PLOS

研究论文

随机森林、声音象征与宝可梦的进化

Alexander James Kilpatrick [1☯]*, **Aleksandra Ćwiek**[2☯], **Shigeto Kawahara**[3]

1 国际传播，名古屋商科大学，名古屋，爱知县，日本，**2** 莱布尼茨普通语言学中心，Berlin，德国，**3** 文化语言学研究所，庆应义塾大学，东京，日本

☯ 这些作者对本研究的贡献相同。* alexander_Kilpatrick@nucba.ac.jp

Check for updates

开放获取

引用： Kilpatrick AJ、Cwiek A、Kawahara S（2023）随机森林、声音象征与宝可梦进化。《PLOS ONE》18(1)：e0279350。https://doi.org/10.1371/journal.pone.0279350

编辑： Maki Sakamoto，电气通信大学，日本

收稿日期： 2022年7月12日

接受日期： 2022年12月6日

发表日期： 2023年1月4日

同行评审历史： PLOS 认可同行评审流程透明化的益处；因此，我们允许在最终发表的文章旁一并发布同行评审和作者回复的全部内容。本文的编辑历程见此处：https://doi.org/10.1371/journal.pone.0279350

版权： © 2023 Kilpatrick 等人。本文为开放获取文章，依据知识共享署名许可协议发布，允许在任何媒介中不受限制地使用、分发和复制，但须注明原作者和来源。

数据可用性声明： 所有数据、脚本及其实现说明均可在以下 OSF资料库中获取。https://osf.io/pe24w/?view_仅=02e9327a7bd54b9280b57434a90ed83a

摘要

本研究构建了经过训练、能够利用声音象征对样本进行分类的机器学习算法，并报告了一项旨在测量这些算法对声音象征的理解能力、并与人类参与者进行比较的实验结果。研究使用宝可梦（虚构的电子游戏角色）的名称及其进化状态来训练随机森林。满足特定游戏内条件时，宝可梦会发生进化。进化会改变宝可梦的外观、能力和名称。在实验1中，我们使用构成日语、中文和韩语宝可梦名称的声音训练了三个随机森林，以将宝可梦分为进化前和进化后类别。随后，我们利用一项引出实验的结果训练了第四个随机森林；在该实验中，日本参与者为此前未见过的宝可梦命名。在实验2中，我们将名称长度作为一项特征，重新训练这些随机森林，并在一项分类实验中将随机森林的表现与人类进行比较；该实验要求日本参与者将实验1中引出的名称分为进化前和进化后类别。实验2揭示了实验1中存在的过拟合问题，我们通过一种新颖的交叉验证方法解决了这一问题。结果表明，随机森林能够高效地学习系统性的声音—意义对应模式，并且对样本的分类准确率高于人类参与者。

引言

自然语言处理（NLP）是一个结合计算语言学与人工智能的研究领域，致力于赋予计算机理解语言的能力，使其理解语言的方式与人类大致相同。本研究检验自然语言处理算法能否利用声音象征对样本进行分类；声音象征是人类语言中一种在自然语言处理领域长期遭到忽视的特征。而在现代语言学中，

人们通常认为声音与意义之间的关系是任意的 [1],但越来越多的研究揭示了声音与意义之间存在系统性关系，其中一些关系具有跨语言的普遍性。例如，许多语言的使用者倾向于将包含 [i] 与

资助： AK - 获得日本学术振兴会（日本东京）资助，_资助编号：20K13055https://www.jsps.go.jp/english/index.html 资助方未参与研究设计、数据收集与分析、发表决定或稿件撰写。

利益冲突： 作者声明不存在任何利益冲突。

小型物体，而包含 [a] 的词通常与较大型物体相关联 [2–5]。人类在婴儿期便能理解某些声音象征关联，据称这些关联有助于语言发展并促进词语习得 [6–9]。因此，如果自然语言处理算法旨在像人类一样理解语言，那么理解声音象征就十分重要。本研究关注随机森林算法（下文简称 RF： [10]），这是一种通常用于分类和回归任务的集成机器学习算法。该研究建立在 Winter 和 Perlman 最近的研究之上（[11]；另见 [12]），他们使用随机森林证明，英语单词中的大小与音位之间存在系统性的声音象征关系。

接下来，我们使用名为 Pokémon 的角色的虚构名称来构建并测试随机森林。Pokémon 最初于1996年作为一款电子游戏发行，是一个极受欢迎的跨媒体系列，尤其在其发源地日本 [13]。本研究测量随机森林的分类准确率，并将其与日本大学生的准确率进行比较。随机森林仅根据构成名称的声音，将 Pokémon 分为进化前和进化后类别。在实验1中，我们使用日语、普通话（以下简称中文）和韩语（以下简称韩语）Pokémon 名称中的声音，构建了三个随机森林。这些随机森林使用各自数据集的一个子集进行训练，再用剩余数据进行测试。所有随机森林对 Pokémon 的分类准确率都高于随机水平，其中日语随机森林的表现最佳，因此后续实验仅针对日本参与者和日语 Pokémon 名称进行。随后，我们使用一项引出实验的结果测试日语随机森林：实验中，日本参与者需要为此前未见过、并与进化前／进化后对应图并列呈现的 Pokémon 命名。此外，我们还使用引出实验中的回答构建了另一个随机森林，并在引出实验回答和官方日语名称上对其进行测试。在实验2中，我们对实验1中的随机森林重新训练，将名称长度纳入考量。重新训练后的随机森林揭示了一个由决策树缺乏变异性导致的过拟合问题。我们通过交叉验证来解决这一问题，具体做法是构建

使用具有不同初始值的多个随机森林（MRF），对数据进行随机化划分，将其分为训练子集和测试子集。随后，将日语 MRF 中随机森林的平均准确率与一项分类实验的结果进行比较；在该实验中，日本参与者被要求将实验1引出的回答归入进化前或进化后类别。接着，将分类实验中人类参与者的结果与 MRF 的结果进行比较。总之，实验1检验随机森林能否利用构成名称的声音学习做出分类判断，以及这种学习是否适用于诱发样本；实验2则测量 MRF 相对于人类的表现。

声音象征

现代语言学理论的一个标准假设是，声音与意义之间的关系是任意的 [1,14]。尽管语言无疑能够以任意方式将声音与意义联系起来，但过去几十年来，越来越多的研究揭示了声音与意义之间存在系统性的关联 [15–17]。一个广为人知的例子是 *takete-maluma* 效应 [18]，即观察到清音阻碍音通常与形状崎岖的物体相关，而含有响音的名称则更常与圆形物体相关。研究表明，这种效应跨语言存在 [19–23]。虽然声音与意义之间的关系可能具有系统性，但其本质通常是随机的 [24]；也就是说，声音与意义的关系表现为一种概率分布，其中存在统计偏斜，但这种关系未必适用于所有词汇项。例如，英语形容词 *tiny*、*mini*和*itsy bitsy*都符合高前

元音与小巧之间的对应关系符合上文讨论的模式，而英语形容词*small* 则明显是这一概括的例外 [11]。声音象征显然对语言习得过程十分重要 [21,25]；象征性词语在面向儿童的言语和早期婴儿言语中都更为常见 [26,27],事实上，研究表明，婴儿对声音象征很敏感 [6–9]。

宝可梦音象学是声音象征学中一个相对较新的分支，研究名为宝可梦的电子游戏角色名称与其属性之间的声音象征关系。在电子游戏中，玩家角色会收集宝可梦，并用它们与其他玩家对战。随着宝可梦获得经验，许多宝可梦可以进化。宝可梦进化会永久改变宝可梦；它们通常会变得更大、更强壮，名称也会改变。宝可梦音象学研究表明，在英语和日语中，宝可梦的进化状态可以通过一些声音象征手段传达，例如名称长度增加、浊阻音的使用增多，以及元音使用上的变化；其中，高前元音 [i] 通常与进化前的宝可梦相关联 [28–31]。基于这些已确立的关系，以及参与者很可能熟悉这一主题，我们认为宝可梦进化适合作为测试案例，用于衡量随机森林与人类理解声音象征的能力（另见 [11]）。

随机森林

随机森林最初由 Breiman 提出 [10],，是一种集成方法机器学习算法，通常用于分类和回归任务。自问世以来，随机森林一直是机器学习领域中的常用工具，近期的几篇综述文章也证明了其有效性[32–34]。通常，随机森林通过使用三分之二的数据构建多棵决策树，再用剩余数据进行测试。决策树本身是一种非参数监督机器学习算法，其结构类似流程图，每个内部节点代表对特征的一项测试。决策树在每个节点上根据各特征对任务的重要性进行划分。划分最终会到达决策树中的一个终端节点，该节点展示决策过程的结果。决策树非常实用；它们具有尺度不变性，能够抵御无关特征的干扰，并且

具有天然的可解释性。然而，决策树对噪声和离群值较为敏感，因此容易出现数据过拟合，从而限制其对未见样本的泛化能力 [35,36]。过拟合是一种建模误差，发生于函数与有限的数据点过度拟合时。这会导致模型在训练数据集上表现良好，但对其他数据集的泛化能力可能较差。为解决过拟合问题，随机森林采用自助聚合（装袋法: [37]）和随机子空间法 [38]。装袋法通过对多个决策树的投票结果（分类任务）或输出（回归任务）取平均值，来提高算法的稳定性和准确率。在装袋法中，样本通常以有放回的方式随机分配给各棵树，因此会产生重复样本。随机子空间法通过在每个内部节点随机选取一部分特征来解决这一问题，使模型通过增加决策树的多样性而更好地泛化。换言之，装袋法随机选择样本，而子空间法随机选择特征。通过在这两个维度上随机化决策树，随机森林解决了决策树固有的过拟合问题。

实验1：诱 elicitation

材料与方法

数据、对数据的解释，以及以下算法的详细注释脚本，均可在 OSF 资料库中获取。

官方宝可梦名称数据。 所有数据均来自 Bulbapedia（[39],最近访问于2022年6月）。截至2022年6月，Bulbapedia 已完成（日常用语中称为“主命名空间”）日语、中文、韩语、英语、德语和法语宝可梦的列表。选择日语、中文和韩语名称进行本实验，是因为日语假名、汉语拼音和韩语麦库恩-赖肖尔罗马化都是相当语音化的书写系统。我们为每种语言创建了一个算法，用于统计每个名称中各个音出现的次数。上述 OSF 资料库中提供了这些算法及其实现方式的详细说明。由此得出的分析几乎完全是音位分析，只有中文声调被视为独立特征，韩语塞音的浊音化也作了单独处理。在韩语 [40] 和中文 [41],中，浊塞音与清塞音之间不存在音系对立。然而，韩语塞音在元音间出现时会系统性地浊音化 [40],，这一点也反映在韩语的麦库恩-赖肖尔罗马化中。鉴于研究已表明，浊塞音在其他语言中包含有关宝可梦进化的信息 [29,31],，因此在韩语中将元音间塞音单独计数。

截至2022年6月，共有905种宝可梦，分属八个世代。本研究仅考察宝可梦进化前和进化后的名称。部分宝可梦不会进化，因此不纳入本研究。核心电子游戏系列的第六世代引入了一种名为超级进化 的机制，可暂时改变某些宝可梦的形态。本研究不考虑超级进化，因为这是一种暂时性形态变化，除了添加*mega*等前缀外，对宝可梦名称影响甚微。其他未纳入分析的宝可梦包括中间阶段进化形态。电击兽 就是中间阶段宝可梦的一个例子，它于电子游戏系列的第一世代引入。它的进化前形态电击怪于第二世代引入，而进化后形态电击魔兽则于第四世代引入。本研究将电击兽 排除在分析之外，因为它被视为中间阶段形态，尽管其他宝可梦后来被加入进化

追溯到整个家族。Kawahara 和 Kumagai [28]分析了 Pokémon 名称中的声音与 Pokémon 进化之间的关系，其中没有排除处于中间阶段的 Pokémon。为此，他们根据进化等级而非进化前和进化后的二元类别，将数据分为四个类别。随机森林能够进行多类别分类；然而，我们在当前分析中选择了二元分类，因为尽管这些数据在技术上属于计数数据，但几乎完全是二元的（例如，日语数据集中 96.7% 的所有数据点要么是 0，要么是 1）。因此，鉴于声音象征模式很可能在中间阶段和最终阶段类别之间呈标量变化，使用二元分类器是合理的。删除中间阶段的 Pokémon 以及没有进化家族的 Pokémon 后，共得到 628 个独特的 Pokémon 名称，其中 303 个属于进化前，325 个属于进化后。分布偏斜的原因在于，某些进化前的 Pokémon 可能会进化为多个进化后变体。

引出实验。 本实验已获名古屋商科大学伦理审查批准。编号：21048。

引出实验有两个主要目标。第一个目标是确定，使用官方宝可梦名称数据构建的随机森林，能否用于分类从参与者那里引出的名称，反之亦然。换言之，官方名称与参与者提供的名称之间是否有足够的重合，使每个模型都能用于对来自另一数据集的宝可梦进行分类。第二个目标是为分类实验（实验2）提供刺激材料，以测量人类参与者与机器学习算法的表现。为了公平地测量分类准确率，

重要的是要在人类和机器学习算法都未曾接触过的数据上进行测试，因此需要引出的样本。

引出实验使用 Google Forms 进行。每份 Google Forms 问卷都包含一段简短的说明，随后是二十张宝可梦风格的图片。'按照 Kawahara与Kumagai [28],所述的方法，这些图片并非现有的宝可梦形象，而且参与者很可能此前从未见过它们。说明中指出，只有日语母语者可以参与调查。参与者获知，他们需要为二十只新的宝可梦命名。参与者明确获知，他们将看到进化前和进化后的宝可梦图片。参与者需要用片假名为宝可梦命名，因为片假名是日语中书写宝可梦名称和临时词所用的文字。参与者被要求不要使用现有词语（日语或其他语言）为宝可梦命名。关于宝可梦命名，参与者没有收到其他说明（例如长度限制）。调查开始前，没有询问参与者是否熟悉宝可梦系列。所有说明均以日语撰写。参与者获知，他们完全自愿参加，并且可以随时退出调查。研究人员以口头方式征得同意，并向参与者说明，参加调查本身即表示同意。除用于确保学生不会重复填写调查的学生邮箱地址外，未收集任何个人数据。分析前，这些邮箱地址均已删除。

每张图像都包含一只进化前和一只进化后的宝可梦，并排呈现。进化前的宝可梦总是位于进化后宝可梦的左侧，而且显示得明显更小（见图 1），与其进化后的对应宝可梦相比尤为如此。每张图像中都有一个箭头指向要命名的宝可梦。箭头指向进化前宝可梦的图像后面，总会紧接着一张相同的图像，唯一不同的是箭头指向进化后的宝可梦。试次并未随机排序，进化前的图像总是先于进化后的图像呈现。进化前的宝可梦总是呈现在左侧，进化后的宝可梦则总是

呈现在右侧。图像由一位半职业艺术家（Devian- tArt 用户：Involuntary-Twitch）创作，样本见图 1。这些图像与早期几代宝可梦游戏中用于呈现宝可梦的像素化图像非常相似。

参与者通过在学生公告栏发布通知，从名古屋商科大学招募。学生无论以金钱或其他形式，均未因参与而获得报酬。人类参与者需要对相关主题有一定了解，因为虚构名称中的音象征关系可能不遵循自然语言中的规律。鉴于宝可梦在日本广受欢迎，且参与者是日本大学生，因此选取宝可梦作为测试案例，以评估随机森林与人类的准确率。共有49名学生参与调查。总计记录了980份回答；然而，其中一些回答为空，另一些则包含重复名称，其分布情况表明参与者在填写调查时可能相互商量过。以上回答均予以剔除，最终得到967个独一无二的名称（进化前482个；进化后485个）。诱发名称采用与官方日语宝可梦名称相同的算法进行转写。诱发实验收集到的名称中，没有一个是现有宝可梦的名称。

随机森林。使用 Ranger 软件包 0.13.1 [42] 构建并测试了随机森林。每个随机森林所包含的树的数量通过在不同树数量取值下、使用不同随机化起始点（set.seed）构建九个随机森林进行手动调参。通过考察袋外平均准确率及其标准差确定最优取值。袋外误差是指错误分类的样本。对于所有随机森林，20,000 棵树



图1. 实验1中使用的进化前和进化后宝可梦角色的示例刺激对。经艺术家许可复制这些图像。

https://doi.org/10.1371/journal.pone.0279350.g001

被确定为合适的数量，因为我们观察到增加树的数量并不会降低袋外误差，而且使用 Altmann 方法计算特征重要性时 [43]在 20,000 棵树的情况下已接近构建随机森林所用计算机的处理能力。与每个节点所考察的特征数量、样本比例和节点大小相关的超参数使用 tuneRanger 软件包 05 [44]进行调参。实质上，调参过程确定了树之间存在多大程度的变异性。变异性很高的树

会产生高度可变的结果，但在处理包含许多不重要特征或空值的数据集时，可能会遇到问题。决策树的低变异性会产生更稳定的算法，但可能掩盖较弱特征的重要性，因为它们往往会与强特征配对。随机森林的准确率通过将测试数据输入模型并评估袋外误差来确定。袋外误差总体反映了算法的准确率，但不会说明哪些特征对分类很重要；这需要通过特征重要性来确定。计算特征重要性有多种方法，本研究采用置换法。在置换法中，逐一随机打乱每个特征，然后在其他所有特征保持不变的情况下重建算法。特征重要性根据随机化导致的袋外误差增幅计算。随机森林可解释性面临的一个问题是，特征重要性无法说明方向性。例如，对分类很重要的声音可以视为将每个样本“拉”向某一类别；特征重要性反映了这种“拉力”的强度，却无法说明这种“拉力”是指向进化前还是演化后类别。本研究报告了语音在进化前后类别中的分布，以表示方向性；但应注意，这些指标未必相同。

实验1共构建了六个随机森林。结果部分介绍的前三个随机森林使用随机抽取的子集进行训练，该子集包含三分之二的日语、中文和韩语宝可梦名称。第四个随机森林使用引出实验结果的三分之二进行训练。随后，使用每个数据集剩余的三分之一子集测试这四个随机森林。接着，我们计算每个随机森林的特征重要性，以考察潜在的跨语言模式，以及日语宝可梦数据与引出数据之间的模式。其余两个随机森林分别使用全部官方日语宝可梦名称和引出实验收集的全部样本构建。随后，使用相反数据集测试这两个随机森林。换言之，一个随机森林使用全部官方

名称构建，并使用引出的回答进行测试；另一个则使用全部引出样本构建，并使用官方名称进行测试。这样做是为了确定两个数据集之间是否有足够的重合度，使算法能够有效地对相反数据集进行分类。

结果

在官方宝可梦名称上训练和测试的三个随机森林，都以高于随机水平的准确率对宝可梦进行了分类。鉴于进化前和进化后的宝可梦分布不均，任何仅凭直觉将样本归入多数类别的模型，其准确率都会约为52%（袋外误差为48%），具体取决于训练子集和测试子集的划分方式。日语随机森林的准确率最高（袋外误差为29.05%），其次是中文随机森林（袋外误差为39.05%），最后是韩语随机森林（袋外误差为40.95%）。日语随机森林的混淆矩阵见表1，其特征重要性见表2。请注意，在实验2中，我们报告了随机化初始值不同的多随机森林的结果；这些随机森林在划分训练数据和测试数据时采用了不同的初始值

表1. 日语随机森林的混淆矩阵。

		分类	
		进化前	进化后
样本	进化前	69	38
	进化后	23	80

https://doi.org/10.1371/journal.pone.0279350.t001

表2. 日语随机森林中，特征重要性大于0.1%的特征及其特征重要性（Importance）和*p* 值。

特征	重要性	*p* 值
/m/	0.78%	0.030
/ɴ/[a]	0.63%	0.020
/:/[b]	0.45%	0.049
/g/	0.40%	0.049
/a/	0.39%	0.139
/ɾ/	0.35%	0.089
/Q/[c]	0.24%	0.059
/ɸ/	0.20%	0.079
⁀/tɕ/	0.19%	0.069
/d͡ʒ/	0.19%	0.129
/d/	0.19%	0.228

[a] /ɴ/ 表示韵尾鼻音。[b] /:/ 表示长元音的后半部分。[c] /Q/ 表示叠音辅音的起始部分。

https://doi.org/10.1371/journal.pone.0279350.t002

subset子集d the RFs themselves. Th结果 lts of the MRFs (OOB 误差.M = 34.07%，标准差 = 2.48%），这表明该结果是由训练子集和测试子集之间一次特别有利的划分所导致的离群值。中文数据集也采用了这一流程（袋外误差：均值 = 40.85%，标准差 = 3.35%），韩语数据集亦然（袋外误差：均值 = 43.28%，标准差 = 3.09%）。在诱发名称上训练和测试的随机森林（诱发随机森林）对样本的分类准确率高于偶然水平。与官方数据集一样，各类别的分布并不均衡；若采用朴素模型，在引出数据中有 50.16% 的样本（袋外误差为 49.84%）能够被正确分类。诱发随机森林的袋外误差为 30.96%。我们计算了每个模型的特征重要性，以确定哪些声音有助于分类。特征重要性及显著性通过训练子集上的 Altmann [43]置换法计算。置换是指逐个随机打乱特征；随后针对每个特征重新构建随机森林。特征重要性指被随机打乱的特征所导致的袋外误差增幅。Altmann置换法通过执行多次置换来估算更精确的*p*值。特征重要性的显著性则通过置换检验对有偏测量指标进行归一化计算。该方法会为每个特征得出显著性结果，而不是为随机森林本身得出结果 [43]。本研究中的所有随机森林均采用 Altmann置换法，迭代次数设为 100。方向性根据数据训练子集中各特征的分布确定。日语训练子集中特征的分布见图2。在日语随机森林中，最重要的特征是双唇鼻音（/m/）、韵尾鼻音（/ɴ/）、长元音（/:/）和浊软腭塞音（/g/）。在这些特征中，只有 /m/ 在演变前样本中的出现频率更高。

与日语随机森林相同，中文随机森林中大多数重要特征的分布偏向演化后类别。中文随机森林的混淆矩阵见表3，其特征的特征重要性得分见表4，分布见图3。声调是随机森林中的一项重要特征；降调在进化后样本中出现得更频繁，而轻声在进化前样本中出现得更

频繁。软腭鼻音（/ŋ/）也被发现是中文随机森林中的一项重要特征。

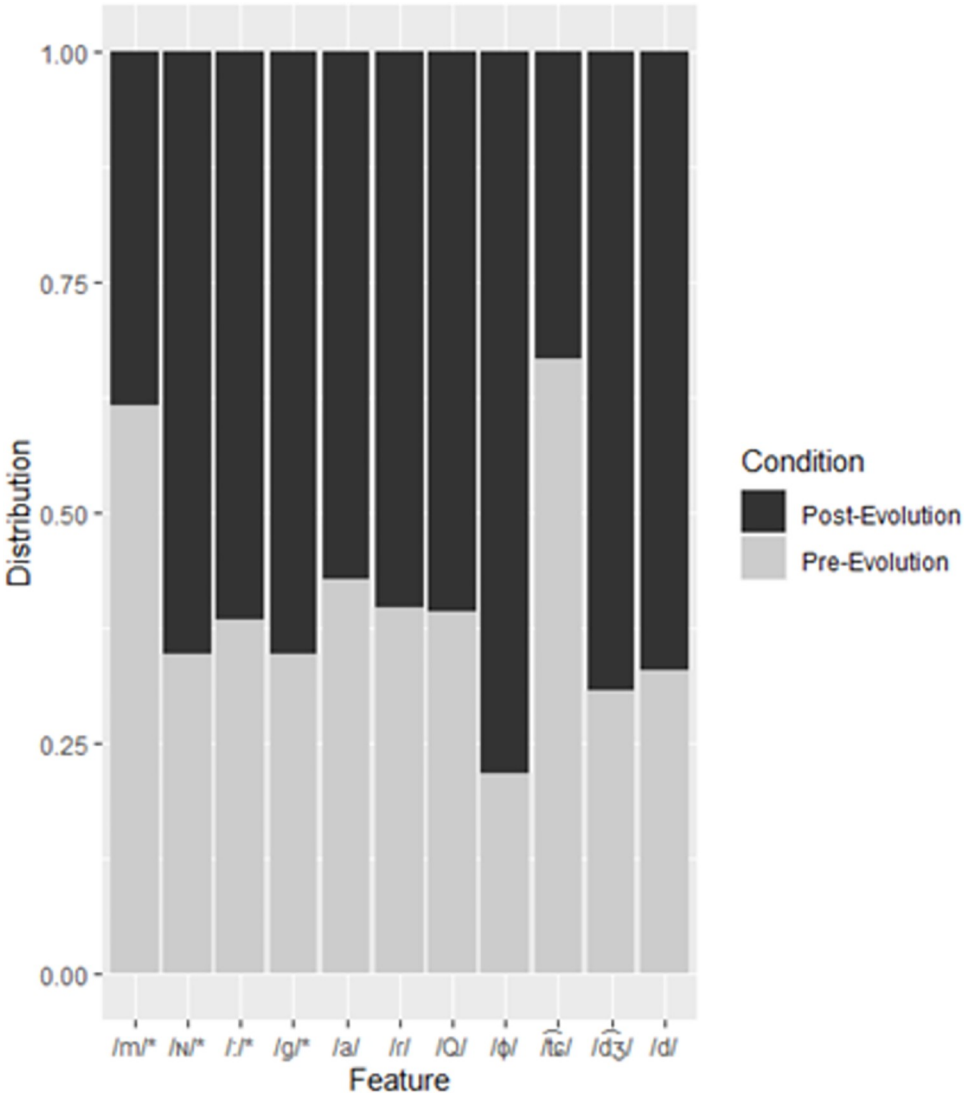

图2. 日语训练子集中进化前后类别的特征分布。特征按重要性从左到右排列。星号表示显著特征。

https://doi.org/10.1371/journal.pone.0279350.g002

在韩语随机森林中，元音 /ɯ/、/a/、/ʌ/ 和 /u/ 都很重要，浊唇软腭近音 /w/ 也同样重要。有趣的是，虽然闭后不圆唇元音 /ɯ/ 在进化后样本中出现得更频繁，但闭后圆唇元音 /u/ 在演化前样本中出现得更频繁。表5展示了韩语随机森林的混淆矩阵，表6展示了特征重要性和 p 值，图4 展示了分布。

表3. 中文随机森林中，特征重要性超过 0.1% 的特征的重要性(Importance)和*p* 值。

		分类	
		进化前	进化后
样本	进化前	53	50
	进化后	29	78

https://doi.org/10.1371/journal.pone.0279350.t003

表4. 中文随机森林中特征重要性大于0.1%的特征的重要性（Importance）和*p*值。

特征	重要性	*p*值
降调	0.88%	0.020
/ŋ/	0.87%	0.010
⁀/tɕ/	0.20%	0.089
/ɕ/	0.14%	0.109
/e/	0.13%	0.238
/o/	0.13%	0.257
轻声	0.13%	0.188

https://doi.org/10.1371/journal.pone.0279350.t004

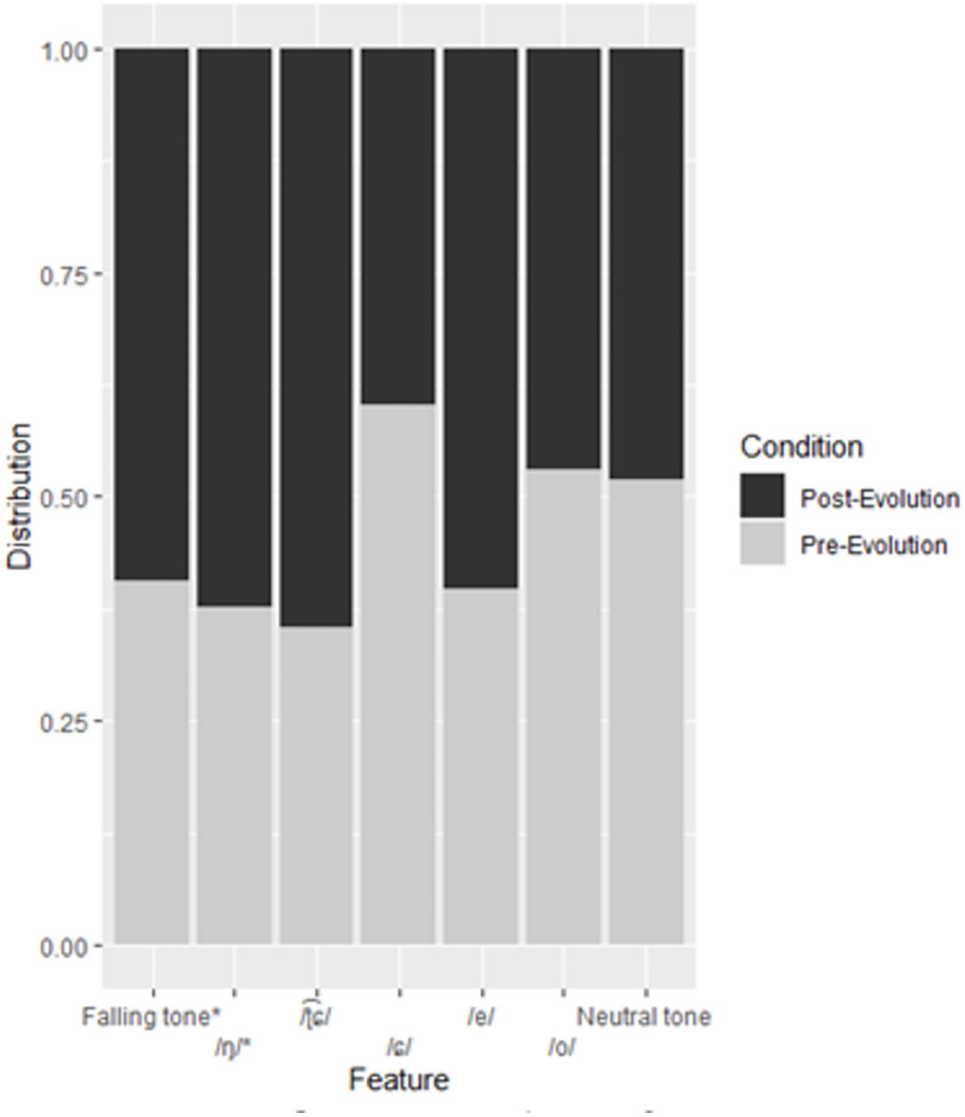

图3. 中文训练子集中进化前后类别的特征分布。特征从左到右按重要性递减排列。星号表示显著特征。

https://doi.org/10.1371/journal.pone.0279350.g003

表5. 韩语随机森林的混淆矩阵。

		分类	
		进化前	进化后
样本	进化前	57	50
	进化后	36	67

https://doi.org/10.1371/journal.pone.0279350.t005

表6。韩语随机森林中，特征重要性大于 0.1% 的特征的重要性（Imp.）和 p值（p）。

特征	重要性	*p* 值
/ɯ/	2.558%	<0.001
/a/	1.326%	0.0297
/w/	0.226%	0.0693
/ʌ/	0.207%	0.1287
/u/	0.113%	0.1782

https://doi.org/10.1371/journal.pone.0279350.t006

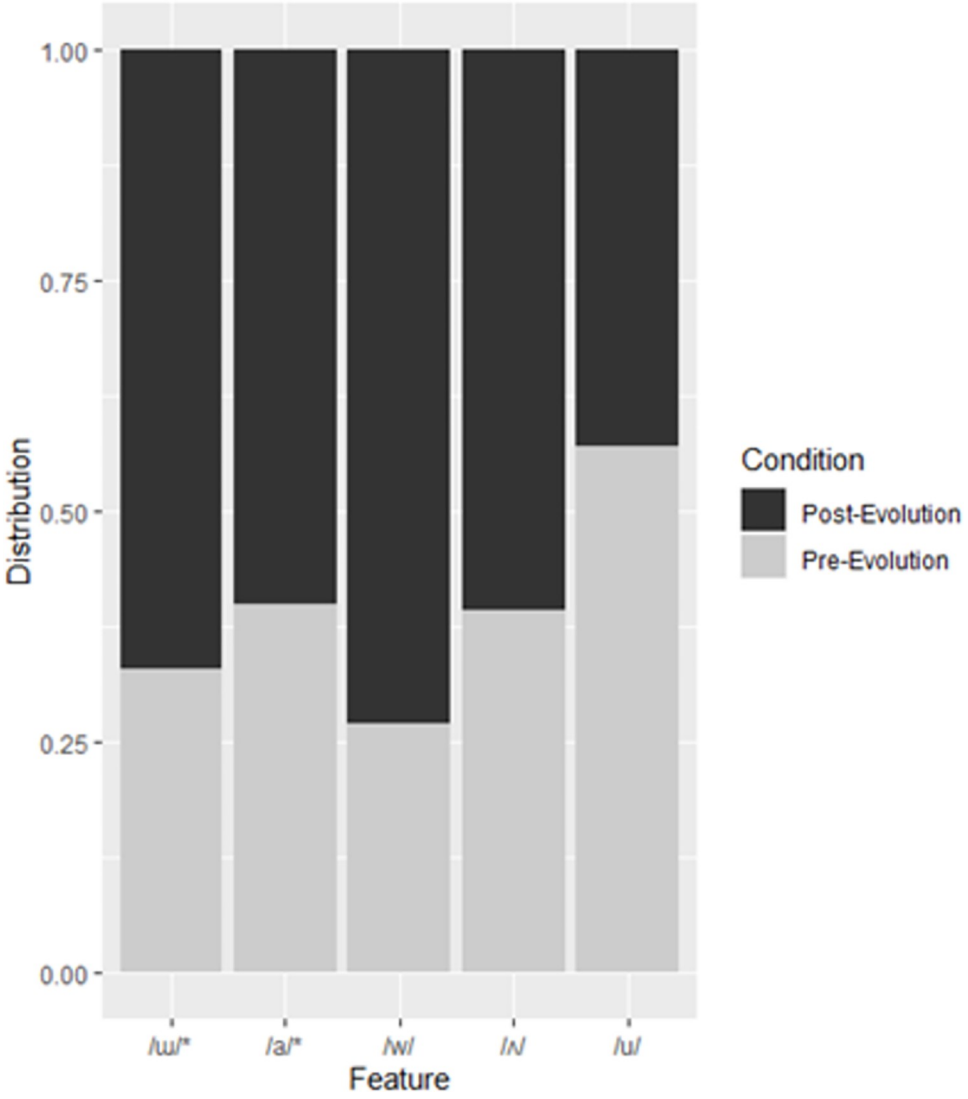

图4。韩语训练子集中进化前后类别的特征分布。特征按照重要性从左到右排列。星号表示具有显著性的特征。

https://doi.org/10.1371/journal.pone.0279350.g004

表7. 诱发随机森林的混淆矩阵。

		分类	
		进化前	进化后
样本	进化前	103	55
	进化后	45	120

https://doi.org/10.1371/journal.pone.0279350.t007

表 8. 诱发随机森林中重要性大于 0.1% 的特征的重要性（Imp.）和 p 值（p）。

特征	重要性	*p* 值
/d/	1.263%	<0.001
/ɯ/	1.059%	<0.001
/a/	0.918%	<0.001
/g/	0.838%	<0.001
/d͡ʒ/	0.448%	<0.001
/o/	0.219%	0.2277
/ɴ/	0.180%	0.2376
/:/	0.140%	0.2772
/z/	0.138%	0.1287

https://doi.org/10.1371/journal.pone.0279350.t008

日语随机森林中重要的大多数特征在诱发随机森林中也同样重要。这些特征包括浊塞音（/g/ 和 /d/）、开前不圆唇元音（/a/）、韵尾鼻音（/ɴ/）以及长元音（/:/）。有趣的是，在诱发随机森林中特征重要性得分高于 0.1% 的所有特征，在进化后 Pokémon 中出现得更频繁。根据诱发实验数据构建并测试的随机森林的混淆矩阵列于表 7 和 8展示了特征重要性得分，而图5 展示了分布图。

鉴于日语随机森林和诱发随机森林的特征重要性模式相当相似，我们想要检验这些随机森林能否准确地对来自相反数据集的样本进行分类。两个模型共有的重要特征包括非唇音浊阻音，如 /d/ 和 /g/、韵尾鼻音、长元音以及低前元音 /a/。有趣的是，所有这些特征的分布偏向演化后类别。我们使用每个现有随机森林对其相反数据集的全部数据进行测试（而非仅测试子集）。日语随机森林对诱发样本的分类准确率为 61.43%（袋外误差为 38.57%）；诱发随机森林对官方日语 Pokémon 名称样本的分类准确率为 66.72%（袋外误差为 33.28%）。相比之下，朴素模型的准确率预计分别为 52% 和 50.16%。使用官方日语 Pokémon 名称训练、并使用诱发样本测试的随机森林混淆矩阵见表9。表10显示了使用诱发样本训练、并使用官方日语 Pokémon 名称测试的随机森林混淆矩阵。

讨论

以上所有随机森林的表现都优于朴素算法。对于日语、中文和韩语随机森林，朴素算法的袋外误差预计为 48%。日语随机森林最为准确（袋外误差为 29.05%），而中文（袋外误差为 39.05%）和韩语（袋外误差为 40.95%）的误差率也远低于 48%。对于诱发随机森林，朴素算法的袋外误差预计为 50%，而其实际袋外误差为 30.96%。需要记住的是，这些随机森林

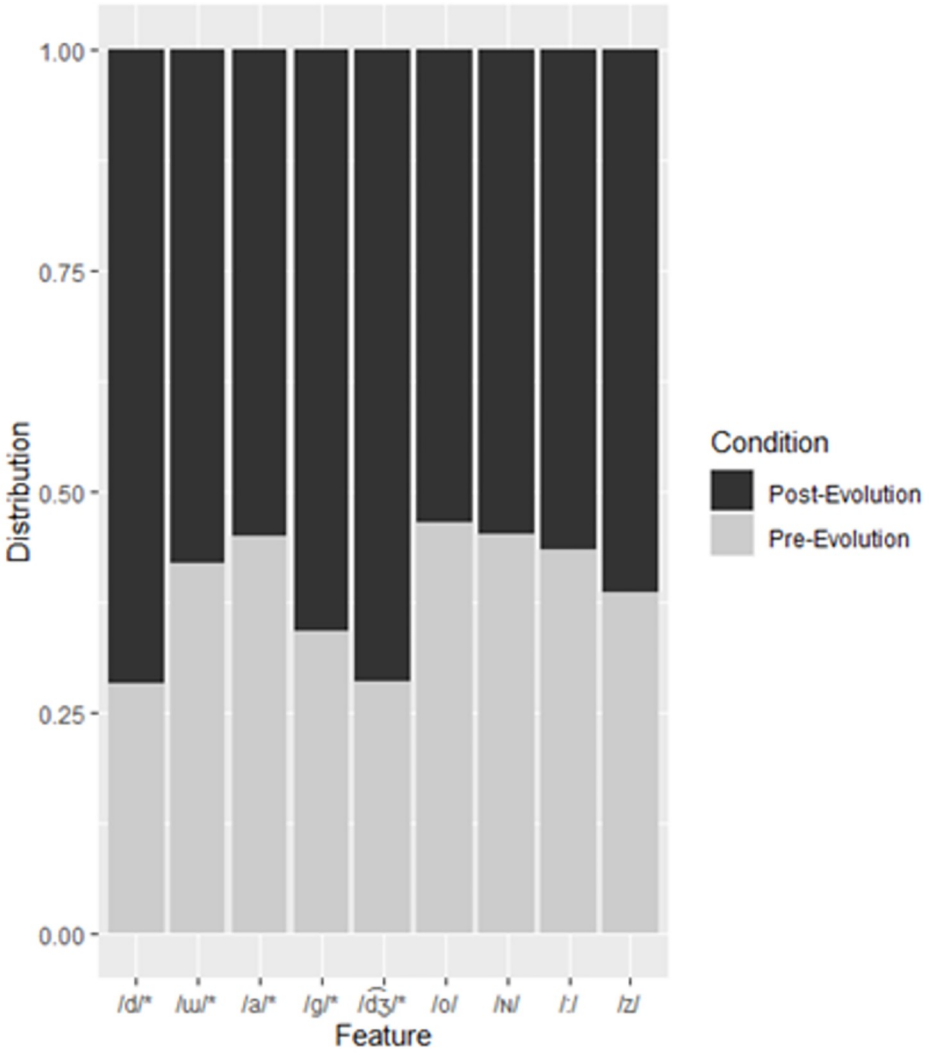

图5. 诱发训练子集中各特征在进化前后类别中的分布。特征按重要性从左至右排列。星号表示显著特征。

https://doi.org/10.1371/journal.pone.0279350.g005

仅使用三分之二的数据进行训练，因此，考虑到它们只有 419 个样本可供学习，这些随机森林仍是高效的学习器。使用日语官方数据训练、并用引出数据测试的随机森林，使用全部 628 个日语官方样本进行训练，并用全部 967 个引出的回答进行测试。尽管可用于学习的样本更多，但使用官方名称训练、并用引出的回答测试的随机森林（袋外误差为 38.57%），准确度低于使用官方名称进行训练和测试的随机森林（袋外误差为 29.05%）。同样，使用..... 训练的随机森林

表9. 使用官方日语 Pokémon 名称训练、并使用诱发样本测试得到的随机森林混淆矩阵。

		分类	
		进化前	进化后
样本	进化前	245	237
	进化后	136	349

https://doi.org/10.1371/journal.pone.0279350.t009

表10. 使用诱发样本训练并使用官方日文宝可梦名称测试的随机森林混淆矩阵。

		分类	
		进化前	进化后
样本	进化前	180	123
	进化后	83	242

https://doi.org/10.1371/journal.pone.0279350.t010

使用全部引出的回答进行训练、并在官方名称上进行测试的随机森林（袋外误差为33.28%），准确度低于使用引出的回答进行训练和测试的随机森林（袋外误差为 30.96%）。尽管存在这些差异，官方名称与引出的回答仍足够相似，因此表现优于简单算法。

随机森林的特征重要性得分揭示了宝可梦进化状态与构成其名称的声音之间有趣的关系，其中一些关系在不同语言中都成立。虽然高前元音在任何随机森林中的特征重要性都未超过 0.1%，但低前元音 /a/ 和高后不圆唇元音 /ɯ/ 在日语、韩语和诱发随机森林中都很重要，而且在所有情况下，其分布都偏向进化后类别。音位 /a/ 表示进化后的宝可梦，这一结果与一个广为人知的观察相符：包含 [a] 的临时词比包含 [i] [2,45,46] 的临时词更大，因为进化后的宝可梦通常比其进化前的对应形态更大。有趣的是，高后圆唇元音 /u/ 在中文模型中很重要，但其分布偏向进化前类别。研究发现，元音在韩语模型中很重要，尤其是 /ɯ/、/a/、/ʌ/ 和 /u/。研究发现，韩语元音在“明”元音与“暗”元音之间存在声音象征对应关系 [47]。这些对应关系与跨语言模式相悖。例如，明元音被定义为低元音，据称能反映体型较小、移动较快的实体；而暗（或高）元音则据称能反映体型较大、移动较慢的实体 [48]。我们的研究结果不支持这一观察。尽管暗元音 /ɯ/ 和 /ʌ/ 的分布偏向进化后类别，明元音 /a/ 的分布也偏向进化后类别，而暗元音 /u/ 的分布则偏向进化前类别。/a/ 对韩语模型很重要且分布偏向进化后类别这一发现，与 [5] 的研究结果一致；该研究发现，韩语听者认为包含 [a] 的临时词更大。长元音在日语和诱发随机森林中都很重要，而且在两种情况下，其分布都偏向进化后类别。这一发现也体现在以往的宝可梦研究中[29,30]，这些研究也表明长元音与增大有关

大小。这些研究倾向于认为，这可以用数量象似性来解释，即较大的物体通常与较长的名称相关联 [49]。实验2将进一步探讨这一点。最后，中文随机森林中的声调对模型很重要。降调在中文随机森林中的特征重要性最高，并且倾向于归入演化后类别。另一方面，轻声则倾向于归入进化前类别。在一项类似的宝可梦研究中，Shih 等人 [50]发现，降调似乎与力量增强、进化阶段以及雄性比例增加相关。这似乎比 Ohala的频率编码假说 [51]所预测的更为复杂，因为该假说仅指出低声调应体现较大，而高声调应体现较小，却未对声调的基频轮廓作出预测。Shih 等人 [50]提出，降调在所有汉语声调中的基频变化幅度最大，这或许可以解释为什么这种声调在中文中会象征性地与大相联系。

日语鼻音 /ɴ/ 和汉语鼻音 /ŋ/ 在日语、中文和诱发随机森林中都很重要，并且在所有情况下都偏向进化后。这是一个有趣的发现

鉴于这两种辅音只能出现在韵尾位置，尽管韩语中的韵尾鼻音 /ŋ/ 的特征重要性并未超过 0.1%。跨语言来看，鼻辅音通常与大型实体相关联 [2,52],，这可能是由于它们的低频能量所致 [2]。然而，在日语中，双唇辅音已被发现与可爱、柔和的形象相关联 [53],这或许可以解释为什么 /m/ 在日语模型中既很重要，又偏向进化前类别。通过考察日语中可爱度与唇音性之间的关系，韩语模型中的高后元音呈现出一个有趣的案例。在韩语模型中，两种高后元音都被发现很重要。高后圆唇元音 /u/ 偏向进化前类别，而高后不圆唇元音 /ɯ/ 则偏向进化后类别。这一结果表明，可爱度与唇音性之间的关联或许是跨语言的；不过，这一推测尚不确定，因为韩语中的唇音-软腭近音也偏向进化后，并且在模型中很重要。Berlin [2]认为，鉴于鼻辅音的低频能量，它们可能暗示体型较大；然而，双唇鼻音 /m/ 在日语随机森林中被发现很重要，却偏向进化前类别。与 Shih 等人 [31],的发现一致，他们发现浊塞音能够体现日语和英语宝可梦名称中的体型，浊塞音 /d/ 和 /g/ 在日语和诱发随机森林中都很重要。韩语中的元音间塞音因这些位置保留了系统性的浊音化而被单独统计；然而，它们的特征重要性并未超过 0.1 %。

实验2：分类

随机森林

实验1中介绍的随机森林仅使用构成宝可梦名称的声音来构建。实验2中，我们为这些随机森林加入名称长度这一额外特征并重新构建。此前的随机森林未纳入长度，因为先前研究表明，长度很可能是一项非常重要的特征 [29,31],，若将其纳入，很可能会掩盖其他特征的重要性。除此之外，实验2中的随机森林与实验1采用相同的方法，但多个随机森林（MRF）彼此独立构建的情况除外

。在 MRF 中，用于将数据划分为训练子集和测试子集的随机化起始值，以及每个随机森林的随机化起始值，均设为该随机森林在序列中的序号。因此，每个 MRF 中的第一个随机森林的 set.seed 值为 1，而第九个随机森林的 set.seed 值为 9。对于 MRF，只对第一个随机森林进行调参，并将得到的超参数设置应用于每个 MRF 中的全部九个随机森林。这是因为，据我们所知，无法使调参过程具有可复现性，而且各个随机森林的数据来自同一来源。我们使用每次随机划分所得的测试子集，对每个 MRF 进行九次测试。当使用一个数据集的全部数据测试另一个数据集的全部数据时（例如，测试官方日语 MRF 对引诱日语 MRF 的准确率），我们会针对每项测试运行九次迭代，并为 MRF 使用不同的随机化起始值。

分类实验

本实验已获名古屋商科大学伦理审查批准。编号：21057。

在分类实验中，我们取用实验1中引出的回答，请日本参与者将其分类为进化前或进化后宝可梦。我们从967条引出的回答中随机抽取了100个样本。

随机抽样过程中没有控制分布，因为随机森林中子集的划分并不控制分布。这些样本被用于制作五份 Google Forms，每份包含二十个诱发名称。采用这种设计，而不是从整个数据集中随机抽取二十个名称，是为了模拟随机森林中决策树进行的投票过程。下文的结果部分将对此作进一步讨论。表单用日语向参与者说明，他们需要将新的宝可梦归入进化前或进化后类别。选项以按钮形式呈现，标签分别为進化前 [进化前] 和進化後 [进化后]。问卷完成前，没有询问参与者是否熟悉宝可梦系列。我们为五份 Google Forms 分别生成了二维码。这些二维码被打印在传单上，并分发给爱知县立大学和名古屋商科大学的日本大学生。除二维码外，传单上没有其他信息，只有标题ポケモンクイズ [宝可梦问答]。传单在社团活动和排定的课程开始前发给学生。学生没有在课堂上获得填写问卷的时间。共有119名参与者回应了问卷，其中有10次未能指定类别，因此总计获得2,370条回答。与实验1一样，参与者被告知参与完全自愿，且可以随时退出问卷。参与者以口头方式表示同意。除学生电子邮件地址外，未收集任何个人数据；收集电子邮件地址是为了确保学生没有重复填写问卷。分析前，这些地址均已删除。我们要求参加过实验1的学生不要参加实验2。

结果

实验2旨在比较随机森林与人类在将宝可梦分类为进化前和进化后类别时的表现。在分类实验中，人类参与者可以获取名称长度信息，因此算法中也纳入了长度，以便随机森林能够利用这一信息。下文将考察四个数据集中进化前后名称长度的分布。随后，重新构建此前所有随机森林

，以纳入长度这一因素，检验其对袋外误差的影响。我们还计算长度的特征重要性，以确定其对袋外误差的贡献程度。最后，我们报告分类实验的结果，并比较人类参与者与机器学习算法的准确率。除汉语声调外，各数据集的长度均按其中所有声音的总数计算。在对宝可梦名称中的声音象征关系进行探索时，Kawahara 和 Kumagai [28]将日语名称的长度按莫拉数计算，因为音拍是心理语言学上最显著的韵律单位 [54]。尽管决策树具有尺度不变性，我们仍按特征数计算长度，以使日语长度参数与中文和韩语参数保持一致。计算长度时排除了汉语声调，因为声调是基频轮廓的测量指标，不像其他语音那样会增加名称的整体长度。尽管如此，中文名称仍比其他所有数据集中的名称都长：进化前（均值 = 8.76，标准差 = 2.09）和进化后（均值 = 9.31，标准差 = 2.09）的宝可梦名称的声音数中位数均为九个。日语、韩语和诱发数据集中的长度相近：所有进化前名称的声音数中位数均为七个，所有进化后名称的声音数中位数均为八个。进化前后长度的均值差异在诱发数据集中最大（1.52），其次是日语数据集（0.9）、韩语数据集（0.73），最后是中文数据集（0.55）。各数据集中长度的均值、中位数和标准差见表11。图6以箱线图呈现了中文、日语和韩语名称按进化状态划分的长度。

表11. 韩语随机森林中特征重要性大于0.1%的特征的重要性（Imp.）和 p值（p）。

《语言》	测量指标	进化前	进化后
中文	中位数	9	9
	Mean	8.76	9.31
	标准差	2.09	2.09
日语	中位数	7	8
	Mean	7.28	8.18
	标准差	1.38	1.26
韩语	中位数	7	8
	Mean	7.33	8.06
	标准差	1.88	1.81
诱发	中位数	7	8
	Mean	6.79	8.31
	标准差	2.07	2.38

https://doi.org/10.1371/journal.pone.0279350.t011

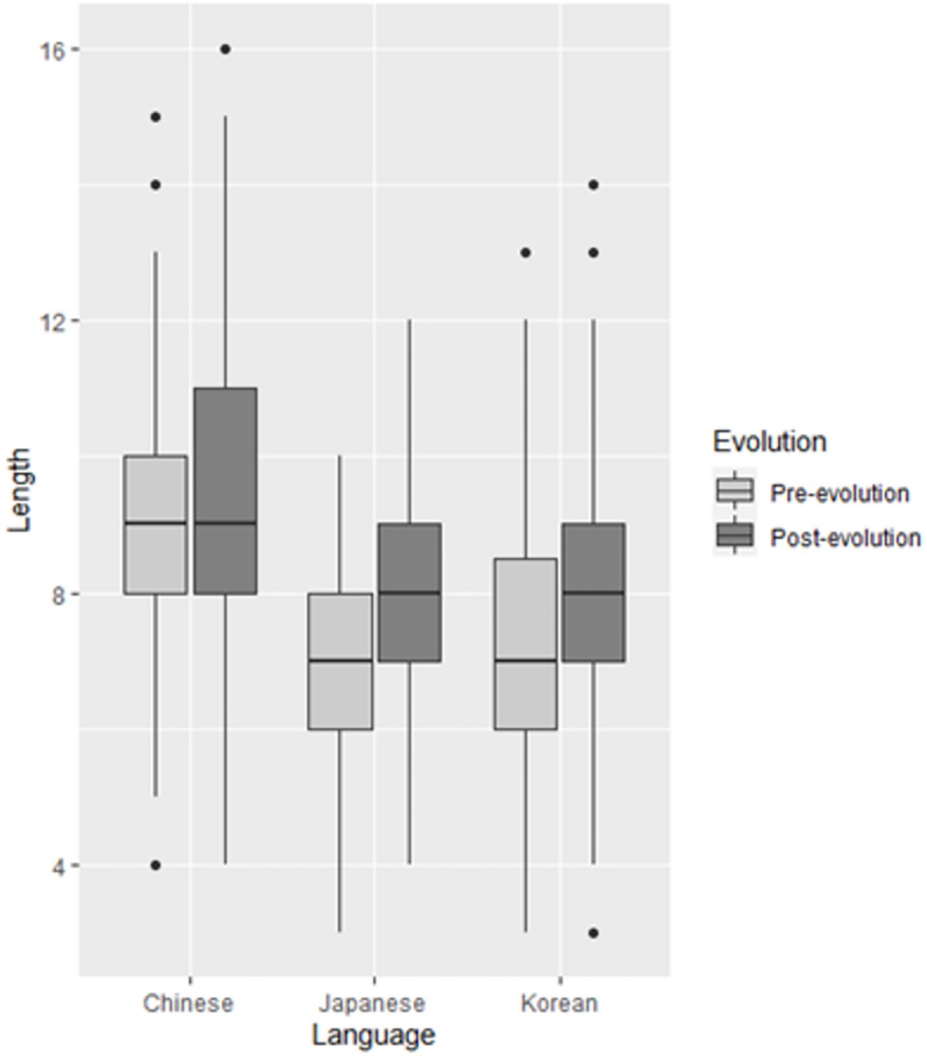

图6. 三种语言中进化前和进化后Pokémon的长度箱线图。

https://doi.org/10.1371/journal.pone.0279350.g006

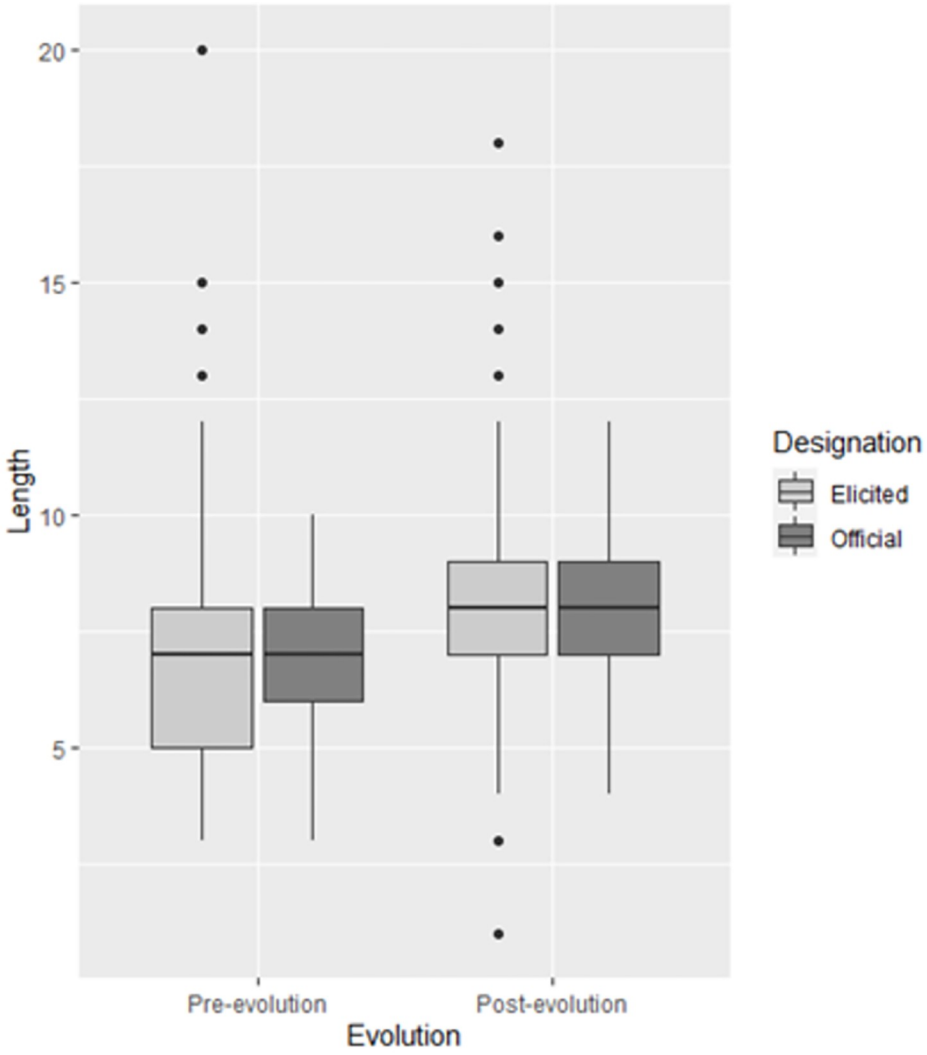

图 7. 官方日语宝可梦名称长度（官方）与引出名称长度（引出）的箱线图并列展示。

https://doi.org/10.1371/journal.pone.0279350.g007

官方日语宝可梦数据集与引出数据集的长度分布极为相似。尽管引出数据集中有更多离群值，但中位数、上四分位数和下四分位数，以及最小值和最大值（不包括离群值）几乎完全相同。图 7并列展示了官方日语宝可梦名称长度和引出名称长度的箱线图，以说明这些相似之处。

上一节的随机森林中未纳入长度，因为先前的研究已明确表明，长度会是一个重要特征 [29,31]，且很可能会掩盖语音的重要性。因此，纳入长度理应提高模型的准确率（或降低袋外误差）。然而，事实并非如此。我们重新构建了所有先前的随机森林，将长度特征纳入其中。这些随机森林按照实验1中所述的相同调参过程进行了调参。长度的特征重要性见表12。 表12 还列出了纳入长度（+L）和未纳入长度（-L）的随机森林的袋外误差率。纳入长度后，除中文随机森林外，所有随机森林的袋外误差均有所增加。由于进化前后宝可梦的平均长度差异最小，中文随机森林理应是受长度影响最小的。

表12. 实验1中构建的随机森林的袋外误差率（-L OOB）、使用长度构建的随机森林的袋外误差（+L OOB），以及长度在这些随机森林中的特征重要性（L Imp）。

RF	-L OOB[a]	+L 袋外[b]	L Imp[c]
中文	41.90%	41.43%	0.25%
日语	29.05%	30.95%	4.74%
韩语	40.95%	41.43%	1.33%
诱发	30.34%	30.96%	6.66%

https://doi.org/10.1371/journal.pone.0279350.t012

在中文数据集中最小。令人困惑的是，长度被证明是随机森林中的一个重要特征，但其影响并未体现在 +L与-L随机森林之间的袋外误差差异上。

为了探究这一现象的潜在解释，我们考察了构建随机森林时所采用的随机化过程。此前所有随机森林使用的 set.seed 值都相同，只有在调整每个森林的树的数量（num.trees）时例外。该值用作生成随机数的起始值，既用于划分训练数据和测试数据，也用于随机森林本身。在实验1的方法部分，我们通过对每个 num.trees 值运行九次迭代，并为随机森林的随机化使用不同的 set.seed 值，来调整 num.trees。我们将这一方法应用于训练子集和测试子集之间的划分随机化，发现袋外误差存在相当大的变异。我们对本研究中展示的每个随机森林都运行了九次迭代。不过，在此处，我们调整了子集划分和随机森林各自的 set.seed 值。对于 L 和 -L 随机森林，set.seed 值均为 1–9，因此产生了相同的九种子集划分。这些结果见 +表13。调参过程中的随机化无法控制；每次进行调参时，即使使用相同的数据，也会返回不同的超参数值。鉴于数据的性质保持不变，MRF 使用的超参数值取自此前的随机森林。

为了了解实验1中的日语随机森林为何取得如此低的袋外误差，我们考察了 -L 日语 MRF 中各特征的平均特征重要性值，并将其与 -L 日语随机森林中各特征的重要性进行比较。表 14展示了日语 MRF 的混淆矩阵。表 15 列出了日语随机森林中的特征重要性值，以及日语 MRF 中的平均特征重要性值。由此可见，实验1中所述的日语随机森林过分强调了特征 /m/、/ɸ/ 和 /Q/ 的重要性，却低估了 /ɴ/、/ː/、/ɾ/、/d/、/ɯ/、/o/ 和 /s/ 的重要性。由于后三个特征在日语随机森林中的特征重要性均未超过 0.1%，因此未在前面的表格和图表中列出。

鉴于子集的随机化对袋外误差有如此大的影响，我们重新进行了测试

使用引出数据训练的日语随机森林，以及使用官方日语 Pokémon 数据训练的诱发随机森林，分别采用 -L 和 +L 数据集。我们对整个日语

表13. 多重随机森林的结果。不含长度构建的随机森林的平均袋外误差（-L OOBM）及其标准差（-L OOBSD）、含长度构建的随机森林的平均袋外误差（+L OOBM）及其标准差（+L OOBSD），以及 +L 个多随机森林中长度的平均特征重要性（L ImpM）。

MRF	-L OOB[M]	-L 袋外[SD]	+L 袋外[M]	+L 袋外[SD]	L Imp[M]
中文	40.85%	3.35%	39.36%	4.52%	0.51%
日语	34.07%	2.40%	31.69%	3.01%	5.47%
韩语	43.28%	3.09%	40.85%	2.85%	2.44%
诱发	36.29%	1.98%	32.47%	2.44%	6.99%

https://doi.org/10.1371/journal.pone.0279350.t013

表 14. 在官方宝可梦名称的多个包含长度这一特征的子集上训练和测试的日语 MRF 混淆矩阵。

		分类	
		进化前	进化后
样本	进化前	609	314
	进化后	285	682

https://doi.org/10.1371/journal.pone.0279350.t014

表 15. 基于日语官方宝可梦名称训练的日语随机森林的声音特征重要性（RF Imp）、基于官方名称训练的日语 MRF 的声音平均特征重要性（MRF ImpM），以及日语 MRF 的平均标准差（MRF ImpSD）。星号表示平均 p 值小于 0.05。

特征	RF Imp	MRF Imp ^{M}	MRF Imp ^{SD}
/m/	0.78%*	0.39%*	0.22%
/ɴ/	0.63%*	1.29%*	0.38%
/ː/	0.45%*	0.83%*	0.21%
/g/	0.40%*	0.45%	0.20%
/a/	0.39%	0.37%	0.23%
/ɾ/	0.35%	0.67%*	0.26%
/Q/	0.25%	0.01%	0.05%
/ɸ/	0.21%	0.09%*	0.11%
⁀/tɕ/	0.19%	0.10%	0.07%
/d͡ʒ/	0.19%	0.11%	0.14%
/d/	0.19%	0.50%	0.26%
/ɯ/		0.51%*	0.29%
/o/		0.27%	0.18%
/s/		0.10%	0.07%

https://doi.org/10.1371/journal.pone.0279350.t015

表 16。在所有官方日语 Pokémon 名称上训练并在所有引 elicited样本上测试的多随机森林混淆矩阵。

		分类	
		进化前	进化后
样本	进化前	3055	1510
	进化后	1283	2855

https://doi.org/10.1371/journal.pone.0279350.t016

表 17. 使用所有诱发样本训练、并在所有官方日语宝可梦名称上测试的多随机森林混淆矩阵。

		分类	
		进化前	进化后
样本	进化前	1436	1291
	进化后	585	2340

https://doi.org/10.1371/journal.pone.0279350.t017

数据集，并分别在诱发和日语多随机森林上进行测试。 表 16展示了使用官方日语 Pokémon 名称训练、并在诱发样本上测试的随机森林。 表 17展示了使用诱发样本训练、并在官方名称上测试的随机森林的混淆矩阵。

表18. 使用实验1的引出数据测试日语 MRF，以及使用日语数据测试引出式MRF的结果。其中包括不将长度作为特征的两个 MRF（-L OOBM）和将长度作为特征的两个 MRF（+L OOBM），以及它们的标准差（-L OOBSD，+L OOBSD）。

训练数据	测试数据	-L 袋外^{M}	-L 袋外^{SD}	+L 袋外^{M}	+L 袋外^{SD}
日语	诱发	37.31%	1.26%	32.09%	0.75%
诱发	日语	35.86%	2.01%	33.19%	2.03%

https://doi.org/10.1371/journal.pone.0279350.t018

为了进一步探究过拟合问题，我们重新构建了 -L 日语 MRF；不过这次我们跳过了调参过程，并采用 Ranger 软件包中的默认超参数设置。这样做是因为我们认为，过拟合最可能的原因是调参过程建议的超参数设置导致决策树的变异性较低。未经调参的 -L 日语 MRF（袋外误差M = 35.94%，标准差 = 1.75%）的准确度低于经过调参的 MRF（袋外误差M = 34.07%，标准差 = 2.4%），但标准差更低，这表明过拟合在单个随机森林中的情况较少。随后，我们使用全部官方日语名称重新创建了未经调参的 MRF，并用全部诱发名称对其进行测试，发现了相同的模式：未经调参的 MRF（袋外误差M = 38.24%，标准差 = 0.42%）准确度较低，但比表*18*中所示的经过调参的 MRF（袋外误差M.31%，标准差.26%）更稳定。

我们考虑了一种可能的替代解释，来说明调优后的多随机森林标准差较高：子集划分随机化所造成的变异，可能是由于测试子集和训练子集中进化前／进化后宝可梦的比例过高或过低所致。我们构建了一个简单的回归模型，用于预测测试子集中进化后宝可梦比例增加对从多随机森林中提取的所有日语、中文和韩语随机森林的袋外误差所产生的影响。由于引出的回答中，样本分布到进化前／进化后类别的情况有所不同，因此未纳入诱导数据。未观察到子集中的分布与袋外误差之间存在相关性，$F(1,25) = 0.31$，$p = 0.581$，$R.01$。因此，我们认为，随机化子集时准确率的变异很可能是由决策树的变异性较低所导致的过拟合造成的。

在分类实验中，119名日本参与者分别将二十个名称归入进化前或进化后类别。这二十个名称取自实验1结果中的100个随机抽取样本。参与者将诱发的宝可梦名称归入进化前和进化后类别时，准确率尚可（M = 61.58%，SD = 17.84%）。这项评估依据的是各参与者的个体回答，并以其均值准确率为准。考虑到声音象征关联，这种评估人类能力的方式可以说并不公平

由言语社群而非个体说话者决定。在为分类任务构建的随机森林中，每棵决策树都会对样本的分类结果投票。随机森林根据多数投票来确定分类结果。为了将这种方法应用于分类实验的结果，我们将每个回答视为一票，并分析了多数投票的结果。简单来说，我们针对每个样本考察的是众数，而非均值。通过多数投票，分类实验的参与者能够准确分类71%的样本。随后，使用以官方日语名称构建的多随机森林中的每个随机森林，对相同的100个样本进行测试。多随机森林的分类准确率远高于人类，正确分类样本的比例为75.88%（袋外误差M = 24.12%，标准差 = 1.61%）。

讨论

实验2旨在检验机器学习算法的表现是否与人类相当，不过，可能并不立即清楚，表18中列出的哪些多随机森林

应当作为衡量算法准确率的公平标准。我们考虑算法的结果 f
使用所有官方日语宝可梦名称的长度和声音训练的 MRF，并在实验2所用的100个样本上进行测试（袋外误差 $M = 24.12\%$），是衡量算法性能最公平的指标，因为该模型利用了人类参与者也能获得的最多信息进行训练，并在实验2所用的相同样本上进行测试。将人类参与者的回答转换为袋外误差后可以看出，人类参与者（袋外误差 $M = 38.42\%$，$SD = 17.84\%$）的准确率远低于算法（袋外误差 $M = 20.12\%$，$SD = 1.61\%$），即便采用多数投票法也是如此（袋外误差 $= 29\%$）。

算法在宝可梦分类任务中的准确率高于单个人类参与者，这一发现有悖直觉，尤其考虑到训练 MRF 所用的数据有限。对此的一种解释是，人类参与者并非始终全力以赴，而机器学习算法则会。这种投入不足可能源于缺乏理据、没有认真对待调查，或其他任何难以纳入考量的因素。不过，我们认为，以下原因表明，这并不能完全解释分类准确率的差异。首先，这项分类实验是自愿参加的；参与者没有因参加实验而获得金钱或其他形式的奖励。虽然纸质问卷在课前发放，但学生并没有在课堂上留出时间完成实验，完全需要自行安排时间完成。此外，任务很简短，只需大约2–3分钟。最后，之所以特意选择这一主题，是因为它对该样本群体具有吸引力，而且为他们所熟悉。基于这些因素，我们预计参与者会有较高的兴趣，且许多人会认真投入实验。

因此，我们认为，另一种解释或许更能说明参与者与算法准确率之间的差异：人类容易受到认知偏差的影响，而机器学习算法则不会。例如，人们常常会把过度简化的形象或观念套用到某类人或事物上，这被称为刻板印象。从随机森林的视角来看，

刻板印象就是将某种特征过度套用于某个类别。其他认知偏差表明，人类并不能凭直觉理解概率；鉴于声音象征具有随机性而非确定性，这一点尤为重要 [24]。这些偏差包括近因偏差（也称为可得性偏差），即人们预期近期发生过的事件会再次发生，而不考虑其发生概率；还包括合取谬误，即认为具体情况比一般情况更有可能，即便该具体情况包含了一般情况 [55]。事实上，其他研究表明，机器学习算法的表现可以优于人类（近期综述见 [56] ）。例如，McKinney等人 [57]提出了一种机器学习算法，在乳腺癌预测表现方面优于六名乳腺X线摄影图像专家阅片者。与放射科专家相比，该算法的假阳性和假阴性绝对数量均有所降低。考虑到任务的性质和人类参与者的情况，我们可以相当有把握地认为，表现差异并非源于放射科医生缺乏兴趣或理据。因此，我们必须考虑，本研究中人类参与者与算法之间的袋外误差差异，可能是由于学习效率以及学习成果应用方面存在差异。

实验1中报告的随机森林未纳入长度，因为我们希望单独考察语音的特征重要性，而描述性统计表明，长度可能是一个重要特征，会掩盖较弱特征的重要性。事实上，长度在所有多随机森林中都被发现很重要。在诱发式（6.99%）和日语（5.47%）多随机森林中，长度最为重要，这表明词长在日语中承载着相当多的声音象征信息。它在韩语

（2.44%）和中文（0.51%）多随机森林中的重要性则较低。在实验1中将长度与其他特征分离，并在实验2中将其纳入随机森林，揭示了导致采用多随机森林的过拟合问题。实验1中的不含长度的日语随机森林（袋外误差率为29.05%）表现优于实验2中的含长度日语随机森林（袋外误差率为30.95%），尽管长度在含长度日语随机森林中是一个非常重要的特征（4.54%）。过拟合最可能的解释是决策树之间的差异很小。我们通过使用 Ranger 软件包的默认超参数设置重新构建日语随机森林，对这一假设进行了检验。运行未经调优的多随机森林后，得到的随机森林更加稳定，准确率仅略低于调优后的对应模型。这一发现支持我们的假设：实验1中的过拟合是决策树之间差异不足所致。决策树差异不足可能是因为每个节点上考察的特征数量（mtry）过多；由于数据集中空值比例很高（82.26%），调优过程建议采用这一设置。

我们探讨了一个解决这一问题的潜在方案，即使用默认超参数设置构建每个随机森林；但在所有情况下，这都会导致袋外误差增大。另一个潜在方案是，不报告声音本身，而是报告音系特征，以减少空值的数量。这样既能减少空值的数量，也能减少特征的数量，从而降低数据分辨率的精细程度。我们采用了另一种方法：构建由相互独立的随机森林组成的 MRF，并为数据划分为子集的过程以及随机森林本身的随机化设置不同的起始值。乍一看，MRF 似乎是堆叠随机森林（SRF：[58]），但事实并非如此。堆叠法 [37]是一种通过将较弱模型组合成超级学习器来提高算法准确率的方法 [59]。例如，Hänsch [58]通过使用先前随机森林的估计值，依次向 SRF 中添加随机森林，以提高最终模型的准确率。我们的方法更类似于 k 折交叉验证：将数据随机划分为 k 个组或折，然后通过部分拉丁方重新组合数据，生成多个训练/测试子集，用于构建和测试该算法的多次迭代 [60]。本文未采用 k 折交叉验证

本研究中，如果用户遵循三分之二子集规则，那么他们在选择迭代次数时会受到限制。

结论

本研究使用宝可梦名称构建并测试机器学习算法。这些算法旨在将宝可梦分类为进化前和进化后类别。在实验1中，算法使用构成日语、中文和韩语宝可梦名称的语音来构建。对这些算法进行的特征重要性计算显示，虽然某些声音象征模式在不同语言中都存在，但许多重要特征因语言而异。实验1还包括一项引出实验，让日本参与者为此前未见过的宝可梦命名。随后，我们使用完整的官方日语宝可梦名称数据和引出的回答构建随机森林，并在各自相反的数据集上对其进行测试。这些测试的袋外误差表明，这些数据集中的声音象征模式相当相似，这说明这些声音象征模式要么已存在于日语中，要么参与者熟悉宝可梦的命名惯例。先前研究显示，在非真实词宝可梦命名实验中，对宝可梦的熟悉程度与声音象征效应大小无关 [61,62]，这表明其结果并非源于对宝可梦名称的既有认知。在实验2中，所有算法都经过重新构建，将名称长度纳入特征。这揭示了一个过拟合问题，我们通过多随机森林予以解决。随后，我们将多随机森林的表现与日本参与者的表现进行比较。结果表明，多随机森林的准确率高于人类。

据称，随机森林适用于“小样本量、高维度”数据集 [63]，，例如本研究中的数据集。然而，实验2揭示了实验1中一个明显的过拟合案例。将长度作为特征的随机森林显示，长度很重要，但这种重 +要性并不总能反映在袋外误差中。举例来说，日语 = 31L 多随机森林（袋外误差 = 29.69%）的表现不如实验1中的 -L 随机森林（袋外误差.05%）。既然长度在多随机森林中被发现很重要，这就说明单个随机森林发生了过拟合，因为决策树缺乏变异性。进一步的证据可见于以下准确率差异：以官方日语宝可梦 = 38名称训练、用于分类诱发名称的随机森林（袋外误差.57%），以及以官方日语宝可梦 ^{M} = 37名称训练、用于分类诱发名称的 -L 多随机森林（袋外误差.31%）。换言之，实验1中的日语随机森林在分类自身测试子集时比日语多随机森林更准确，但在分类引出的样本时准确率较低，因为其函数与初始数据集过于紧密地吻合，导致其分类外部样本的能力下降。

声音象征是对声音与意义之间系统性关系的研究。这些关系并非确定性的，而是随机的，因此需要通过统计分析加以观察。本文详细介绍了随机森林算法，这些算法能够从此类随机关系中学习，并将所学知识应用于分类任务。该任务是将宝可梦分类为进化前后类别。这一发现对自然语言处理研究领域具有重要意义，拓展了 Winter 和 Perlman 的研究成果， [11]并表明，机器学习算法能够依据声音象征原则（至少在很大程度上）作出分类决策；如果算法的目标是像人类一样理解和使用语言，就应该这样做。这些算法通过特征重要性展示其分类决策的依据；特征重要性是一项有用的指标，可用于衡量特定语言特征的声音象征特质。在评估普遍的声音象征模式时，这一点尤其有用。本文还揭示了随机森林中固有的过拟合问题：当随机森林由变异性较低的决策树构建时便会出现这一问题。本文

这一问题通过在多个随机森林之间随机分配训练集和测试子集划分得到解决。机器学习算法被证明是这项任务中高效的学习器，其分类准确率高于人类参与者，尽管它们只能从数量有限的样本中进行学习。

作者贡献

概念化： Alexander James Kilpatrick、Shigeto Kawahara。

数据整理： Alexander James Kilpatrick。

形式分析： Alexander James Kilpatrick。

经费获取：Alexander James Kilpatrick .

调查：Alexander James Kilpatrick。

方法学：Alexander James Kilpatrick。

项目管理：Alexander James Kilpatrick .

资源：Alexander James Kilpatrick。

软件：Alexander James Kilpatrick。

监督： Alexander James Kilpatrick。**验证：** Alexander JamesKilpatrick。**可视化：** Alexander James Kilpatrick。

撰写——初稿： Alexander James Kilpatrick, Aleksandra Ćwiek。

撰写——审阅与编辑： Alexander James Kilpatrick, Aleksandra Ćwiek, Shigeto Kawahara.

参考文献

1. De Saussure F.《普通语言学教程》。巴黎：Payot；1916年。2. Berlin B. 民族动物学命名法第一次大会。《皇家人类学研究所杂志》。2006年；12:S23–44。3. Blasi DE, Wichmann S, Hammarström H, Stadler PF, Christiansen MH. 数千种语言中均有证据表明存在声音—意义联想偏向。《美国国家科学院院刊》。2016年；113(39):10818–23。https://doi.org/10.1073/pnas.1605782113 PMID: 27621455 4. Newman SS. 语音象征的进一步实验。《美国心理学杂志》。1933年；45(1):53–75。5. Shinohara K, Kawahara S. 声音象征的跨语言研究：大小意象。载于《伯克利语言学会年会论文集》2010；36(1):396–410。6. Imai M, Kita S. 语言习得与语言演化的声音象征引导假说。《英国皇家学会哲学汇刊B辑：生物科学》。2014年9月19日；369 (1651):20130298。https://doi.org/10.1098/rstb.2013.0298 PMID: 25092666 7. Ozturk O, Krehm M, Vouloumanos A. 婴儿期的声音象征：4个月大婴儿中声音—形状跨模态对应关系的证据。《实验儿童心理学杂志》。2013年；114(2):173–86。https://doi.org/10.1016/j.jecp.2012.05.004 PMID: 22960203 8. Perniss P, Vigliocco G. 象似性之桥：从体验世界到体验语言。《英国皇家学会哲学汇刊B辑：生物科学》。2014年；369 (1651):20130300。9. Sidhu DM, Williamson J, Slavova V, Pexman PM. 对四个数据集中的象似语言发展的考察。《儿童语言杂志》。2022年；49(2):382–96。https://doi.org/10.1017/S0305000921000040 PMID: 34176538 10. Breiman L. 随机森林。《机器学习》。2001年；45(1):5–32。11. Winter B, Perlman M. 英语词汇中的大小声音象征。《Glossa：语言学杂志》。2021年；6(1)。12. Miura, T, Murata, M, Yasuda, S, Miyabe, M, Aramaki, E. (2012). 声音象征的机器学习再现：生成最强的宝可梦 [声音象征的机器学习再现：生成最强的宝可梦]。语言处理学会第18届年会论文集 [第18届年会论文集的

自然语言处理协会], 2012；18:65–68。13. Bainbridge J. "'这是一个宝可梦的世界'：宝可梦系列与环境"。《国际文化研究杂志》。2014年；17(4):399–414。14. Hockett CD. "言语的起源"。《科学美国人》。1960；203(3):88–97。PMID: 14402211 15. Kawahara S. "声音象征与理论音系学"。《语言与语言学指南》。2020年；14(8):e12372。16. Sidhu DM, Pexman PM. "声音象征联结的五种机制"。《心理学通报与评论》。2018年；25(5):1619–43。https://doi.org/10.3758/s13423-017-1361-1 PMID: 28840520 17. Spence C. "跨模态对应关系：教程式综述"。《注意、知觉与心理物理学》。2011年；73(4):971–95。https://doi.org/10.3758/s13414-010-0073-7 PMID: 21264748 18. Köhler W. "格式塔心理学"。纽约：NY；1947。19. Bremner AJ, Caparos S, Davidoff J, de Fockert J, Linnell KJ, Spence C. "纳米比亚也有'Bouba'和'Kiki'吗？一个遥远文化群体与西方人做出相似的形状—声音匹配，但不同的形状—味觉匹配"。《认知》。2013年；126(2):165–72。https://doi.org/10.1016/j.cognition.2012.09.007 PMID: 23121711 20. Chen YC, Huang PC, Woods A, Spence C. "当'Bouba'等于'Kiki'：声音—形状对应关系中的文化共性与差异"。《科学报告》。2016年；6(1):1–9。21. Ćwiek A, Fuchs S, Draxler C, Asu EL, Dediu D, Hiovain K, 等。"bouba/kiki效应在不同文化和书写系统中都很稳健"。《英国皇家学会哲学汇刊B辑》。2022年；377 (1841):20200390。https://doi.org/10.1098/rstb.2020.0390 PMID: 34775818 22. Rogers SK, Ross AS. "跨文化检验Maluma-Takete现象"。《知觉》。1975年；4 (1):105–106。https://doi.org/10.1068/p040105 PMID: 1161435

23. Styles SJ, Gawne L. maluma/takete何时会失效？两种关键失效情况和一项荟萃分析表明 t
音系学和音位组合学很重要。i-Perception。2017年；8(4):2041669517724807。
https://doi.org/10.1177/2041669517724807PMID：28890777**24.** Kawahara S、Katsuda H、Kumagai G。使用最大熵模型解释声音象征的随机性。Open Linguistics。2019年；5(1):109–20。**25.** Maurer D、Pathman T、Mondloch CJ。boubas的形状：幼儿中的声音—形状对应关系
s
和成年人。发展科学。2006年；9(3):316–22。https://doi.org/10.1111/j.1467-7687.2006. 00495.xPMID：16669803**26.** Perry LK、Perlman M、Lupyan G。英语和西班牙语中的象似性及其与词类的关系
d
习得年龄。《PloS one》。2015年；10(9):e0137147。https://doi.org/10.1371/journal.pone.0137147PMID：26340349**27.** Perry LK、Perlman M、Winter B、Massaro DW、LupyanG。儿童和成年人的言语中的象似性。发展科学。2018年；21(3):e12572。https://doi.org/10.1111/desc.12572 PMID：28523758**28.** Kawahara S、Kumagai G。在宝可梦名中表达进化：实验性探索。《日语语言学》期刊。
l
语语言学。2019年；35(1):3–8。
29. Kawahara S、Moriya M. 2021年，59(3):577–607英语宝可梦名中表达进化 · 语言学 ·

30. Kawahara S, Noto A, Kumagai G. 《宝可梦》名称中的声音象征模式。Phonetica. 2018年; 75 (3):219–44.
31. Shih SS, Ackerman J, Hermalin N, Inkelas S, Kavitskaya D. 宝可梦名称：一项关于声音象征的研究 m
与《宝可梦》名称。美国语言学会论文集。2018年; 3(1):42:1–6.
32. Bao Begin M, Brăguţ 随机森林回归 Test 2016年2月)，197–187。《PRS摄影测量与遥感杂志》。2016年；114：24–31。**34.** Ziegler A, König IR. 使用随机森林进行数据挖掘：面向实际应用的当前方案。《Wiley跨学科评论：数据挖掘与知识发现》。2014年；4(1)：55–63。**35.** Al-Akhras M, El Hindi K, Habib M, Shawar BA. 用于避免决策树过拟合的实例约简。《智能系统杂志》。2021年；30(1)：438–59。**36.** Kotsiantis S. B.（2013年）。决策树：近期综述。《人工智能评论》，39(4)，261–283。**37.** Breiman L. 装袋预测器。机器

学习。1996年；24(2):123–40. 38. Ho TK. 用于构建决策森林的随机子空间法。《IEEE模式分析与机器智能汇刊》。1998；20(8):832–44. 39. Bulbagarden. Bulbapedia：由社群共同编写的宝可梦百科全书[互联网] 2004 [更新于2022年4月 22]；网址：https://bulbapedia.bulbagarden.net/wiki/Main_Page。40. Sohn HM. 韩语。Routledge；2019年。41. Duanmu S. 标准汉语音系。牛津大学出版社；2007年。42. Wright MN, Ziegler A. ranger：在 C++ 和 R 中快速实现高维数据随机森林。《统计软件杂志》。2017年；77(1):1–17. https://doi.org/10.18637/jss 43. Altmann A, Toloşi L, Sander O, Lengauer T. 置换重要性：一种经过校正的特征重要性度量。《生物信息学》。2010年；26(10):1340–7. https://doi.org/10.1093/bioinformatics/btq134 PMID: 20385727 44. Probst P, Wright MN, Boulesteix AL. 随机森林的超参数与调优策略。《Wiley跨学科评论：数据挖掘与知识发现》。2019年；9(3):e1301. 45. Sapir E. 语音象征研究。《实验心理学杂志》。1929年；12(3):225. 46. Ultan R. 大小—声音象征。《人类语言共性》。1978年；2:525–68. 47. Cho YM, Gregory KI. 二十世纪后期的韩语音系学。《语言研究》。1997年；33 (4):687–735. 48. Kim KO. 韩语中的声音象征。《语言学杂志》。1977年；13(1):67–75. 49. Haiman J. 语法的象似性：同构与理据。《语言》。1980年；56(3):515–40. 50. Shih SS, Ackerman J, Hermalin N, Inkelas S, Jang H, Johnson J, Kavitskaya D, Kawahara S, Oh M, Starr RL, Yu A. 跨语言与特定语言的声音象征：宝可梦拟声词。手稿。南加州大学、加利福尼亚大学默塞德分校、加利福尼亚大学伯克利分校、庆应义塾大学、新加坡国立大学和芝加哥大学。2019年。51. Ohala JJ. 频率编码假说是音高象征性运用的基础。声音象征

-
象征。1994；2：325–47。

52. Lapolla RJ. 关于语音象征与普通话关系的实验研究。载于：Hinton L、Johanna N、Ohala J，编。《声音象征》。2006年；剑桥：剑桥大学出版社：130–147。53. Kumagai G. 双唇辅音的多潜能性：日语和英语中柔和度与可爱度的意象。Open Linguistics。2020；6(1)：693–707。54. Otake T、Hatano G、Cutler A、Mehler J. 音拍还是音节？日语语音切分。《记忆与语言杂志》。1993；32(2)：258–78。55. Tversky A、Kahneman D. 外延推理与直觉推理：概率判断中的合取谬误。《心理学评论》。1983；90(4)：293–315。56. Webb ME、Fluck A、Magenheim J、Malyn-Smith J、Waters J、DeschênesM、Zagami J. 面向人类学习者的机器学习：机遇、问题、张力与威胁。《教育技术研究与发展》。2021年；69(4)：2109–30。57. McKinney SM、Sieniek M、Godbole V、Godwin J、Antropova N、Ashrafian H、Back T、Chesus M、Corrado GS、Darzi A、Etemadi M. 用于乳腺癌筛查的人工智能系统的国际评估。《自然》。2020；577(7788)：89–94。https://doi.org/10.1038/s41586-019-1799-6 PMID: 31894144 58. Hänsch R. 堆叠随机森林：更准确且校准效果更佳。载于 IGARSS 2020–2020 IEEE国际地球科学与遥感研讨会。2020；1751–1754，https://doi.org/10. 1109/IGARSS39084.2020.9324475 59. Van der Laan MJ、Polley EC、Hubbard AE. 超级学习器。《遗传学与分子生物学统计应用》。2007；6(1)：1544–6115，https://doi.org/10.2202/1544-6115.1309 PMID: 17910531 60. Gareth J、Daniela W、Trevor H、Robert T.《统计学习导论：R语言应用》。Spinger；2013年。61. Kawahara S、Kumagai G. 运用声音象征推断宝可梦的属性：浊音化与唇音性的影响。《音声研究》[speech research]。2019年；23：111–6。62. Godoy MC、de Souza Filho NS、de Souza JG、Franc,a HA、Kawahara S. 全部都要命名：巴西葡萄牙语宝可梦名称声音象征的实验研究。《心理语言学研究杂志》。2020年10月；49(5)：717–40。63. Strobl C、Malley J、Tutz G. 递归划分导论：分类与回归树、装袋法和随机森林的原理、应用及特征。《心理学方法》。2009；14(4)：323–

348. https://doi.org/10.1037/a0016973PMID：19968396