数据集 1 上，TCNLDA 与其他基准模型的 ROC 和 PR 曲线对比

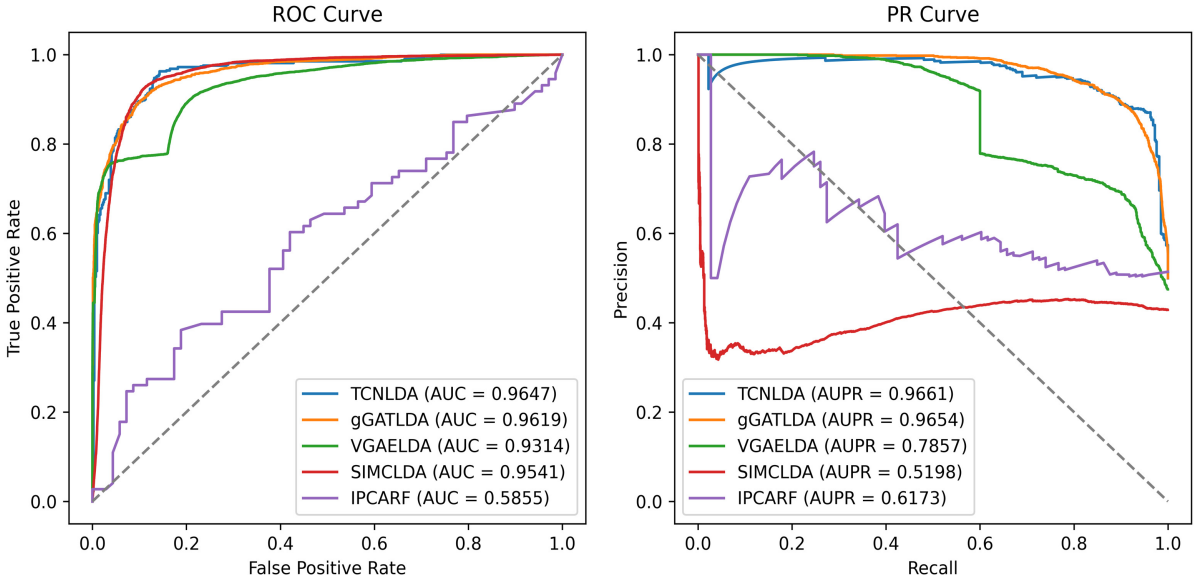

Figure 4. The ROC and PR curves of TCNLDA compared to other benchmark models on Dataset 2
图 4. 在数据集 2 上，TCNLDA 与其他基准模型的 ROC 和 PR 曲线对比

如图 3，图 4 所示，相比于其他五个模型，在两个数据集中 TCNLDA 的 AUC 和 AUPR 值皆是最优。在数据集 1 中，TCNLDA 的 AUC 值为 0.9762，比 gGATLDA、VGAELDA、SIMCLDA 和 IPCARF 分别高出 1.38%、2.35%、4.85%、47.53%。TCNLDA 的 AUPR 值为 0.9773，比 gGATLDA、VGAELDA、SIMCLDA 和 IPCARF 分别高出 1.23%、3.41%、8.12%、49.16%。在数据集 2 中，TCNLDA 的 AUC 值为 0.9647，比 gGATLDA、VGAELDA、SIMCLDA 和 IPCARF 分别高出 0.29%、3.58%、1.11%、64.77%。TCNLDA 的 AUPR 值为 0.9661，比 gGATLDA、VGAELDA、SIMCLDA 和 IPCARF 分别高出 0.07%、22.96%、85.86%、56.50%。本文模型较高的 AUC 和 AUPR，说明 TCNLDA 在排序能力、鲁棒性以及正例预测精确率上要优于其他五个模型。然后，在 Precision, Recall, F1-score, Accuracy, Kappa coefficient, and MCC 这六个评价指标上，将 TCNLDA 与其他五个模型进行了对比，如图 5，图 6。可以看出，TCNLDA 在数据集 1 上的 Precision，Recall，F1-score，Kappa coefficient，MCC 等评价指标基本优于其他五个模型，仅有 Recall 略逊色于模型 VGAELDA。然而，TCNLDA 在数据集 2 上各评价指标全部都优于其他五个模型。这表明，TCNLDA 在整体正确率、正例准确率和覆盖率、平衡性、一致性以及全面综合性能等各方面要优于其他模型。因此可以证明 TCNLDA 在预测 LDAs 的方面具有很强的竞争力。

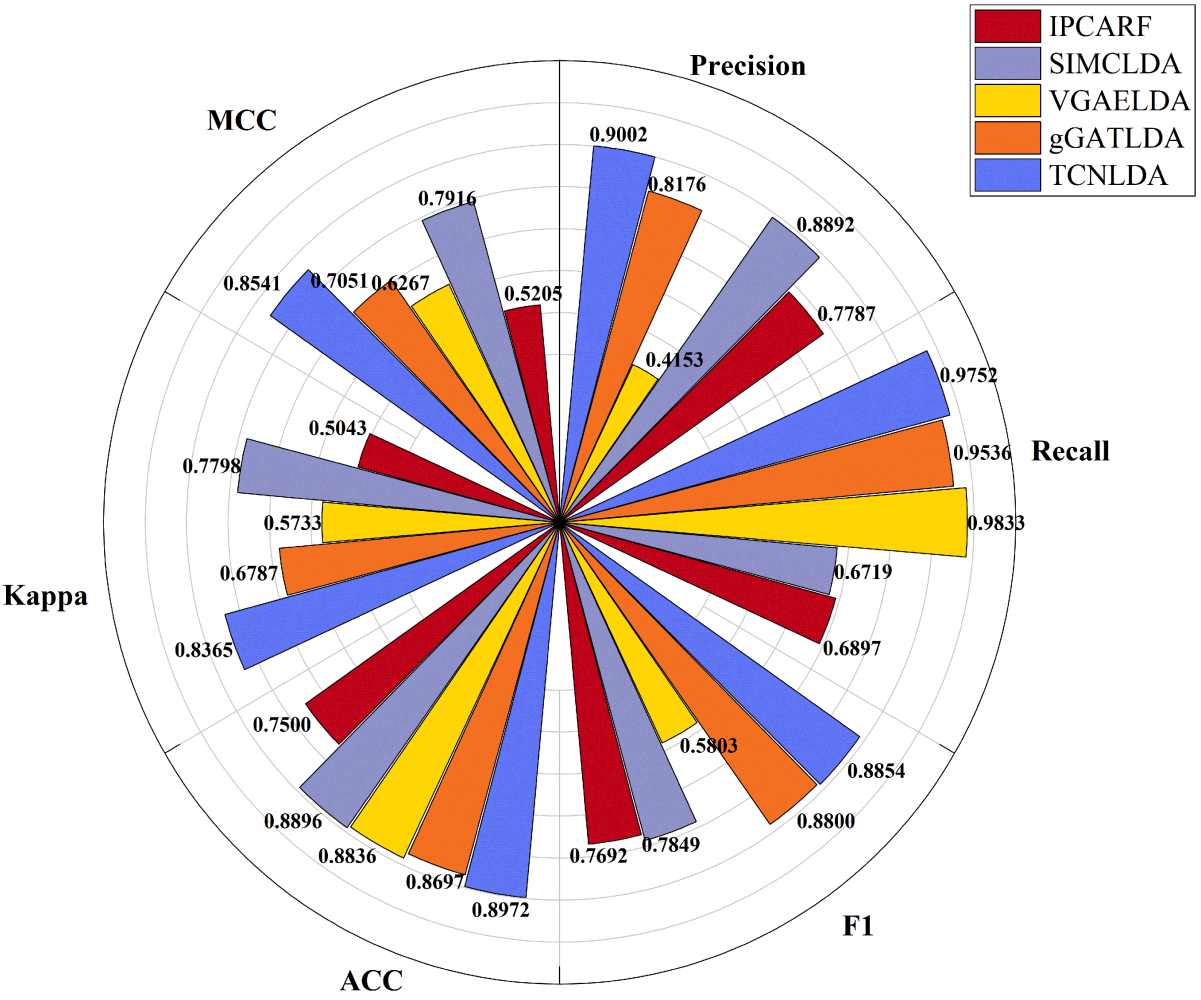

Figure 5. Comparisons of precision, recall, F1-score, accuracy, Kappa coefficient, and MCC among various models on Dataset 1
图 5. 数据集 1 上各种模型的精度、召回率、F1 分数、准确率、Kappa 系数和 MCC 的比较

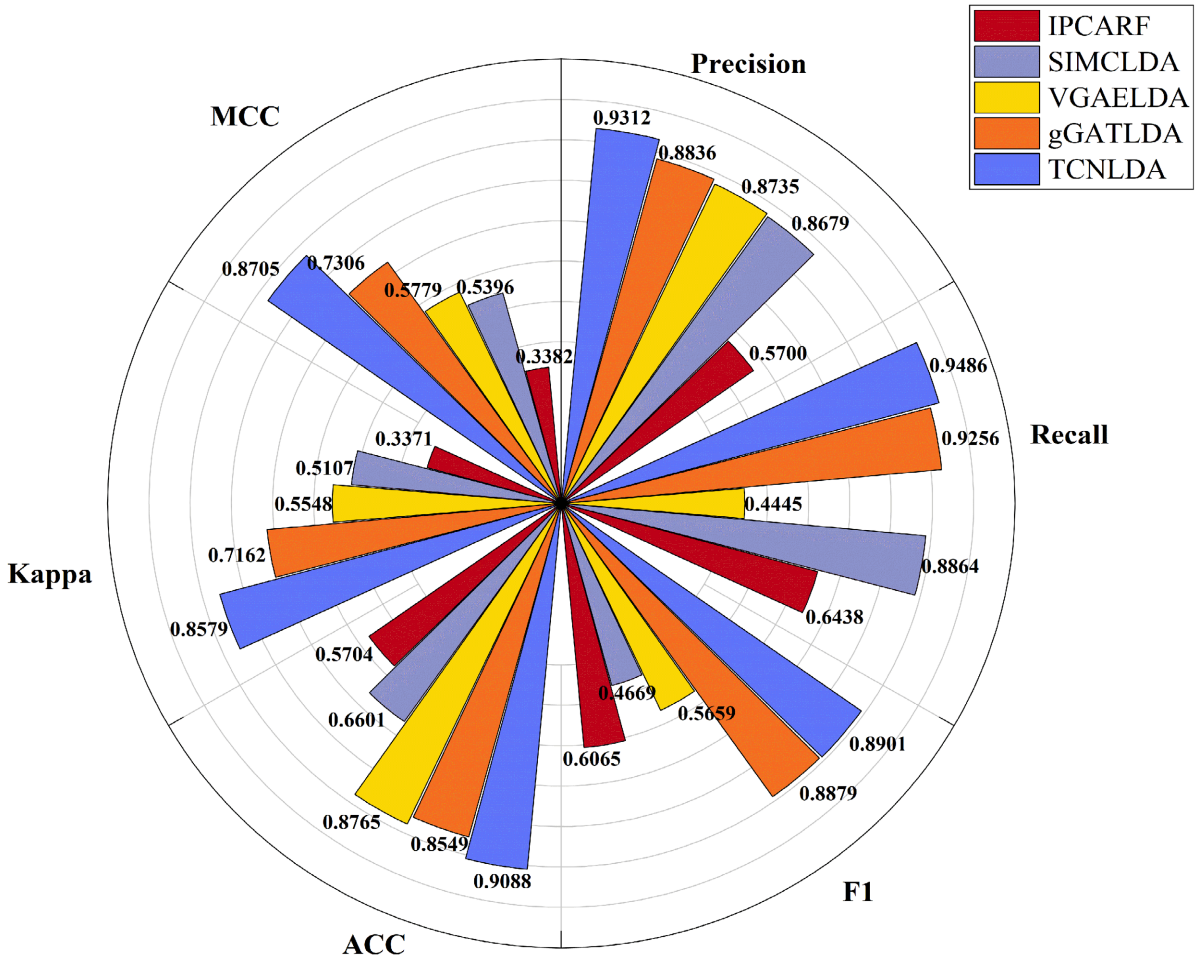

Figure 6. Comparisons of precision, recall, F1-score, accuracy, Kappa coefficient, and MCC among various models on Dataset 2
图 6. 数据集 2 上各种模型的精度、召回率、F1 分数、准确率、Kappa 系数和 MCC 的比较

3.4. 消融实验

消融实验经常被用于在 LDAs 预测中验证模型各部分的存在有效性。在本实验中，首先要验证自编码器、TCN 加入的 Dropout 层和时序卷积网络本身的有效性。因此，将原始模型与其去掉自编码器(TCNLDA-noAE)，去掉 Dropout 层(TCNLDA-noD)和 TCN 变成卷积神经网络(TCNLDA-CNN)后的变体模型进行了比较。在两个数据集中，比较了模型所得出的 AUC 值和模型训练至收敛所需要的时间，如表 2 所示。在表中可以看出 TCNLDA 在两个数据集上的 AUC 都优于其它三个变体。虽然 TCNLDA 和变体 TCNLDA-noAE 的各项指标差别不大，但其所耗时间远远小于去掉自编码器的变体模型，这说明自编码器可以对数据进行降维，节省训练时间和计算成本。Dropout 层可以在训练过程中防止模型过拟合。此外，虽然变体 TCNLDA-CNN 训练所耗时间最少，但是 CNN 模型过于简单，无法学习到 LDAs 中的复杂特征，因此其 AUC 值很差，而相比于 CNN，TCN 拥有更好的预测性能和鲁棒性。

3.5. 案例分析

为了进一步检验 TCNLDA 在特定疾病的新型 LDAs 方面的预测能力，我们在数据集 1 上进行了胃癌(GC)和肺癌(LC)的案例研究，具体步骤如下：1) 将数据集中所有已知的 LDAs 作为阳性样本，随机抽取

Table 2. The impact of AE, BN layers and dropout layers on model performance and computational time on Dataset 1 and Dataset 2
表 2. 数据集 1 和数据集 2 上的 AE 层、BN 层和剔除层对模型性能和计算时间的影响

Model \ Dataset	数据集 1		数据集 2	
	AUC	时间(分钟)	AUC	时间(分钟)
TCNLDA	0.9762	11.28	0.9647	19.49
TCNLDA-noAE	0.9699	463.73	0.9652	728.25
TCNLDA-noD	0.9490	11.15	0.9056	18.91
TCNLDA-CNN	0.7239	2.14	0.5961	4.59

等量的未知 LDAs 作为阴性样本，将阳性样本和阴性样本合并用于模型的训练。2) 将所有未知的 lncRNA 与两种特定疾病之间的关联分别用于模型测试。3) 使用 TCNLDA 对测试样本输出预测得分，然后进行排序，并选取与这些疾病相关的排名前 10 位的 lncRNA。然后通过 PubMed 文献检索实验证据，每种疾病的具体分析如表 3 所示。

Table 3. Top 10 TCNLDA predicted lncRNAs associated with GC and LC on Dataset 1
表 3. 数据集 1 中与 GC 和 LC 相关的前 10 个 TCNLDA 预测 lncRNA

疾病	序号	LncRNA	证明(PMID)
Gastric cancer (GC)	1	MALAT1	32104001
	2	NEAT1	未证实
	3	LINC01133	30134915
	4	ERICH1-AS1	未证实
	5	HOTAIR	30810117
	6	CCAT2	29435046
	7	UCA1	29805620
	8	LSINCT5	30127643
	9	NPTN-IT1	25674261
	10	MIR124-2HG	未证实
Lung cancer (LC)	1	OIP5-AS1	29897167
	2	SNHG12	30719111
	3	MALAT1	31133357
	4	CCAT2	30214594
	5	KCNQ1OT1	未证实
	6	MIR155HG	32432745
	7	HOTAIR	32248643
	8	LINC01133	26840083
	9	GAS5	30926767
	10	XIST	29812958

胃癌是全球第五大常见癌症，也是癌症死亡的第三大常见原因[36]。根据 TCNLDA 预测的与胃癌相关的前 10 个 lncRNA 中有 7 个得到确认。例如，沉默 CCAT2 基因可以抑制胃癌 BGC-823 细胞的增殖，以及诱导胃癌 BGC-823 细胞的凋亡和自噬[37]。LSINCT5 的激活将影响 GC 细胞迁移和侵袭，其可能成为新的 GC 疗法的靶点[38]。MALAT1 可以调节奥沙利铂对胃癌的耐药性[39]。

肺癌是世界上癌症死亡的主要原因之一[40]，其包括小细胞肺癌(SCLC)和非小细胞肺癌(NSCLC)。模型预测与肺癌相关的前 10 个 lncRNA 中有 9 个得到文献证实。例如，GAS5 的上调可以抑制 NSCLC 的生长、迁移和侵袭[41]。HOTAIR 表达水平的变化会影响 NSCLC 细胞的迁移和侵袭能力[42]。OIP5-AS1 基因的反义是一种 lncRNA，其在肺癌组织中高表达与肿瘤大小和肿瘤生长速度相关[43]。

上述预测结果可以进一步指导那些未被证实过的 LDAs 在生物医学实验中进行验证，使实验方向更加明确，减少不必要的成本损耗。

4. 总结与讨论

随着生物医学领域的不断发展，越来越多人发现 lncRNA 在许多疾病的发病和治疗等生物学过程中起到重要作用。同时，miRNA 作为 lncRNA 与疾病关联的中间介质也不断被验证。本文提出一种基于自编码器和时序卷积网络的 LDAs 预测模型 TCNLDA。首先，TCNLDA 使用矩阵融合的方法将多种 lncRNA 和疾病相似性矩阵分别融合为 lncRNA 相似性矩阵和疾病相似性矩阵，它们与 lncRNA-疾病关联矩阵、lncRNA-miRNA 相似性矩阵和 miRNA-疾病关联矩阵共同构建了 lncRNA-疾病对。然后，自编码器用于特征提取，并将提取好的特征输入时序卷积网络中进行训练并输出预测得分。多个实验结果表明，TCNLDA 相比于其他基准模型和近些年先进模型拥有更优越的性能，其各部分都在模型中都具有重要作用，并且能够很好地预测出新的 LDAs。

然而，本文模型中还存在一些问题需要进一步研究。首先是数据正负样本不平衡的问题，和 LDAs 预测相关领域的许多研究一样采用了欠采样以平衡正负样本。但是欠采样的方法必然会舍弃一部分数量过多的负样本，因此会丢失一些负样本所包含的特征，对模型的最终预测结果产生一定程度的误差。此外，本模型验证的数据集仅包含很小一部分现实中的 lncRNA 和疾病，并且数据量的不断扩增是未来 LDAs 预测领域的趋势。因此，在未来的研究中，我们将会尝试整合出包含更多 lncRNA 和疾病的数据集进行 LDAs 预测研究，同时尝试探索新的模型结构，使之可以在不平衡的样本中得到较高的预测性能。

参考文献

[1] Nagano, T. and Fraser, P. (2011) No-Nonsense Functions for Long Noncoding RNAs. *Cell*, **145**, 178-181. https://doi.org/10.1016/j.cell.2011.03.014

[2] Zhao, T. (2019) Long Noncoding RNA and Its Role in Virus Infection and Pathogenesis. *Frontiers in Bioscience*, **24**, 777-789. https://doi.org/10.2741/4750

[3] Tüncel, Ö., Kara, M., Yaylak, B., Erdoğan, İ. and Akgül, B. (2022) Noncoding RNAs in Apoptosis: Identification and *Turkish Journal of Biology*, **46**, 1-40.

[4] Chen, J., Ao, L. and Yang, J. (2019) Long Non-Coding RNAs in Diseases Related to Inflammation and Immunity. *Annals of Translational Medicine*, **7**, 494-494. https://doi.org/10.21037/atm.2019.08.37

[5] Yan, J., Qu, W., Li, X., Wang, R. and Tan, J. (2024) GATLGEMF: A Graph Attention Model with Line Graph Embedding Multi-Complex Features for ncRNA-Protein Interactions Prediction. *Computational Biology and Chemistry*, **108**, Article 108000. https://doi.org/10.1016/j.compbiolchem.2023.108000

[6] Wang, F., Lin, H., Su, Q. and Li, C. (2022) Cuproptosis-Related lncRNA Predict Prognosis and Immune Response of Lung Adenocarcinoma. *World Journal of Surgical Oncology*, **20**, Article No. 275. https://doi.org/10.1186/s12957-022-02727-7

[7] Yan, J., Wang, R. and Tan, J. (2023) Recent Advances in Predicting lncRNA-Disease Associations Based on Computational Methods. *Drug Discovery Today*, **28**, Article 103432. https://doi.org/10.1016/j.drudis.2022.103432

[8] Sun, J., Shi, H., Wang, Z., Zhang, C., Liu, L., Wang, L., *et al*. (2014) Inferring Novel lncRNA-Disease Associations Based on a Random Walk Model of a lncRNA Functional Similarity Network. *Molecular BioSystems*, **10**, 2074-2081. https://doi.org/10.1039/c3mb70608g

[9] Zhang, J., Zhang, Z., Chen, Z. and Deng, L. (2019) Integrating Multiple Heterogeneous Networks for Novel lncRNA-Disease Association Inference. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **16**, 396-406. https://doi.org/10.1109/tcbb.2017.2701379

[10] Xi, J., Wang, M. and Li, A. (2017) Discovering Potential Driver Genes through an Integrated Model of Somatic Mutation Profiles and Gene Functional Information. *Molecular BioSystems*, **13**, 2135-2144. https://doi.org/10.1039/c7mb00303j

[11] Xi, W., Zhou, F., Gao, Y., Liu, J. and Zheng, C. (2023) LDCMFC: Predicting Long Non-Coding RNA and Disease Association Using Collaborative Matrix Factorization Based on Correntropy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **20**, 1774-1782. https://doi.org/10.1109/tcbb.2022.3215194

[12] Lu, C., Yang, M., Li, M., Li, Y., Wu, F. and Wang, J. (2020) Predicting Human lncRNA-Disease Associations Based on Geometric Matrix Completion. *IEEE Journal of Biomedical and Health Informatics*, **24**, 2420-2429. https://doi.org/10.1109/jbhi.2019.2958389

[13] Tan, J., Li, X., Zhang, L. and Du, Z. (2022) Recent Advances in Machine Learning Methods for Predicting lncRNA and Disease Associations. *Frontiers in Cellular and Infection Microbiology*, **12**, Article 1071972. https://doi.org/10.3389/fcimb.2022.1071972

[14] Sheng, N., Huang, L., Lu, Y., Wang, H., Yang, L., Gao, L., *et al*. (2023) Data Resources and Computational Methods for lncRNA-Disease Association Prediction. *Computers in Biology and Medicine*, **153**, Article 106527. https://doi.org/10.1016/j.compbiomed.2022.106527

[15] Chen, X. and Yan, G. (2013) Novel Human lncRNA-Disease Association Inference Based on lncRNA Expression Profiles. *Bioinformatics*, **29**, 2617-2624. https://doi.org/10.1093/bioinformatics/btt426

[16] Yao, D., Zhan, X., Zhan, X., Kwoh, C.K., Li, P. and Wang, J. (2020) A Random Forest Based Computational Model for Predicting Novel lncRNA-Disease Associations. *BMC Bioinformatics*, **21**, Article No. 126. https://doi.org/10.1186/s12859-020-3458-1

[17] Xuan, P., Cao, Y., Zhang, T., Kong, R. and Zhang, Z. (2019) Dual Convolutional Neural Networks with Attention Mechanisms Based Method for Predicting Disease-Related lncRNA Genes. *Frontiers in Genetics*, **10**, Article 416. https://doi.org/10.3389/fgene.2019.00416

[18] Fan, Y., Chen, M. and Pan, X. (2021) GCRFLDA: Scoring lncRNA-Disease Associations Using Graph Convolution Matrix Completion with Conditional Random Field. *Briefings in Bioinformatics*, **23**, bbab361. https://doi.org/10.1093/bib/bbab361

[19] Fu, G., Wang, J., Domeniconi, C. and Yu, G. (2017) Matrix Factorization-Based Data Fusion for the Prediction of lncRNA-Disease Associations. *Bioinformatics*, **34**, 1529-1537. https://doi.org/10.1093/bioinformatics/btx794

[20] Ning, S., Zhang, J., Wang, P., Zhi, H., Wang, J., Liu, Y., *et al*. (2015) Lnc2Cancer: A Manually Curated Database of Experimentally Supported lncRNAs Associated with Various Human Cancers. *Nucleic Acids Research*, **44**, D980-D985. https://doi.org/10.1093/nar/gkv1094

[21] Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., *et al*. (2012) LncRNADisease: A Database for Long-Non-Coding RNA-Associated Diseases. *Nucleic Acids Research*, **41**, D983-D986. https://doi.org/10.1093/nar/gks1099

[22] Li, J., Liu, S., Zhou, H., Qu, L. and Yang, J. (2013) Starbase V2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and Protein-RNA Interaction Networks from Large-Scale CLIP-Seq Data. *Nucleic Acids Research*, **42**, D92-D97. https://doi.org/10.1093/nar/gkt1248

[23] Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., *et al*. (2013) HMDD V2.0: A Database for Experimentally Supported Human MicroRNA and Disease Associations. *Nucleic Acids Research*, **42**, D1070-D1074. https://doi.org/10.1093/nar/gkt1023

[24] Gao, Y., Shang, S., Guo, S., Li, X., Zhou, H., Liu, H., *et al*. (2020) Lnc2Cancer 3.0: An Updated Resource for Experimentally Supported lncRNA/circRNA Cancer Associations and Web Tools Based on RNA-Seq and scRNA-Seq Data. *Nucleic Acids Research*, **49**, D1251-D1258. https://doi.org/10.1093/nar/gkaa1006

[25] Lin, X., Lu, Y., Zhang, C., Cui, Q., Tang, Y., Ji, X., *et al*. (2023) LncRNADisease V3.0: An Updated Database of Long Non-Coding RNA-Associated Diseases. *Nucleic Acids Research*, **52**, D1365-D1369. https://doi.org/10.1093/nar/gkad828

[26] Cui, C., Zhong, B., Fan, R. and Cui, Q. (2023) HMDD V4.0: A Database for Experimentally Supported Human MicroRNA-Disease Associations. *Nucleic Acids Research*, **52**, D1327-D1332. https://doi.org/10.1093/nar/gkad717

[27] Wang, D., Wang, J., Lu, M., Song, F. and Cui, Q. (2010) Inferring the Human MicroRNA Functional Similarity and Functional Network Based on MicroRNA-Associated Diseases. *Bioinformatics*, **26**, 1644-1650. https://doi.org/10.1093/bioinformatics/btq241

[28] Liang, Y., Zhang, Z., Liu, N., Wu, Y., Gu, C. and Wang, Y. (2022) MAGCNSE: Predicting lncRNA-Disease Associations Using Multi-View Attention Graph Convolutional Network and Stacking Ensemble Model. *BMC Bioinformatics*, **23**, Article No. 189. https://doi.org/10.1186/s12859-022-04715-w

[29] Lu, C. and Xie, M. (2023) LDAEXC: LncRNA-Disease Associations Prediction with Deep Autoencoder and XGBoost Classifier. *Interdisciplinary Sciences*: *Computational Life Sciences*, **15**, 439-451. https://doi.org/10.1007/s12539-023-00573-z

[30] Xuan, P., Pan, S., Zhang, T., Liu, Y. and Sun, H. (2019) Graph Convolutional Network and Convolutional Neural Network Based Method for Predicting lncRNA-Disease Associations. *Cells*, **8**, Article 1012. https://doi.org/10.3390/cells8091012

[31] Shi, Y., Lei, M., Ma, R. and Niu, L. (2019) Learning Robust Auto-Encoders with Regularizer for Linearity and Sparsity. *IEEE Access*, **7**, 17195-17206. https://doi.org/10.1109/access.2019.2895884

[32] Lu, C., Yang, M., Luo, F., Wu, F., Li, M., Pan, Y., *et al*. (2018) Prediction of lncRNA-Disease Associations Based on Inductive Matrix Completion. *Bioinformatics*, **34**, 3357-3364. https://doi.org/10.1093/bioinformatics/bty327

[33] Zhu, R., Wang, Y., Liu, J. and Dai, L. (2021) IPCARF: Improving lncRNA-Disease Association Prediction Using Incremental Principal Component Analysis Feature Selection and a Random Forest Classifier. *BMC Bioinformatics*, **22**, Article No. 175. https://doi.org/10.1186/s12859-021-04104-9

[34] Shi, Z., Zhang, H., Jin, C., Quan, X. and Yin, Y. (2021) A Representation Learning Model Based on Variational Inference and Graph Autoencoder for Predicting lncRNA-Disease Associations. *BMC Bioinformatics*, **22**, Article No. 136. https://doi.org/10.1186/s12859-021-04073-z

[35] Wang, L. and Zhong, C. (2022) gGATLDA: LncRNA-Disease Association Prediction Based on Graph-Level Graph Attention Network. *BMC Bioinformatics*, **23**, Article No. 11. https://doi.org/10.1186/s12859-021-04548-z

[36] Smyth, E.C., Nilsson, M., Grabsch, H.I., van Grieken, N.C. and Lordick, F. (2020) Gastric Cancer. *The Lancet*, **396**, 635-648. https://doi.org/10.1016/s0140-6736(20)31288-5

[37] Yu, Z., Wang, Z., Lee, K., Yuan, P. and Ding, J. (2017) Effect of Silencing Colon Cancer-Associated Transcript 2 on the Proliferation, Apoptosis and Autophagy of Gastric Cancer BGC-823 Cells. *Oncology Letters*, **15**, 3127-3132. https://doi.org/10.3892/ol.2017.7677

[38] Qi, P., Lin, W., Zhang, M., Huang, D., Ni, S., Zhu, X., *et al*. (2018) E2F1 Induces LSINCT5 Transcriptional Activity and Promotes Gastric Cancer Progression by Affecting the Epithelial-Mesenchymal Transition. *Cancer Management and Research*, **10**, 2563-2571. https://doi.org/10.2147/cmar.s171652

[39] Feng, C., Zhao, Y., Li, Y., Zhang, T., Ma, Y. and Liu, Y. (2019) LncRNA MALAT1 Promotes Lung Cancer Proliferation and Gefitinib Resistance by Acting as a miR-200a Sponge. *Archivos de Bronconeumología* (*English Edition*), **55**, 627-633. https://doi.org/10.1016/j.arbr.2019.03.018

[40] Nasim, F., Sabath, B.F. and Eapen, G.A. (2019) Lung Cancer. *Medical Clinics of North America*, **103**, 463-473. https://doi.org/10.1016/j.mcna.2018.12.006

[41] Dong, L., Li, G., Li, Y. and Zhu, Z. (2019) Upregulation of Long Noncoding RNA GAS5 Inhibits Lung Cancer Cell Proliferation and Metastasis via miR-205/PTEN Axis. *Medical Science Monitor*, **25**, 2311-2319. https://doi.org/10.12659/msm.912581

[42] Zheng, F., Li, J., Ma, C., Tang, X., Tang, Q., Wu, J., *et al*. (2020) Novel Regulation of miR-34a-5p and HOTAIR by the Combination of Berberine and Gefitinib Leading to Inhibition of EMT in Human Lung Cancer. *Journal of Cellular and Molecular Medicine*, **24**, 5578-5592. https://doi.org/10.1111/jcmm.15214

[43] Wang, M., Sun, X., Yang, Y. and Jiao, W. (2018) Long Non-Coding RNA OIP5-AS1 Promotes Proliferation of Lung Cancer Cells and Leads to Poor Prognosis by Targeting miR-378a-3p. *Thoracic Cancer*, **9**, 939-949. https://doi.org/10.1111/1759-7714.12767