滚蛋吧！Parkinson's doctor

——基于帕金森远程监控数据集的模型分析

strength library and muscle laboratory（第七组）

陈立国 樊可 韩思越 黄永晟 杨远琨 俞铖昊

一. 背景介绍

帕金森氏疾病，是一种常见于中老年群体，严重影响活动能力的慢性疾病。美国国家生物信息中心（NCBI）2018 年 4 月发布的《北美帕金森氏病患病率》调查数据[1]显示，每 1000 位 45 岁以上中老年人群中，就约有 6 位患有帕金森氏疾病，而每 1000 位 80 岁以上老年人群中，竟约有高达 31 位患病者!

同时，随着人们生活水平的逐步提高，越来越多的人也格外地关注自身及身边亲人的身心健康。如此高的患病率，使得大家纷纷时不时会留意身边中老年群体是否有类似患病症状发生，一旦有疑似帕金森氏疾病症状，便不得不去各家医院预约挂号预诊检查，确保没有错过疾病初期最佳药物干预时间，费时费钱又费力。

而对于已确诊的帕金森氏疾病患病者，为了确保疾病处在药物可控状态，也不得不定期千里迢迢跑去专业医院进行症状复诊，挂号与各项指标细查，以明确是否有疾病恶化情况，决定是否需要改换药物与治疗方式，长期高频的医院复诊方式无论对于患者还是医院，都是一笔相当巨额的时间与经济开支。

另一方面，每一位帕金森氏疾病诊断医师，不仅需要应对疑似患者的彻查预诊，还需要兼顾到每一位确诊患者的长期跟踪治疗，明确是否有疾病恶化情况，尽可能避免错过疾病治疗黄金阶段，大把的时间精力投入也在所难免。

Lewin 集团机构（The Lewin Group）2019 年 7 月发布的《帕金森病的经济负担和未来影响》数据[2]显示，全世界帕金森氏疾病每年投入的医疗总成本约为 519 亿美元，其中包括直接医

[1] Marras, C., et al. "Prevalence of Parkinson's disease across North America." *NPJ Parkinson's disease* 4.1 (2018): 21.

[2] Grace Yang, et al. "Economic Burden and Future Impact of Parkinson's Disease." *The Lewin Group, Inc.* 7.5(2019)

疗成本（住院和药物治疗等）254 亿美元，以及 265 亿美元的非医疗成本（工作失误，工资损失，被迫提前退休和家庭护理时间等）。全世界每年在帕金森氏疾病上的医疗投入总成本，甚至超过百度企业目前的总市值（463 亿美元），可见该疾病医疗投入之高，负担程度之深重。

然而，即便在如此高的医疗投入下，仍然有大量的帕金森氏疾病患者，未能在最佳疾病诊断期间内，得到相应治疗。《解放军保健医学杂志》发布的"帕金森病诊治现状调查"文章[3]显示，仅有 3.75％的患者，在帕金森氏疾病发病初期，意识到发病情况并进行及时医疗确诊，帕金森氏疾病患者，就诊时间平均延迟甚至达到了 6.73 个月，即平均每一位患者会错过长达半年的黄金治疗时间与机会。由此，如何降低帕金森氏疾病预诊与长期跟踪复诊治疗的医疗成本，并且尽可能使每一位患者，及时觉察疾病，跟踪疾病发展情况，在最佳疾病诊断时间内获得合适医疗干预手段，已是迫在眉睫的问题。

因此，本案例选择 UCI 机器学习数据库，帕金森远程监控数据集语音数据，以医师评分为因变量，其他变量为自变量进行描述与模型分析。帕金森氏疾病预诊医师及医院机构，既能够借助我们的分析结果辅助就诊者的初步诊断，也能够使用我们训练的语音检测诊断模型，直接对就诊者语音信息进行预诊，作为疾病诊断辅助参考项，更能够通过我们设计的帕金森氏疾病远程线上诊断 APP，长期远程实时线上联系疾病确诊患者，进行长期远程疾病状况跟踪监控与治疗诊断。

帕金森氏疾病确诊患者及身边亲属，也能够根据我们的描述分析结果，初步判断患者病情发展情况，并能够通过高频定时使用远程线上疾病诊断 APP 的方式，足不出户即可长期跟踪病情发展情况，在病情恶化第一时间联系医师复诊，改换合理医疗干预方式。

其他中老年人群或疑似帕金森氏疾病患者及身边家属，也能够通过我们描述分析结果，初步检查身边人群是否存在疑似帕金森氏疾病症状，并能够定期使用远程线上诊断 APP，简单初步诊断自己或身边中老年群体是否患有帕金森疾病，一旦发现疾病症状，能够第一时间进行干预与治疗。

[3] 韩艳, et al. *帕金森病诊治现状调查*. Diss. 2008.

二. 数据介绍

由于帕金森氏疾病医疗数据多包含患者私人信息，受隐私保护需要并未在网络渠道公开。因此本案例选择美国某医院捐赠入 UCI 机器学习数据库匿名帕金森远程监控数据集，共 5875 条语音数据。

数据集因变量医师评分，源于患者全面疾病复查当天，多名专业医师对患者各项生活状况多项专业指标（包括心理状态，日常活动，运动状态等）综合评分汇总结果[4]。当天所录制的语音数据，既作为疾病复查的一小部分，受专业医师评估后被考量入医师评分当中，也被该所医院加以隐私存档用于数据分析。总之，因变量医师评分标准，具备充分的权威性与可信度。

自变量声带信息部分，“基频微扰”（Jitter）取自频率平均绝对差除以平均频率所得值，用以衡量声带粗糙程度[5]，“振幅微扰”（Shimmer）取自各振幅平均绝对差除以平均振幅，用以衡量声带嘶哑程度[6]。“谐噪比”（HNR）反映周期性谐波总能量与非周期性噪波总能量比值大小[7]，“噪谐比”（NHR）反之。两者均反映声带内谐波总能量高低，即声音悦耳程度。可能源于两者复杂计算方式略有差异，真实数据中谐噪比与噪谐比并非呈现倒数关联（本数据中谐噪比与噪谐比乘积均值为 0.52，标准差为 0.45）。

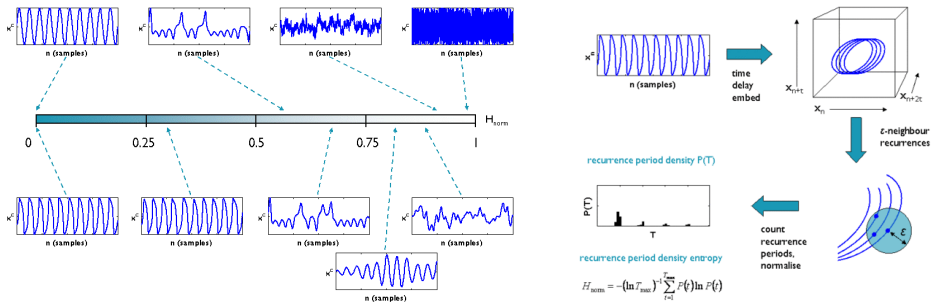


图 1 递归周期密度熵大小衡量，计算方式示意图

语音稳定性部分，“递归周期密度熵”（RPDE）[8]表征时间序列重复相同序列的程度[9]（衡量

[4] Tsanas, Athanasios, et al. "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests." IEEE transactions on Biomedical Engineering 57.4 (2009): 884-893.
[5] The Praat Program: https://www.fon.hum.uva.nl/praat/manual/Voice_2__Jitter.html
[6] The Praat Program: https://www.fon.hum.uva.nl/praat/manual/Voice_3__Shimmer.html
[7] The Praat Program: https://www.fon.hum.uva.nl/praat/manual/Harmonicity.html
[8] Little, Max A., et al. "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection." Biomedical engineering online 6.1 (2007): 23.
[9] Little, Max, et al. "Nonlinear, biophysically-informed speech pathology detection." 2006 IEEE

方式示意图如图 1 所示），“去趋势波动分析”（DFA）表征信号的统计自相关性[10]，“音高周期熵”（PPE）则进一步反映帕金森氏疾病患者发音时对固定音高的控制受损情况，并考虑到受试者的平均音高，而为帕金森氏疾病诊断专门设计的衡量指标[11]。三者从不同维度衡量语音稳定性情况，其中 0 为极端稳定，1 为极端不稳定（如图 1 所示）。篇幅有限各自变量详细计算方式不逐一展开，感兴趣读者可自行查阅参考文献。以上信息可全部由测试者语音音频中直接提取，因此只需测试者远程提供语音音频（包括测试者性别年龄记录），即可提取所有自变量进行帕金森疾病初步诊断分析。

表一 变量表

变量类型		变量名	详细说明	取值范围与备注
因变量		医师评分	连续变量	7~55（单位：分）
自变量	基本信息	性别	定性变量 共 2 个水平	男，女
		年龄	定性变量 共 3 个水平	>70, 60~70, <60（单位：岁）
	声带信息	Jitter(%)基频微扰	连续变量	8.30×10^{-4}~1.00×10^{-1}
		Shimmer:APQ11 振幅微扰	连续变量	2.50×10^{-3}~2.76×10^{-1}
		HNR 谐噪比	连续变量	1.66×10^{0}~3.79×10^{1}
		NHR 噪谐比	连续变量	2.86×10^{-4}~7.48×10^{-1}
		Jitter 异常情况	定性变量 共 2 个水平	异常，非异常
		Shimmer 异常情况	定性变量 共 2 个水平	异常，非异常
		HNR 异常情况	定性变量 共 2 个水平	异常，非异常
		NHR 异常情况	定性变量 共 2 个水平	异常，非异常
	语音稳定性信息	递归周期密度熵	连续变量	1.51×10^{-1}~9.66×10^{-1}
		去趋势波动分析	连续变量	5.14×10^{-1}~8.66×10^{-1}
		音高周期熵	连续变量	2.20×10^{-2}~7.32×10^{-1}

[10] Peng, C-K., et al. "Mosaic organization of DNA nucleotides." Physical review e 49.2 (1994): 1685.

[11] Little, Max, et al. "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease." Nature Precedings(2008): 1-1.

为了更好地进行描述与建模分析，我们基于个人理解，对原数据集自变量做了一定增加与筛减调整。由于原数据集基频微扰，振幅微扰多项检指标存在高度正相关情况（如图 3 所示），我们将基频微扰，振幅微扰各项指标，分别与因变量医师评分进行线性回归，选取拟合效果最佳的基频微扰指标，振幅微扰指标，作为后续描述与模型分析所使用的自变量，即直接筛除其他高度正相关基频微扰，振幅微扰指标。与此同时，根据官方文献对声带信息是否异常的判断方式[12]，我们分别对基频微扰，振幅微扰，谐噪比，噪谐比设置异常情况判断，作为新离散自变量，用于后续描述与模型分析。附录 1 中线性模型假设检验也进一步表明，我们新提取的自变量具备一定程度合理性。预处理后用于下文描述分析与模型分析的变量情况如表一所示。

三. 描述分析

为直观地了解帕金森氏疾病患病程度与年龄，性别，声带，语音稳定性之间的关联，我们对各因变量与自变量分别进行可视化描述分析。先从因变量医师评分开始。如图 2 医师评分直方图所示，医师对于该疾病的评分，总体呈现右偏态分布，右尾较厚，即存在较多症状极端恶化的帕金森氏疾病患者。其中 20 分以下蓝色部分表明患者尚处于帕金森氏疾病早期，而 20 分到 40 分之间绿色部分表明患者处于帕金森氏疾病中期，40 分以上黄色，红色部分表明患者已进入帕金森氏疾病晚期。可见绝大部分复诊患者均处在疾病中期，有相当一部分患者已步入帕金森氏疾病晚期，其中最高得分为 54.99 分，反映该患者帕金森氏疾病已经相当严重。

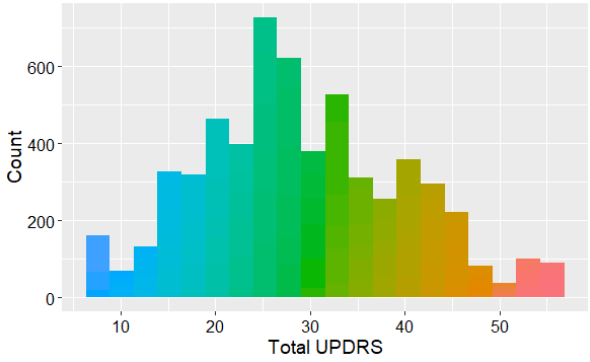


图 2 医师评分直方图

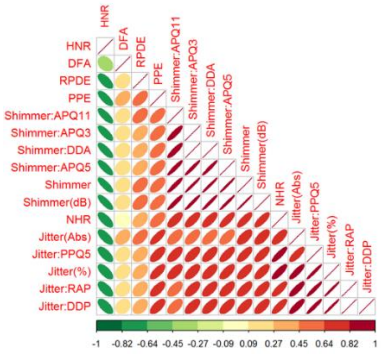


图 3 原数据集自变量相关性图

[12] The Praat Program: https://www.fon.hum.uva.nl/praat/manual/Voice_2__Jitter.html
[12] The Praat Program: https://www.fon.hum.uva.nl/praat/manual/Voice_3__Shimmer.html
[12] The Praat Program: https://www.fon.hum.uva.nl/praat/manual/Harmonicity.html

接下来我们入手各自变量。如图 3 原数据集自变量相关性图所示，除基频微扰（jitter），振幅微扰（shimmer）外，其余自变量不存在明显线性相关情况。因此经过上一部分变量筛选后（筛选后基频微扰，振幅微扰均仅保留一条），各自变量不再存在高度线性相关情况。同时，除谐噪比（HNR）与其他各自变量呈现负相关外，其他各自变量两两均呈现正相关。

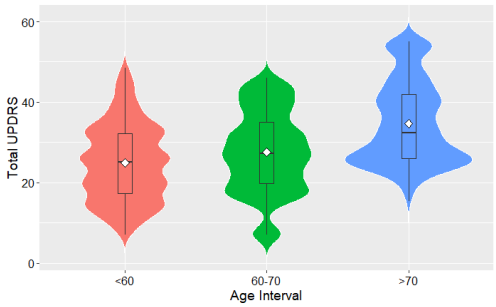


图 4　医师评分与年龄小提琴图

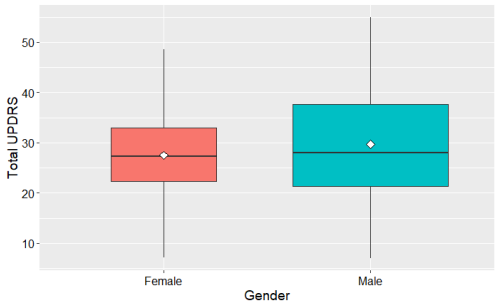


图 5　医师评分与性别箱线图

而后我们逐一考察各因素对帕金森氏疾病严重程度的影响。首先是年龄因素，由图 4 医师评分与年龄小提琴图可见，60 岁以下人群普遍得分在 10 分至 30 分之间，即帕金森氏疾病早期与中期阶段；60-70 岁人群普遍得分在 20 至 40 分之间，大多处于疾病中期阶段；70 岁以上人群普遍得分在 20 至 60 分，即疾病中晚期阶段。显而易见，随着年龄的逐步增加，帕金森氏疾病呈现严重与恶化趋势。因此对于高龄群体，尤其 70 岁以上群体，需要格外地关注帕金森氏疾病的检查与预诊，确诊患者也需要定期进行复诊，以观察是否存在病症恶化情况。

其次是性别因素，由图 5 医师评分与性别箱线图可以看出，无论是中位数，上四分位数箱线和平均值点对应得分，男性的疾病评分总体上均略高于女性。同时箱线图箱宽也显示，在此次随机抽样数据中患有帕金森氏疾病的男性样本数也明显大于女性样本数。由此可以推断，帕金森氏疾病在男性群体当中更为高发与严重，男性群体应当更多地关注该疾病的检查与预诊。

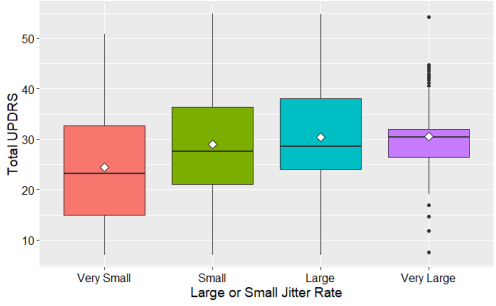


图 6　医师评分与基频微扰箱线图

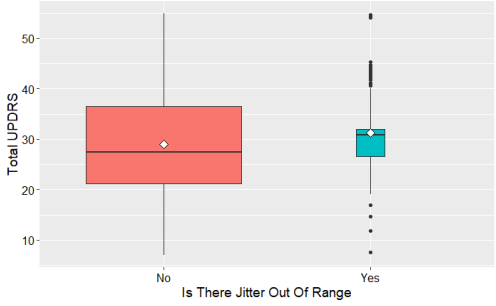


图 7　医师评分与基频微扰异常情况箱线图

接着考察声带信息对帕金森氏疾病严重程度的影响。从反映声带粗糙程度的基频微扰，反映声带嘶哑程度的振幅微扰入手。由图 6 医师评分与基频微扰箱线图可以看出，随着基频微扰值（jitter）的逐步增加，即声带粗糙程度逐步明显，帕金森氏疾病明显呈现严重与恶化趋势。图 7 医师评分与基频微扰异常情况箱线图也能够看出，尽管基频微扰异常情况比例非常小，但基频微扰异常，声带异常粗糙对应患者，得到的诊断得分大体上仍然略高于非异常患者。

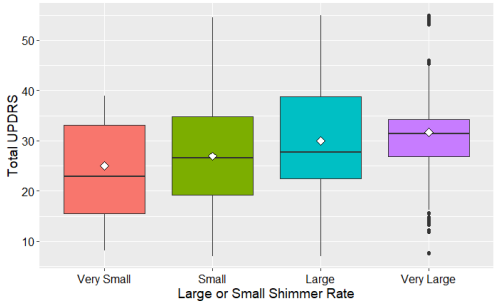


图 8 医师评分与振幅微扰箱线图

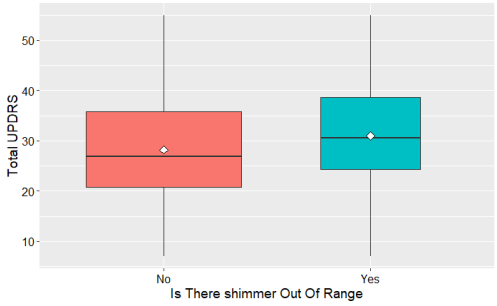


图 9 医师评分与振幅微扰异常情况箱线图

同样，如图 8 医师评分与振幅微扰箱线图所示，当振幅微扰（shimmer）对应值较小时（very small/small），帕金森氏疾病诊断得分偏低，而振幅微扰，即声带嘶哑程度对应值较大时，患者声带越为嘶哑，所对应的帕金森氏疾病越为恶化。同样图 9 医师评分与振幅微扰异常情况箱线图也反映出，声带异常嘶哑的患者音频，得到的诊断得分也略高于非异常患者。由此可见，医学界广泛认定的基频微扰异常判断与振幅微扰异常判断，都能够作为帕金森氏疾病严重程度的有效衡量标准。一旦发现身边中老年群体具备声带粗糙程度，声带嘶哑程度严重恶化情况，都有可能为疾病雏形而需要进一步关注留意与检查诊断。

图 10 医师评分与谐噪比小提琴图

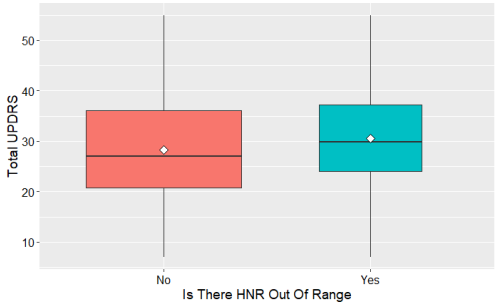


图 11 医师评分与谐噪比异常情况箱线图

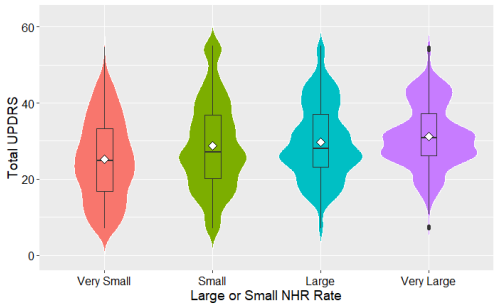


图 12 医师评分与噪谐比小提琴图

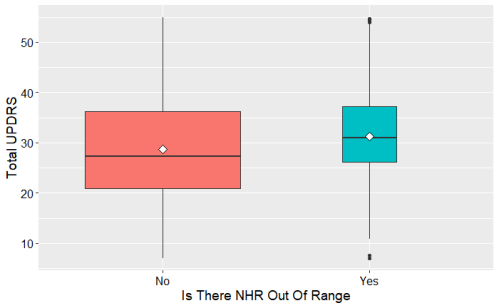
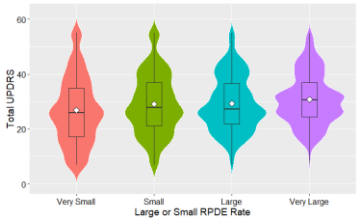
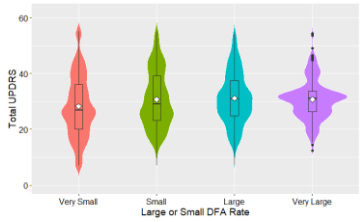


图 13 医师评分与噪谐比异常情况箱线图

与基频微扰，振幅微扰结果相似，图 10 医师评分与谐噪比（NHR）小提琴图，图 12 医师评分与噪谐比（HNR）小提琴图同样显示出，随着谐噪比值逐步增加，噪谐比值逐步降低，所对应的帕金森氏疾病逐步趋向严重与恶化。图 11 图 13 也同样映出，即便都是帕金森氏疾病患者，音频谐噪比，音频噪谐比异常情形均略少于音频谐噪比，音频噪谐比正常情形，而音频谐噪比，音频噪谐比异常患者，诊断得分也有一定偏高。可见，医学界认定的谐噪比异常判断与噪谐比异常判断，都能够作为帕金森氏疾病严重程度的有效衡量标准。





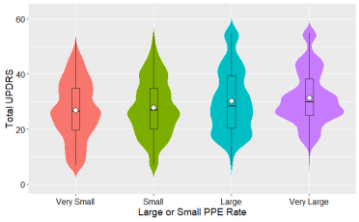


图 14 医师评分-RPDE 小提琴图 图 15 医师评分-DFA 小提琴图 图 16 医师评分-PPE 小提琴图

最后看语音稳定性信息对帕金森氏疾病严重程度的影响。如图 14,15,16 所示，无论是递归周期密度熵（RPDE），去趋势波动分析（DFA）还是音高周期熵（PPE），较小的声波熵（very small/small）对应的患者诊断得分大多分布在 10 至 40 分之间，即帕金森氏疾病早期与中期，而较大的声波熵（very small/small）对应的患者诊断得分大多分布在 20 分以上，即帕金森氏疾病中晚期。随着各个声波熵逐步增加，即随着患者声音稳定性的逐步恶化，帕金森氏疾病同样趋向恶化。可见递归周期密度熵（RPDE），去趋势波动分析（DFA），音高周期熵（PPE）均可作为帕金森氏疾病衡量标准，在与身边中老年人交流过程中一旦发现交流语音存在忽高忽低，忽响忽轻等声音不稳定情况，都需要进一步关注留意帕金森氏疾病检查诊断。

四. 模型分析

（一）模型选择

为了进一步量化各影响因素对于帕金森氏疾病的影响程度，也为帕金森氏疾病远程监控提供模型预测参考方法，我们尝试以医师评分为因变量，其他因素为自变量，建立线性回归，提升树，决策树，神经网络，随机森林和 KNN 模型，以量化各影响因素与疾病关联。经过十折交叉验证后，各个模型的拟合优度 R 方如图 17 所示，R 方越大表明拟合程度越佳。可见，其中 KNN 模型拟合效果最好，达到了 0.569，其次是随机森林，神经网络和决策树模型，普遍在 0.35 左右，因此我们最终选择 KNN 与决策树模型做模型预测。

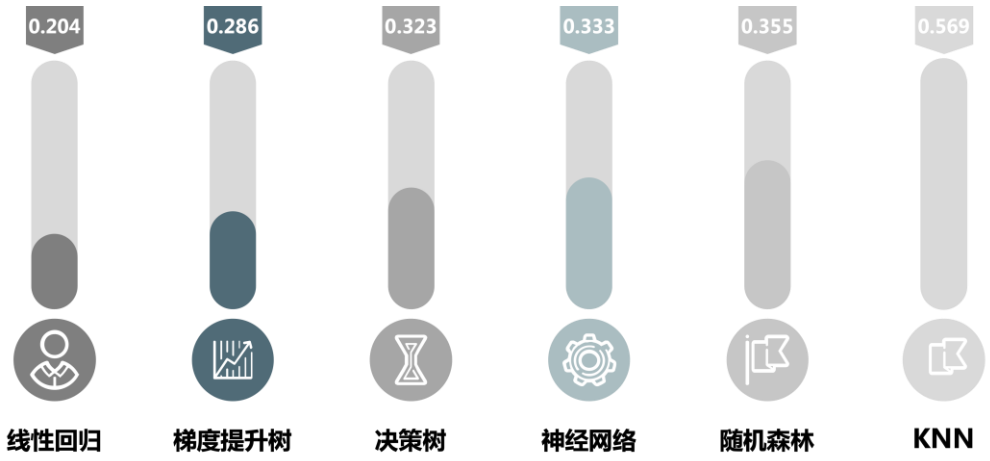


图 17 各模型 R 方对比图

对于各模型结果，我们猜测由于数据集中存在大量连续自变量，因此基于“一刀切决策选择”决策选择的树模型（梯度提升树，决策树，随机森林）效果并不理想。而 K 近邻模型更贴近这个医师评分的内在机理，即通过对比就诊患者与过往相似患者疾病情况（类似选择最近 K 个邻近训练点），推测就诊患者患病程度（类似取平均），相似的症状将导致相似的评分。

（二）变量重要性

为了更直观地反映出各个影响因素对于帕金森氏疾病的影响程度，我们将梯度提升树，神经网络，随机森林模型中各个自变量进行可视化。图 18 梯度提升树（gbm），神经网络（nnet），随机森林（rf）变量重要性柱状图，各个模型起始点（如梯度提升树约 7.2）分别对应考虑所有自变量情况下，梯度提升树，神经网络，随机森林预测结果 RMSE 均方根误差值，

各自变量对应柱的终点（如梯度提升树年龄自变量约为 11）对应仅筛去该自变量情况下，模型预测结果 RMSE 值，以两者之差反映各自变量重要性程度。同时，我们将图 18 的各个模型变量重要性的大小等比缩放后，以图 19 雷达图的方式进一步对比各个模型变量重要性情况。

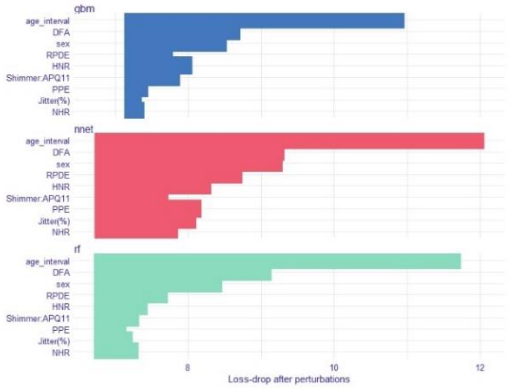


图 18 gbm, nnet, rf 变量重要性柱状图

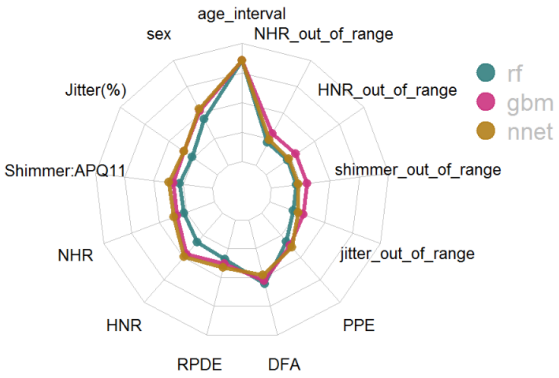


图 19　gbm, nnet, rf 变量重要性雷达图

由图 18 梯度提升树，神经网络，随机森林变量重要性可以看出，年龄和性别两个因素，无论在梯度提升树，神经网络还是随机森林当中，都对于帕金森氏疾病影响最为明显。其次的影响因素是 DFA，即音高周期熵，以及 HNR，即谐噪比，由此表明患者语音信息当中的声带信息与声音稳定性信息，仍然具备一定的影响作用。而图 19 则更为直观地显示，三个模型中拟合程度最佳的随机森林模型，年龄，性别，音高周期熵变量重要性高于其他两个模型（体现于其他自变量相对年龄变量重要性低于其他两个模型），而预测效果最糟糕的梯度提升树振幅微扰，振幅微扰异常情况，谐噪比异常情况，噪谐比异常情况考量的变量重要性高于其他两个模型。由此更能反映年龄，性别，音高周期熵因素在较好的模型预测中，更起到决定作用。

（三）模型检验

接下来我们尝试运用训练得到的模型，对任意一位帕金森氏疾病患者，或疑似帕金森氏疾病患者进行随时随地地远程疾病诊断。用户通过远程 APP 输入一段语言音频，年龄与性别信息，即可通过语言信息处理，得到基频微扰等多个声带，声音稳定性表征信息，再通过我们训练得到的模型，即可给出疾病诊断分数，进而进一步判断帕金森氏疾病严重程度。

我们模拟 Simpson 家庭成员数据作为示范。假设 Simpson 家庭成员信息如表二所示：

表二 帕金森氏疾病模拟数据

name	sex	age	jitter	shimmer	HNR	NHR	RPDE	DFA	PPE
Homer	male	52	0.0018	0.027	25	0.021	0.58	0.72	0.09
Marge	female	49	0.0012	0.012	30	0.011	0.22	0.65	0.07
Abe	male	78	0.0068	0.187	21	0.032	0.58	0.72	0.23
Mona	female	76	0.0036	0.016	24	0.014	0.53	0.64	0.11

首先考察决策树模型，如图 18 决策树输出结果树状图所示，年龄仍然被作为了首要决策因素，其中年龄小于 70 岁患者均分预测在 26 分，大于 70 岁患者均分预测在 35 分。对于大于 70 岁患者，下一决策因素即为性别，其中女性被预测为 29 分，而男性预测均分为 37 分。对小于 70 岁中年人群，下一决策因素为 DFA，以此类推。

使用决策树模型对表二数据进行模型检验，按照图 18 决策树输出结果，四人最终的帕金森氏疾病预测结果分别为 Homer:, Marge:20, Abe:49, Mona:29，由此除 Abe 预测为较为恶化的晚期外，其他三人均处在疾病早期与中期。因此根据决策树模型，我们建议对 Marge 使用疾病初期轻微药物治疗方式，对 Homer 与 Mona 则需要适当增加药物剂量与治疗力度，而对疾病预测结果最为严峻的 Abe 而言，一方面建议进一步去医院进行彻底详细的病症检查，必要时可采取手术治疗，另一方面家属亲人也需多多关注与照料正常起居生活。

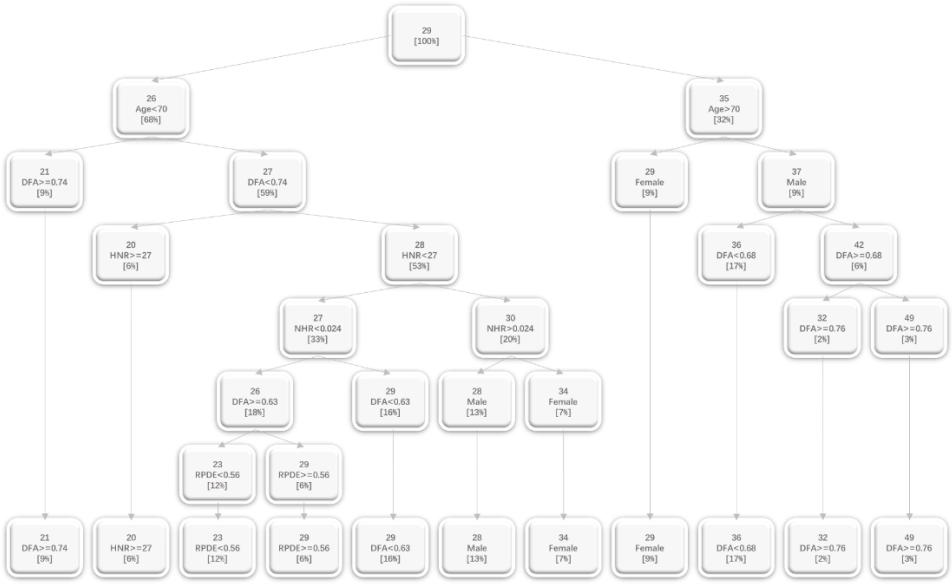


图 18 决策树输出结果树状图

紧接着是拟合效果最好的 KNN 模型。为了更直观地可视化 KNN 模型预测结果，我们使用 KNN 模型一次交叉验证中的测试集数据进行模型检验，测试集样本点与预测结果如图 19 所示，（由于无法将十三个自变量十三维空间一次性可视化，因此选择 HNR 与 DFA 两个维度作为模型检验结果示意）样本点的颜色分别表征预测值大小。可见，一方面，预测的医师评分越高的数据 DFA 越高，HNR 越低。另一方面，该散点图形成了两个簇（DFA>0.7，HNR<23 部分预测得分普遍小于 22，而 DFA<0.62，HNR>20 部分预测得分普遍大于 31），也能我们进一步了解评分与指标的内在机理提供启发。同时我们将上述模型检验测试集预测结果与实际结果绘制散点图进行对比并分性别展示（如图 20 所示）。可见大部分 KNN 模型预测数据都与实际结果较为接近，其中女性的预测结果相对男性更好。

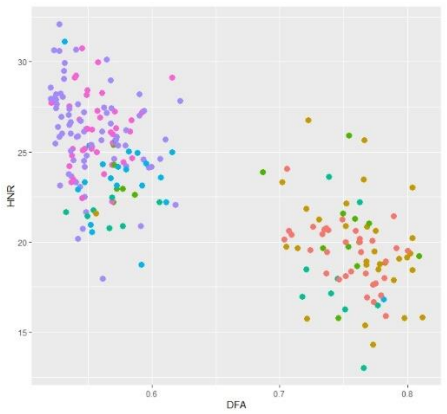


图 19 KNN 空间点对应预测范围散点图示意图

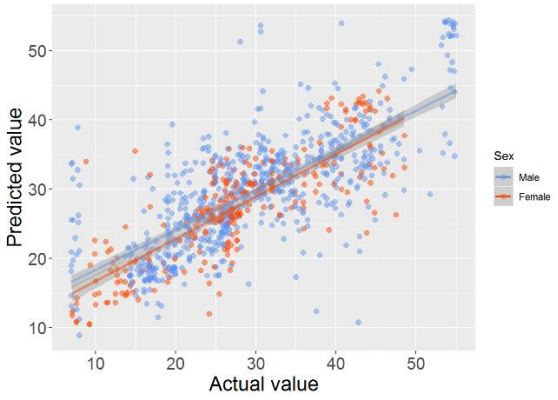


图 20 KNN 预测与实际结果对比散点图

五. 结论建议

为了分析语音信息等不同因素对于帕金森氏疾病患病程度的影响，使得帕金森氏疾病远程诊断成为可能，我们选择 UCI 机器学习数据库帕金森远程监控数据集语音数据进行可视化描述分析，并使用了线性回归，提升树，决策树，神经网络，随机森林和 KNN 模型，分别对数据进行建模分析。

从分析结果中可以看出，年龄与性别两项基本信息仍然对帕金森氏疾病影响程度最大。具体地，老年群体普遍比中年群体帕金森氏疾病病症更为严重，而女性相比男性病症更轻微。其他各项语音信息，例如音高周期熵，谐噪比等，同样具备不可忽略的重要影响。其中基频微扰

越大，即声音越粗糙，帕金森氏疾病病症越严重，振幅微扰越大，即声音越嘶哑，帕金森氏疾病病症越严重，谐噪比越低，即噪音成分越高，帕金森氏疾病病症越严重。而语音稳定性信息，无论是递归周期密度熵，去趋势波动分析，还是音高周期熵，熵越大，语音越不稳定，显示帕金森氏疾病病症即越为严重。

同时，我们构想设计帕金森氏疾病远程线上诊断 APP。任意一位帕金森氏疾病患者或疑似疾病患者，都能够随时随地在移动端输入或当场录入一段语音信息，并输入个人基本信息（性别，年龄等），即可远程提取声带信息与声音稳定性信息各变量，套用事先训练完成的模型，直接得出帕金森氏疾病诊断评估，作为远程线上诊断结果，由此能够节省大量来自医院挂号，医师仪器，诊断分析的人力物力财力。同时，方便快捷低成本的疾病诊断方式，使得帕金森氏疾病患者或其他疑似患者，更易于高频率地进行疾病检查，由此进一步大幅缩短疾病就诊的时间延迟，更多的患者能够有效利用 “黄金时间”控制疾病进一步恶化。另外，也可通过该 APP 建立医师与帕金森氏疾病患者长期远程联系，医师能够根据患者定期远程诊断结果，判断患者疾病演变情况，足不出户即可实时全面兼顾数十数百位帕金森氏疾病患者病症情况，并在患者症状恶化的第一时间内采取对应治疗手段。

并且，我们分别对帕金森氏疾病预诊医师及医院机构，帕金森氏疾病患者及其身边家属，与其他中老年人群及其身边家属分别提出以下建议：

对于帕金森氏疾病预诊医师及医院机构：

1. 疾病预诊前可以先了解预诊者年龄，并且将预诊者年龄与性别列入预诊项，尤其对高龄男性需要更为慎重地进行疾病预诊。

2. 预诊中与患者交流时，可以留意预诊者语音是否粗糙，是否嘶哑，是否有发音不稳定的实际情况，上述情况下预诊者帕金森氏疾病可能性更高，需要更为慎重地进行疾病预诊。

3. 实际预诊流程中，可以引入语音检测环节，预诊者现场录制一段语音，由语音波段中提取声带信息与声音稳定性信息各变量，再使用我们事先训练完成的模型，直接对语音信息进行预诊者帕金森氏疾病诊断评估作为参考项。

4. 对于确诊帕金森氏疾病患者，可以使用我们构想设计的帕金森氏疾病远程线上诊断 APP，建

立与患者之间的长期远程线上联系，督促患者定期使用远程线上诊断 APP 进行初步疾病诊断，并能够根据患者长期疾病诊断情况，在必要时（疾病诊断得分上升时）请患者门诊挂号复查，及时调整药物与治疗方式。

对于诊断确认帕金森氏疾病患者及其身边家属：

1. 需要长期留意患者语言交流过程中，语音粗糙程度，嘶哑程度，声音稳定性的变化情况。一旦出现明显语音粗糙，嘶哑，不稳定变化情况，需要及时复诊。

2. 可以高频定时使用我们构想设计的帕金森氏疾病远程线上诊断 APP，定期录入语音信息进行长期实时跟踪疾病诊断。并且通过该线上诊断 APP 与负责医院医师进行长期远程线上跟踪联系，一旦医师发现某一阶段诊断得分上升，能够第一时间通知患者及时去医院复诊，改换药物或治疗方式，以确保不错过最佳治疗时期。

对于其他中老年人群或疑似帕金森氏疾病患者及其身边家属：

1. 需要长期留意患者语言交流过程中，语音是否粗糙，是否嘶哑，声音是否稳定，是否存在其他疑似帕金森氏疾病症状。一旦出现疑似帕金森氏疾病症状，需要及时进行帕金森氏疾病预诊以避免错过帕金森氏疾病治疗黄金时期。

2. 可以定期使用我们构想设计的帕金森氏疾病远程线上诊断 APP，进行简单帕金森氏疾病初步预诊，一方面能够以低成本的方式换得一份放心，另一方面一旦发现疑似疾病症状也能够及时干预治疗。

有关后续工作展望，一方面可以进一步与多家帕金森氏疾病诊断医院进行合作，得到更为充实的帕金森氏疾病患者语音数据集，另一方面也可结合疾病相关背景，尝试进一步由患者语音音频数据中提取更多有效信息用于模型预测，以提升模型适用性与精准度。同时，我们也可以辅助结合语音方面外，其他可远程获取的该疾病诊断方式信息（例如尝试让患者绘制螺旋线，或设计可佩带设备随时记录患者步态等）全面综合辅助疾病诊断，以更为全面地进行疾病预测与诊断。

附 1：线性模型 jitter, shimmer, HNR, NHR 异常情况假设检验：

首先，设自变量$x_1, x_2, \dots, x_n$与因变量 y 之间满足$\mathrm{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$,有 m 个样本，对每个样本，$\mathrm{y}_i$，$\mathrm{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni} + \varepsilon_i$

H_0是$\beta_0, \beta_1, \beta_2 \dots \beta_n$满足矩阵方程H$\beta$=d, β =$(\beta_0, \beta_1, \beta_2 \dots \beta_n)'$

对于自变量的 m 个样本，对所有 n 的自变量做线性回归，得到$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots \hat{\beta}_n$，$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_n x_{ni}$，$\mathrm{SSE}_0 = \sum_0^n (\hat{y}_i - \mathrm{y}_i)$

假定H_0成立,加入这个约束条件,做线性回归,得到$\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_n$, $\tilde{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{1i} + \tilde{\beta}_2 x_{2i} + \cdots + \tilde{\beta}_n x_{ni}$, $\mathrm{SSE}_1 = \sum_0^n (\tilde{y}_i - \mathrm{y}_i)$

根据理论，设矩阵方程Hβ=d 的解有p_1个自由度，SSE_1/σ^2服从自由度为$m - \mathrm{p}_1 - 1$的卡方分布，SSE_0/σ^2服从自由度为$m - n - 1$的卡方分布,$(\mathrm{SSE}_1 - \mathrm{SSE}_0)/\sigma^2$服从自由度为$n - \mathrm{p}_1$的卡方分布,就是矩阵方程提供的约束数目 k 的自由度的卡方分布,F=$\frac{(\mathrm{SSE}_1 - \mathrm{SSE}_0)/\mathrm{k}}{\mathrm{SSE}_0/(m-n-1)}$满足自由度为 k 和 m-n-1 的 F 分布。

在本次实验中，假设 jitter, shimmer, HNR, NHR 异常情况变量对于线性回归影响均不显著（H_0）, F 应当服从$F(4,5860)$分布,计算可得 F= 5.007594,远远大于$\mathrm{F}_{0.95}(4,5860) = 0.5661682$,因此拒绝$\mathrm{H}_0$，jitter, shimmer, HNR, NHR 异常情况变量有一定程度影响。

附 2：小组分工情况

变量预处理：全员讨论变量处理方式

描述分析：全员讨论图像呈现形式，韩思越负责代码

模型分析：俞铖昊负责回归树模型，樊可负责线性回归，神经网络模型，黄永晟负责 KNN 模型

分析报告：全员讨论整合细节，杨远琨负责最终整合

小组汇报：陈立国负责 PPT 制作，杨远琨与陈立国负责汇报