（Orthogonal experimental design，OED）在本研究中用来选择超参数组合。OED是一种研究多因素、多水平问题的设计方法。其中因素表示影响实验结果的各个变量，在本章节中表示超参数；水平表示每个因素的不同取值或处理条件，在本章节中表示不同超参数的取值。OED 基于正交性从整个测试集合中选取一些分布均匀的代表点进行测试。选择代表点的过程是通过构建正交表来实现的。正交表是一种特殊设计的表格，具有均匀分布和正交性的特点。通过正交表的选择，可以确保各个因素之间的相互独立性，从而消除因素间的相互干扰。应用正交试验的优点是进行试验次数少，效率高。通过汲取部分基于深度学习的研究中关于超参数设置的相关经验[22]，本章节选取了一些超参数进行调整并给出各超参数的一组估计值。这些估计值构成了正交表的整个测试点。成对独立组合测试（PICT）[52,53]工具被选定用来构建正交表，以获得多组超参数的代表性组合。与常见的优化超参数的随机选择和网格搜索方法不同，PICT 是一种用于软件测试领域的选择组合参数技术，用于减少系统测试用例输入的数量。神经网络中大量超参数的选择是 PICT 首选的应用场景。

3.2.5 集成模型

集成模型通过构造和组合多个学习器来完成学习任务[29]。与单个模型的分类性能相比，集成模型往往能取得更好的分类性能和泛化能力[54]。集成模型被用来减少模型的整体误差。集成模型包含多个学习器，每个学习器都是基于超参数的代表性组合训练的最优 DSE-ResNet 模型。

个体学习器在本文中被称为单一最优模型。集成模型使用投票策略来集成所有单一最优模型。每个单一最优模型都会对同一个测试样本给出一个预测值，基于少数服从多数的多模型投票策略，集成模型将得票最多的预测值作为最终输出值。集成模型的使用可以有效提高模型的分类性能和容错能力。

3.3 实验细节

3.3.1 实验设备环境

本章节所提出的模型使用 Keras 框架构建训练。所有实验均在配备 Quadro P2200 显卡和显存为 5G 的服务器上运行。软件环境采用 3.6 版本的 Python 编译器，使用 Pycharm 构建项目。

3.3.2 数据预处理

（1）去噪

原始信号中存在肌肉噪声、电流噪声和基线漂移，这些噪声可能影响模型的分类性能。为了去除这些噪声对十二导联信号的影响，使用 Butterworth[55]带通滤

波器滤除了频率为 0.5Hz 至 49Hz 之外的噪音信号，这个区间源自 CPSC2018[63] 比赛中各团队滤除噪音的大致区间。图 3.6 显示了用 Welch[56]方法计算异常样本导联 I 利用 Butterworth 带通滤波器滤波前后的功率谱密度曲线，根据曲线能够观察到高频噪声被衰减。功率谱密度曲线的可视化应用了不同窗口和不同窗口长度的 Welch 方法。

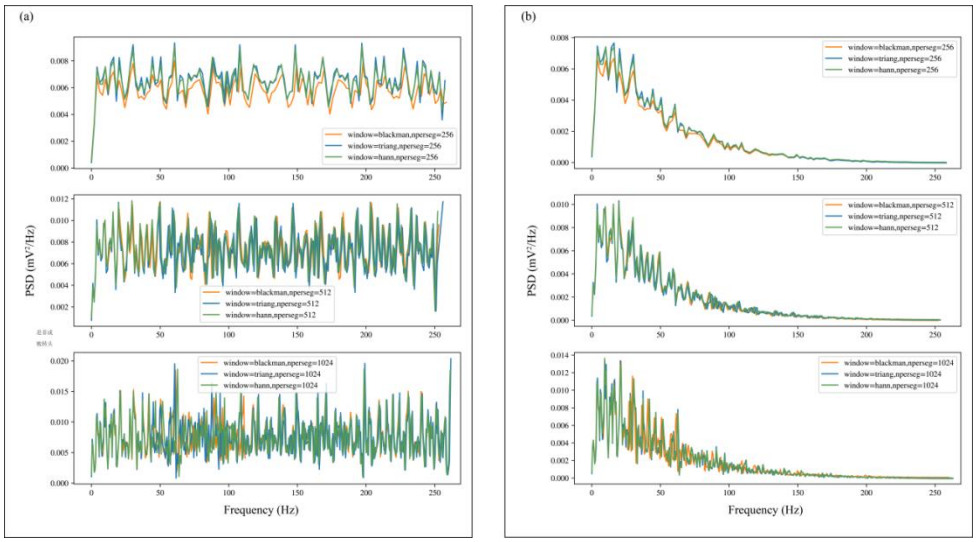

图 3.6 不同窗口和不同窗口长度应用 Welch 得到的功率谱密度曲线

注：（a）异常样本导联 I 信号滤波前的功率谱密度曲线；（b）异常样本导联 I 信号滤波后的功率谱密度曲线。每个子图使用相同的窗口长度和不同的窗口。window 表示窗函数，包括布莱克曼窗、汉宁窗和三角窗。nperseg 表示窗口长度包括 256、512 和 1024。

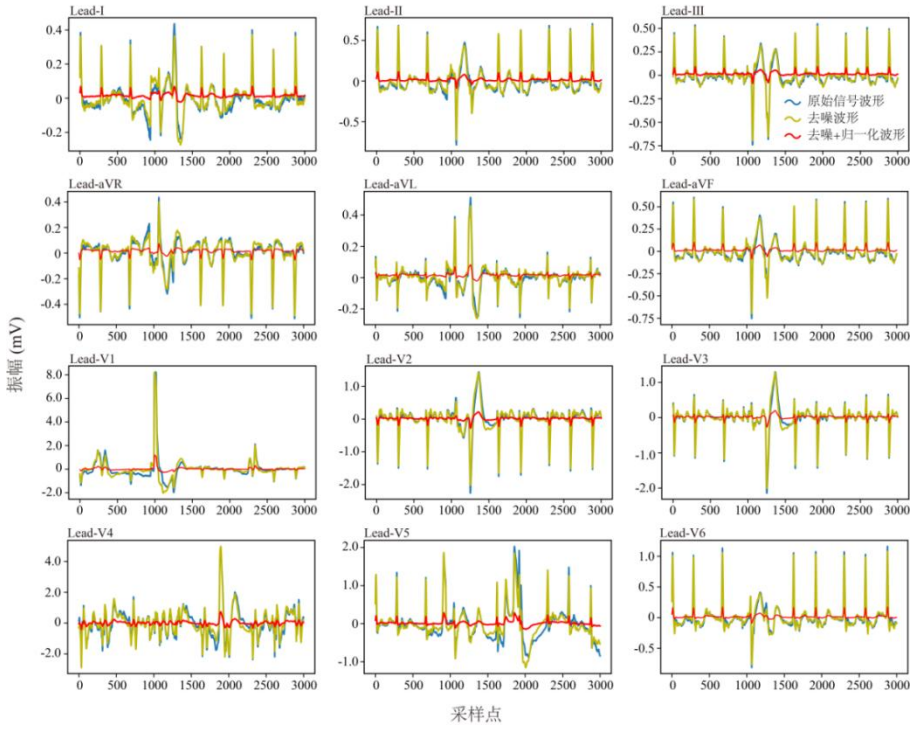

图 3.7 十二导联预处理对比波形图

表 3.3 集成模型与单一最优模型在小样本测试集上的F_1分数

No.	平均F_1	正常及 8 类心律失常									亚异常类型		
		正常	AF	I-AVB	LBBB	RBBB	PAC	PVC	STD	STE	Block	PC	ST
1	0.783	0.739	0.962	0.846	0.786	0.912	0.742	0.854	0.659	0.545	0.885	0.793	0.629
2	0.816	**0.821**	0.963	0.845	0.938	0.922	0.683	0.83	0.75	0.595	0.903	0.761	0.709
3	0.81	0.745	0.955	0.821	**0.97**	0.928	0.682	0.86	0.73	0.595	0.905	0.773	0.695
4	0.776	0.738	0.933	0.824	0.938	0.915	0.692	0.806	0.742	0.4	0.896	0.747	0.688
5	0.835	0.787	0.954	0.876	0.938	**0.941**	0.744	**0.907**	0.763	0.606	**0.926**	0.826	0.728
6	0.824	0.783	0.919	0.851	0.938	0.932	0.738	0.892	0.761	0.6	0.912	0.814	0.727
7	0.817	0.763	**0.969**	0.83	0.941	0.938	0.736	0.876	0.762	0.541	0.911	0.807	0.704
8	0.78	0.743	0.938	0.804	0.811	0.902	0.71	0.87	0.686	0.556	0.867	0.789	0.652
9	0.832	0.76	0.938	0.857	0.941	0.931	0.742	0.914	**0.779**	0.629	0.914	0.824	0.743
10	0.828	0.787	0.942	**0.901**	0.909	0.919	0.742	0.864	0.778	0.606	0.914	0.802	**0.738**
EM	**0.843**	0.787	0.949	0.87	**0.97**	0.935	**0.764**	0.897	0.748	**0.667**	0.922	**0.83**	0.729

3.4.3 在 CPSC2018 隐藏测试集的表现

图 3.8 展示了单一最优模型（learning rate=0.15，dropout=0.5，momentum=0.7）在训练集和验证集上的损失和准确率的变化曲线。验证集主要用来观察训练过程中模型在非训练集的损失和准确率曲线变化情况，通过观察曲线平稳性来控制模型训练的次数，防止出现过拟合现象。图 3.8 中模型损失和准确率曲线从在训练周期（epoch）为 30 次时开始趋于稳定。实验过程中尝试将 epoch 增加到 70 次发现过拟合现象，因此采用早停（early stopping）方法，将训练次数减少到 50 个训练周期。

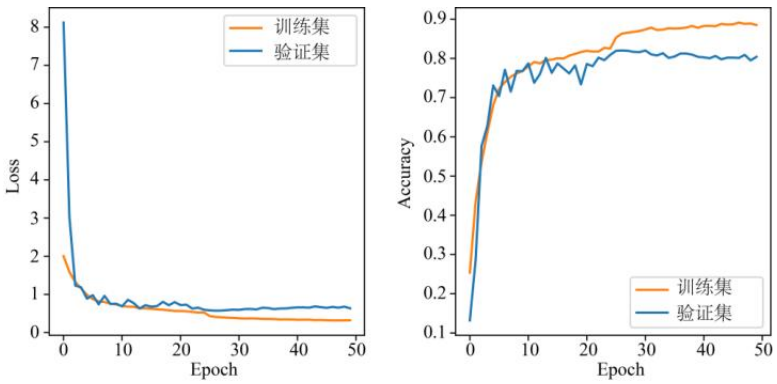

图 3.8 损失和准确率变化曲线

十组不同超参数组合的最优单一模型分别训练完成后，将这些模型组合形成集成模型并提交给 CPSC2018 官方工作人员，得到了基于隐藏测试集（2954 组样本）的测试结果。图 3.9 显示了集成模型在测试集得到的混淆矩阵。对于亚异常型 ST（STD 和 STE 的统称），根据混淆矩阵的结果显示 53 个标有 STD 标签的样本和 27 个标有 STE 标签的样本被预测为正常（Normal），19 个标有 Normal 标签的样本被预测为 STD。引起 ST 段改变的疾病在病理学角度不局限于心律失常还包括心肌梗死、心肌缺血、心包疾病以及药物作用等，这种复杂性导致了能够确定由心律失常而产生 ST 段改变的样本在临床中较少。基于混淆矩阵对 ST 类型的错误判断能够表明 DSE-ResNet 模型对亚异常类型 ST 的识别不敏感，这可能是由于 ST 的训练样本数量稀少导致 DSE-ResNet 能够提取到的特征有限。此外，专业医生对 ST 的诊断意见不一[62]，导致样本标注错误也是原因之一。对于亚异常类型 AF 和 Block，所提出的模型分别获得了 0.944 和 0.913 的F_1分数。

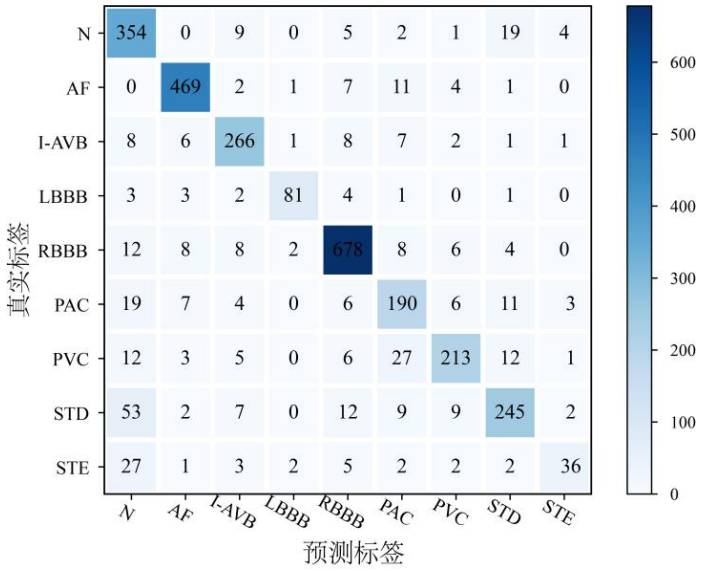

图 3.9 混淆矩阵

根据混淆矩阵计算得到了 DSE-ResNet 在隐藏测试集上的具体分类性能。表 3.4 显示了在 CPSC2018 隐藏数据库中模型对于不同心律失常的准确度、精确率、灵敏度和特异度分数。正常心律和 8 类心律失常的平均准确度和平均特异度分别为 0.965 和 0.979，且均在 LBBB 上达到最大值。值得注意的是 LBBB 在训练集中是所有分类中样本数量最少的类型（仅有 203 组），而 DSE-ResNet 对 LBBB 识别的误诊率非常低。通过寻找医学方向的依据表明，这可能是由于 LBBB 的识别在临床判断中具有多种关键特征，例如 QRS 波的时限异常（男性≥140ms，女性≥130ms）、QRS 波形态异常（导联 V1 的 QRS 波呈 QS 形）以及在导联 I、aVL、V1、V2、V5 和 V6 中至少有两个或两个以上导联存在 QRS 波的切迹或顿挫[66]。

（3）case-3：双极肢体导联被擦除。即导联I、II和III的所有采样点的信号值被填充为 0。

（4）case-4：单极加压肢体导联被擦除。即导联 aVR、aVL 和 aVF 的所有采样点的信号值被填充为 0。

（5）case-5：导联II、aVR、aVL 和 aVF 的所有采样点的信号值被填充为 0。

（6）case-6：胸前导联被擦除。即导联 V1、V2、V3、V4、V5 和 V6 的所有采样点的信号值被填充为 0。图 4.2 展示了不同擦除导联组合方式的原理图。

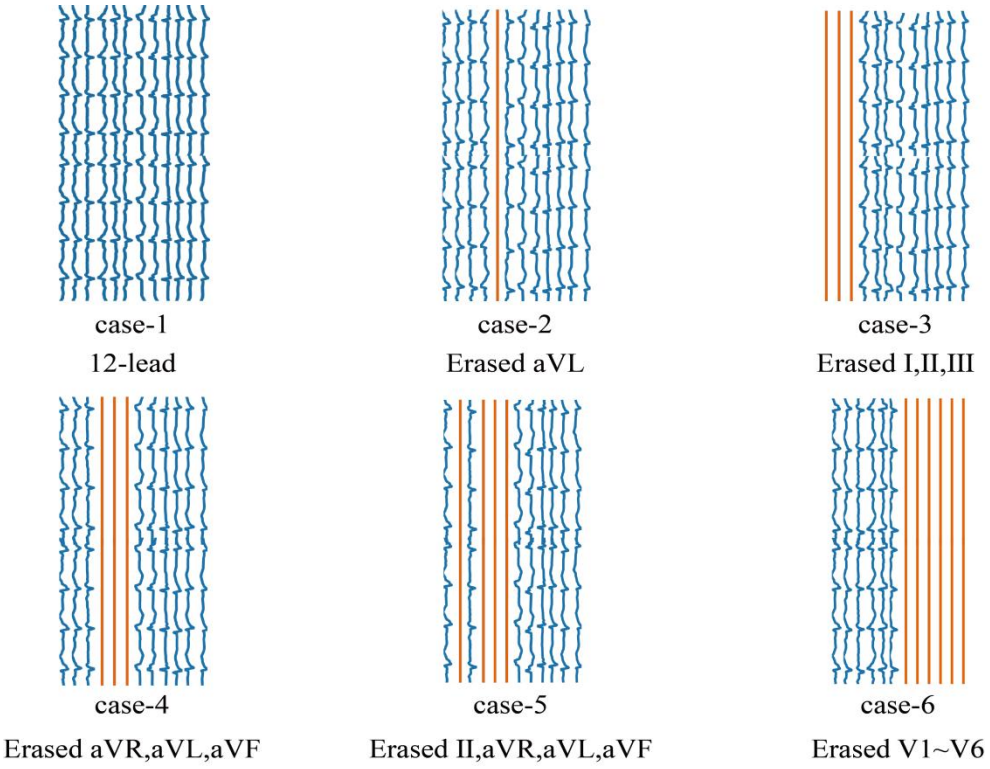

图 4.2 不同擦除导联组合方式原理图

注：导联的排列方式从左至右依次为导联I、II、III、aVR、aVL、aVF、V1、V2、V3、V4、V5 和 V6

4.2.4 改进二维心电图

融合数据库样本的原始十二导联 ECG 信号长度不同。为了保证输入 DNN 的十二导联 ECG 具有相同的维度，需要进行数据预处理。常用方法包括对短时长的样本填充 0 到指定长度，以及截断过长的样本以确保所有样本具有相同的长度。第三章的研究采取了这种方式，但是十二导联 ECG 不同时间段的波形结构不完全相同，填充可能会破坏原始十二导联 ECG 的结构，而裁剪十二导联 ECG 可能会抹除用于确定心律失常类型的关键波形。因此，通过填充或截断来保持样本长度一致是不合理的。

第三章中提及的切片操作，原始样本的十二导联 ECG 经过切片后会变成训练集中的多组样本，使得训练数据中存在同一患者的多个二维心电图样本。过多的

重复同一患者的样本可能导致模型过拟合。且经过维度扩充后二维心电图的维度变为$A_i \in \mathbb{R}^{8192*12*1}$，扩充通道维度为 1 只是为了适应二维 CNN 的训练要求，在本章节中提出了一种改进的二维心电图，称为“Block”。能够保证训练数据中患者样本的唯一性，同时在通道维度添加导联信息实现类似于图片具有 RGB 三通道数据的构造。

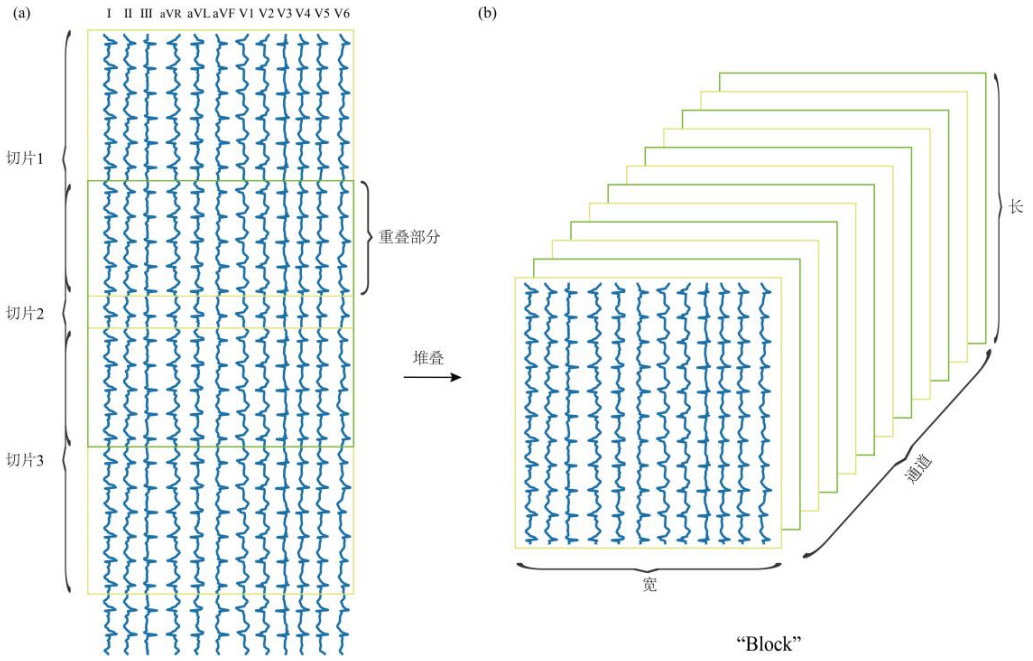

图 4.3 改进的二维化原理示意图

注：（a）切片原则示意图；（b）“Block”示意图

如图 4.3 所示，按照时序顺序十二导联 ECG 被切分并堆叠成一个“Block”。每个 Block 有三个维度：分别表示切片的长度、导联的个数和切片的个数。每个切片包含同一样本不同时间段的十二导联 ECG，经过切片堆叠可以最大程度地保留样本蕴含的信息。每组样本的原始十二导联 ECG 被切分为十二片，切片的长度为 2048（4.096s）。重叠部分的长度由原始信号长度 L 决定，当原始信号长度$L \in (2048, 12 \times 2048]$时，重叠部分长度为：

$$Overlap = \lfloor (12 \times 2048 - L)/11 \rfloor . \quad (4.1)$$

当原始样本长度$L > 12 \times 2048$，重叠部分长度为 0，截取该样本前 12×2048 的长度作为有效信号部分，其余舍弃。经过数据处理共得到 19883 组二维心电图$A \in \mathbb{R}^{2048*12*12}$。不同擦除方式的二维心电图结构示意图如图 4.4 所示。

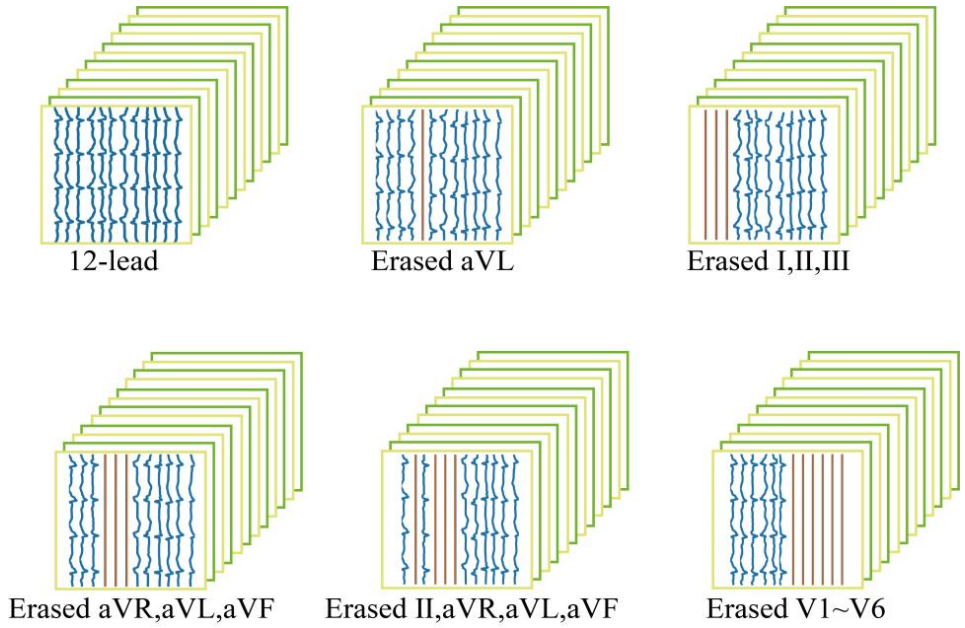

图 4.4 不同擦除方式的二维心电图结构示意图

4.2.5 模型构建

随着算力的提高，能够实现更深层次神经网络的学习。在一定范围内神经网络的深度越深，提取的特征越抽象，特征表达能力越强。在各种神经网络结构中，CNN 由于其具有局部连接和权重共享的特性，可以有效地提取局部特征。RNN 常用于处理一维序列数据[73]，其变体包括长短期记忆网络 LSTM[74]、双向长短期记忆网络 BiLSTM[75]等解决了 RNN 长期依赖的问题。此外，ResNet[76]因为克服了深度学习过程中的梯度消失和爆炸问题而成为深度学习领域最常用的框架。

本章节提出的模型以第三章提出的 DSE-ResNet 模型为基线，延伸和扩展了五种不同结构的 DNN 模型来实现自动心律失常识别。构建的五个 DNN 模型的结构如图 4.5 所示，具体如下：

（1）model-1：该模型为 DSE-ResNet，其中 ResNet 用于提取导联内部和导联间的特征，DSE 用于提取十二导联 ECG 不同时间片段的全局特征，年龄和性别作为辅助特征提高分类性能。

（2）model-2：该模型取自 DSE-ResNet 中的部分结构，包括 ResNet 结构和患者的年龄和性别信息。

（3）model-3：该模型由 ResNet 和 LSTM 组成。LSTM 是 RNN 的一种变体，通过引入门机制（输入门、遗忘门、输出门）和细胞状态，解决了 RNN 无法处理长期依赖的问题，其结构如图 2.11 所示。LSTM 通过各种门函数保留重要特征，以确保它们在长期传播过程中不丢失。由于其独特的特性，LSTM 适用于对时间序列数据（例如十二导联 ECG）进行建模。

（4）model-4：该模型由 ResNet 和 BiLSTM 组成。BiLSTM 是基于 LSTM 的变体，由前向 LSTM 和后向 LSTM 组成，其结构如图 2.12 所示。其优点是可以同时捕获时序数据的前向信息和后向信息。

（5）model-5：该模型为 SE-ResNet。SE 在通道维度上实现了注意力机制[50]。其结构如图 3.4 所示。在本文中，通道维度所存储的数据表示十二导联不同时间段的采样信号。SE 可以通过增加重要切片的权重和弱化次要切片的权重来提高模型的注意力。

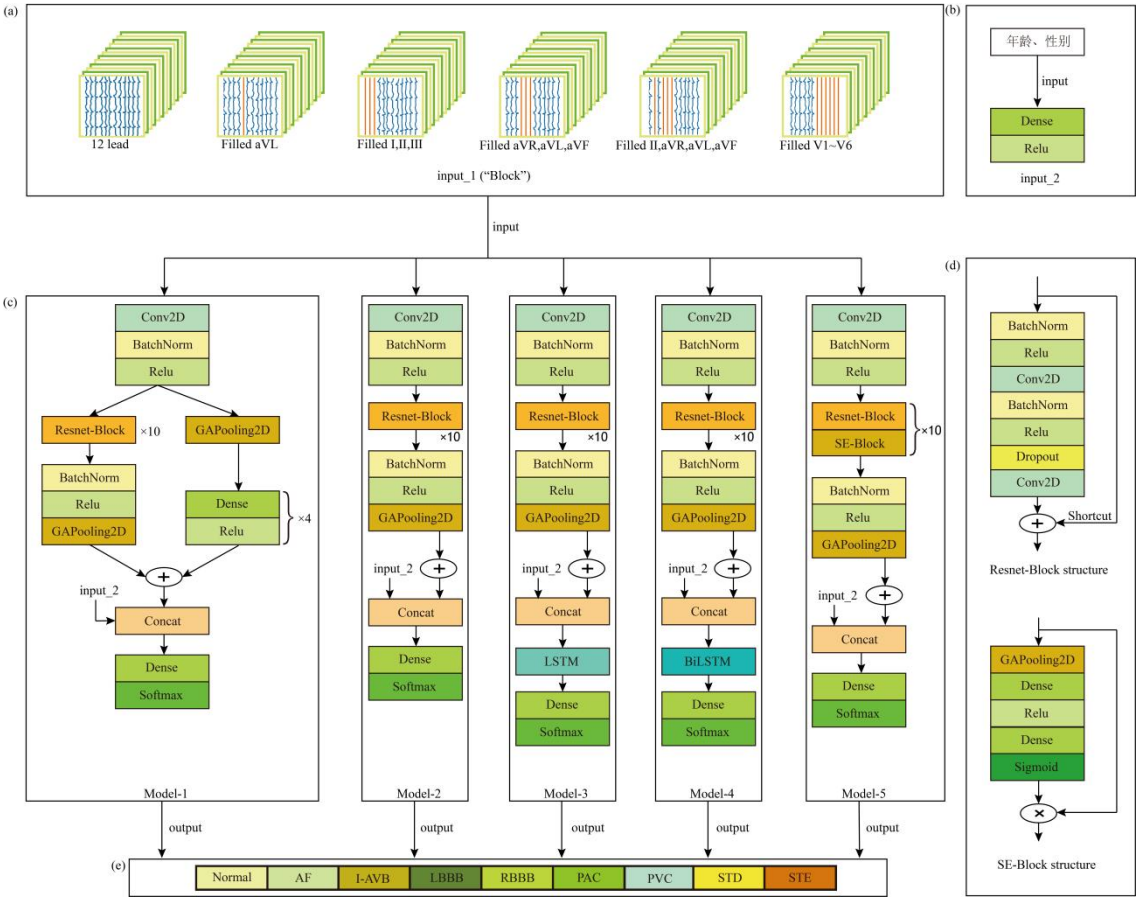

图 4.5 DNN 模型结构

注：（a）包含 5 种不同的擦除方法的二维心电图 A；（b）DNN 的第二部分输入源，包含患者的年龄和性别；（c）不同结构的 DNN 模型结构；（d）ResNet-Block 和 SE-Block 的组成结构；（e）心律失常分类类别

4.2.6 训练和测试

融合数据库样本被划分为三组：训练集、验证集和测试集。所有样本随机打乱并按照标签类别分配 77%为训练集，8%为验证集，15%为测试集。验证集用于监测模型训练过程中的过拟合现象，验证集和测试集中的样本不参与训练过程。

DNN 模型的输入数据来自两部分如图 4.5（a）和图 4.5（b）所示，分别代表二维心电图“Block”和辅助特征。辅助特征包括性别和年龄，其中男性和女性进

平均F_1=0.756)，case-3（擦除导联I、II和III）和 case-5（擦除导联II、aVR、aVL 和 aVF）在 model-2 中达到最高的 *Recall*=0.759，case-3 在 model-4 中取得最优 *Precision*=0.788。

表 4.4 不同擦除组合在不同 DNN 模型的相关评价分数

Precision 表	case-1	case-2	case-3	case-4	case-5	case-6
model-1（DSE-ResNet）	0.77	**0.797**	0.794	0.78	0.761	0.735
model-2（ResNet）	**0.816**	0.781	0.772	0.747	0.767	0.78
model-3（ResNet+LSTM）	0.755	**0.795**	0.771	0.677	0.771	0.708
model-4（ResNet+BiLSTM）	0.774	0.731	**0.788**	0.769	0.774	0.741
model-5（SE-ResNet）	**0.802**	0.692	0.801	0.758	0.798	0.739
平均值	0.783	0.759	**0.785**	0.746	0.774	0.741
Recall 表	case-1	case-2	case-3	case-4	case-5	case-6
model-1（DSE-ResNet）	**0.746**	0.732	0.71	0.7	0.717	0.719
model-2（ResNet）	0.737	0.73	**0.759**	0.707	**0.759**	0.674
model-3（ResNet+LSTM）	**0.737**	0.73	0.728	0.636	0.7	0.712
model-4（ResNet+BiLSTM）	**0.753**	0.694	0.707	0.675	0.701	0.734
model-5（SE-ResNet）	**0.76**	0.689	0.725	0.715	0.744	0.743
平均值	**0.747**	0.715	0.726	0.687	0.724	0.716
宏平均F_1表	case-1	case-2	case-3	case-4	case-5	case-6
model-1（DSE-ResNet）	0.752	**0.758**	0.744	0.727	0.732	0.717
model-2（ResNet）	**0.768**	0.75	0.763	0.719	0.761	0.707
model-3（ResNet+LSTM）	0.743	**0.756**	0.739	0.654	0.727	0.707
model-4（ResNet+BiLSTM）	**0.761**	0.71	0.739	0.714	0.731	0.734
model-5（SE-ResNet）	**0.779**	0.683	0.74	0.734	0.766	0.738
平均值	**0.761**	0.731	0.745	0.71	0.743	0.721

这表明缺少某些导联信息可能会提高 DNN 模型的 *Precision*、*Recall* 或宏平均F_1分数。不同模型对正常和 8 类类型心律失常的F_1分数如表 4.5 所示。显然擦除不同导联组合训练后的不同 DNN 模型对于具有足够样本的类型都显示出优异的性能。然而对于样本数量过少的 STE 类型所有模型都表现出较差的性能。出乎意料的是 PAC 类型在所有擦除组合中也都表现出较差的性能，结合第三章的实验对于 PAC 的分类性能表现良好推断可能是改进后的二维心电图使得深度学习模

型对这种类型的特征学习造成阻碍。

不同擦除组合在所有模型的平均 *Precision*、*Recall* 和F_1分数如图 4.6（a）所示。统计结果表明，在 case-1 中得到最优平均F_1=0.761 和 *Recall*=0.747。与预先想象的结果相反，发现在 case-3 中获得了最佳平均值 *Precision* =0.785。同时 case-3 的模型平均 *Recall* 和F_1仅次于 case-1（没有导联信息被擦除）。case-5 相较于 case-1 模型平均F_1性能降低了 0.018，与 case-3 相比仅降低 0.002。

图 4.6（b）-（f）给出了每个 DNN 模型对于不同擦除组合的 ROC 曲线。对比发现 case-3 在 model-3 中取得最高 AUC，在多数模型中达到了次高的 AUC。case-5 的性能略低于 case-3。

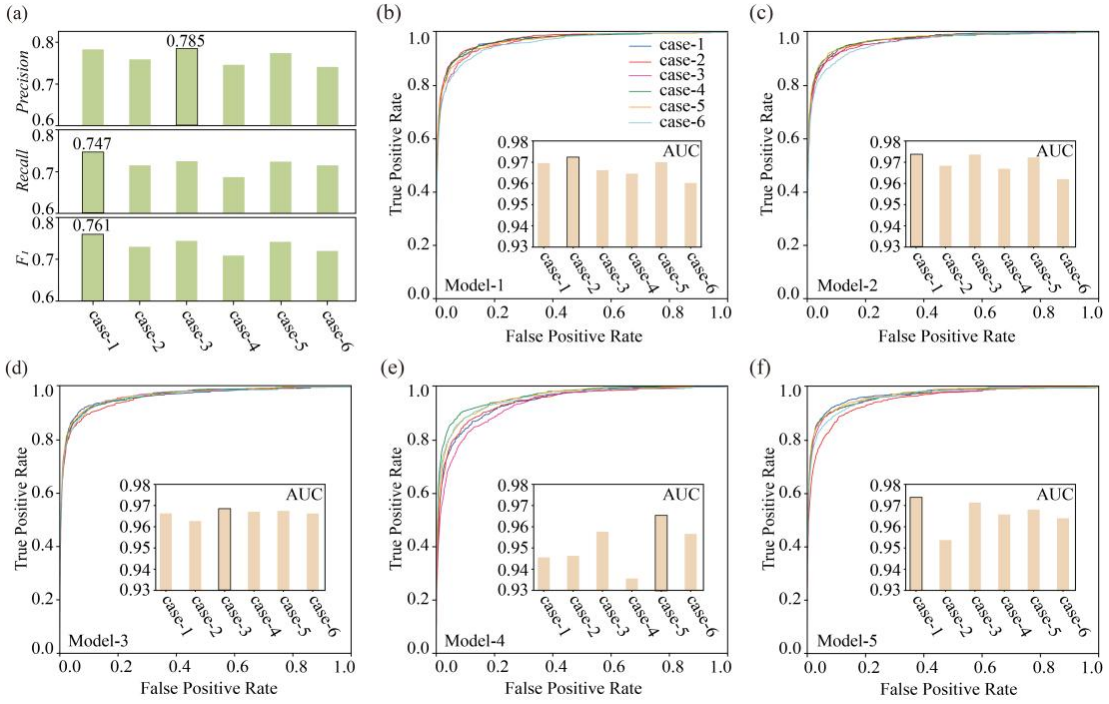

图 4.6 不同 DNN 模型应用不同导联擦除组合的分类性能

注：（a）模型平均 Precision、Recall 和F_1分数图；（b）-（f）不同擦除组合在不同 DNN 模型的 ROC 曲线和 AUC

综合表明，导联I、II、III与导联 aVR、aVL、aVF 存在的逻辑计算关系使得在深度学习过程中存在导联信息冗余。这种冗余仅略微影响 DNN 模型的分类性能。和最初构想的相同，存在逻辑换算关系的导联相关性较强，导致在深度学习过程中相关性较强的导联信号可能变为冗余特征。这为测量不同 ECG 信号的便携式设备应用部分导联信息诊断心律失常类型成为可能。

4.4 补充信息

4.4.1 选择导联 aVL 的过程