从数据的角度回顾腐蚀预测的机器学习

最近机器学习的热度比较火，我前段时间也开始了这方面的学习，幸好，由于本人之前在python的数据分析方面上有一定的功底，sklearn库的上手挺快。最近看到了一篇文章描述比较好几个腐蚀方面的模型性能，其中的分析很精确且全面。接下来我将介绍一下这篇文章并偶尔加上我的一些见解。

### 文章介绍

《Reviewing machine learning of corrosion prediction in a data-oriented perspective》选择最合适的ML方法是一个多方面的问题，这取决于可用的数据量、期望结果的质量以及对结果进行物理解释的必要性。因此，研究重点评估了不同机器学习方法的预测能力，并详细阐述了各种腐蚀主题的回归建模的现状。

### 文献计量学数据挖掘

这篇文章的作者先是采用Web of Science搜索引擎，然后关键字“机器学习”和“腐蚀”进行搜索，而后从返回的164条记录中，发现这两个关键字与“预测”关键字高度相关。最初，研究人员考虑了35篇明确将监督学习用于腐蚀预测的论文。最终选择了16篇采用回归模型并提供可比指标的参考文献，最后，其他3部人工神经网络也被包括在内。

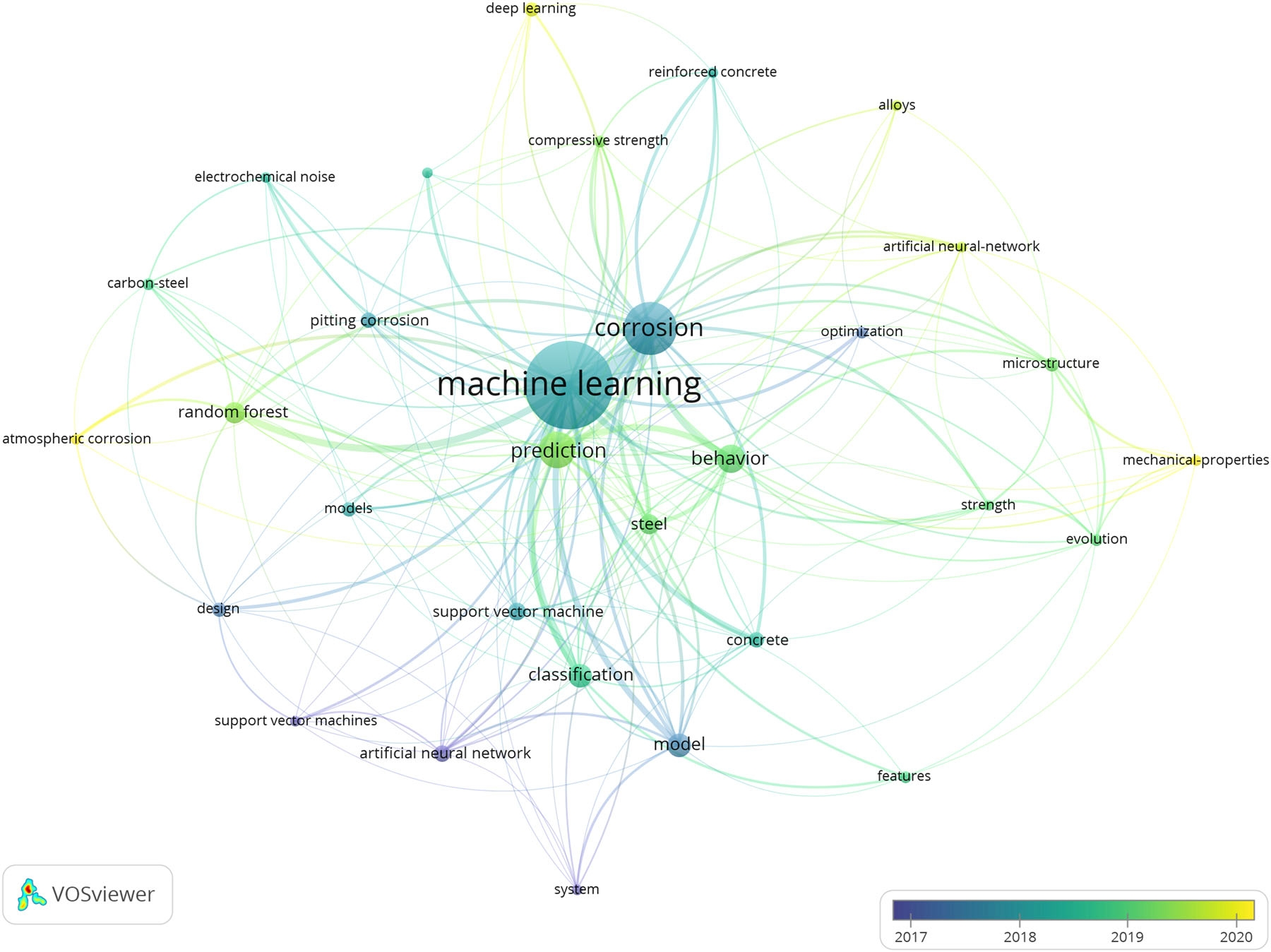


图1 “机器学习”和“腐蚀”出版物的共引网络的可视化

### 机器学习的腐蚀数据库

机器学习非常依赖数据库，没有足够的数据很难进行较为高效的训练，那么模型的拟合优度会非常低。

这篇文章的database处理了47个特性和1个索引列，特征列表描述了调查方法(“定向策略”，“目标”)。它包括数据的基本描述符(“数据收集”，“数据可用性”)，腐蚀(“腐蚀主题”，“材料”)和ML(“模型描述”，“评估方法”，“训练-测试分割”)。

### 时间变量的影响

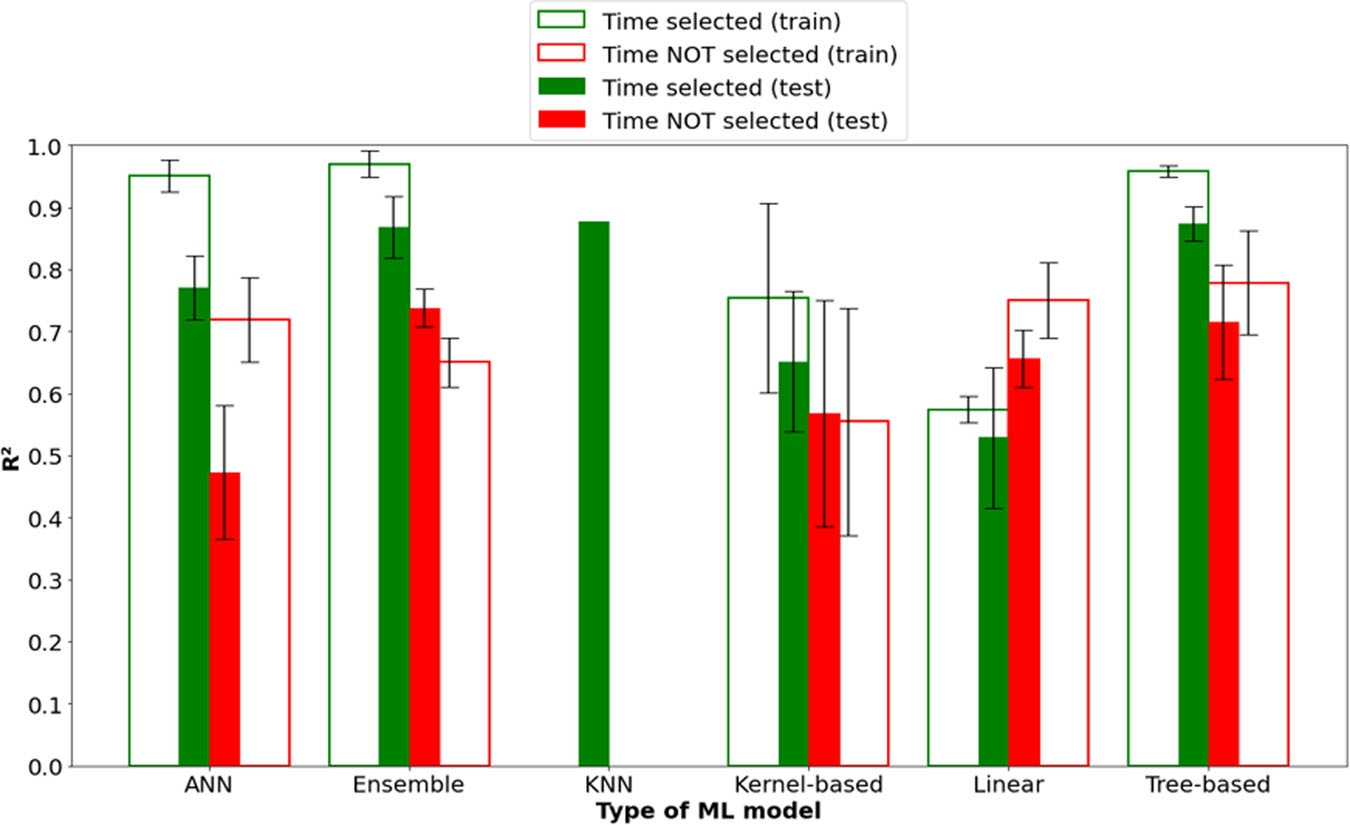
这篇文章让我第一个很欣赏的点，那就是它意识到了线性模型不适用于预测腐蚀，因为腐蚀本质上是以时间序列数据为特征的。时间序列比起其他的特征序列有着一个独有的特征：它有着周期性的变化，可能包括每日，每周，每月，每季节或每年周期。而对于腐蚀情况，时间不同对应着天气因素，使得腐蚀预测结果应当有着一个极其突出的周期性偏向。图2显示，当将时间作为输入时，模型的性能(用R (2)表示)显著提高。

图2 均值R²作为“ML模型类型”的函数，取决于是否选择时间作为特征

可以看出，六个模型在使用时间序列作为特征构建出的模型的拟合优度比之前都有了很大的提升。

唯一可惜的一点是，这篇文章并没有将时间序列处理的细节写在文章中，比如假设检验过程中用的是哪种检验统计量？或是测试相关性方面用的是平均移动值方法还是自相关方法？本人目前学的也不精，现阶段不好献丑。

## 回顾每个腐蚀主题的机器学习

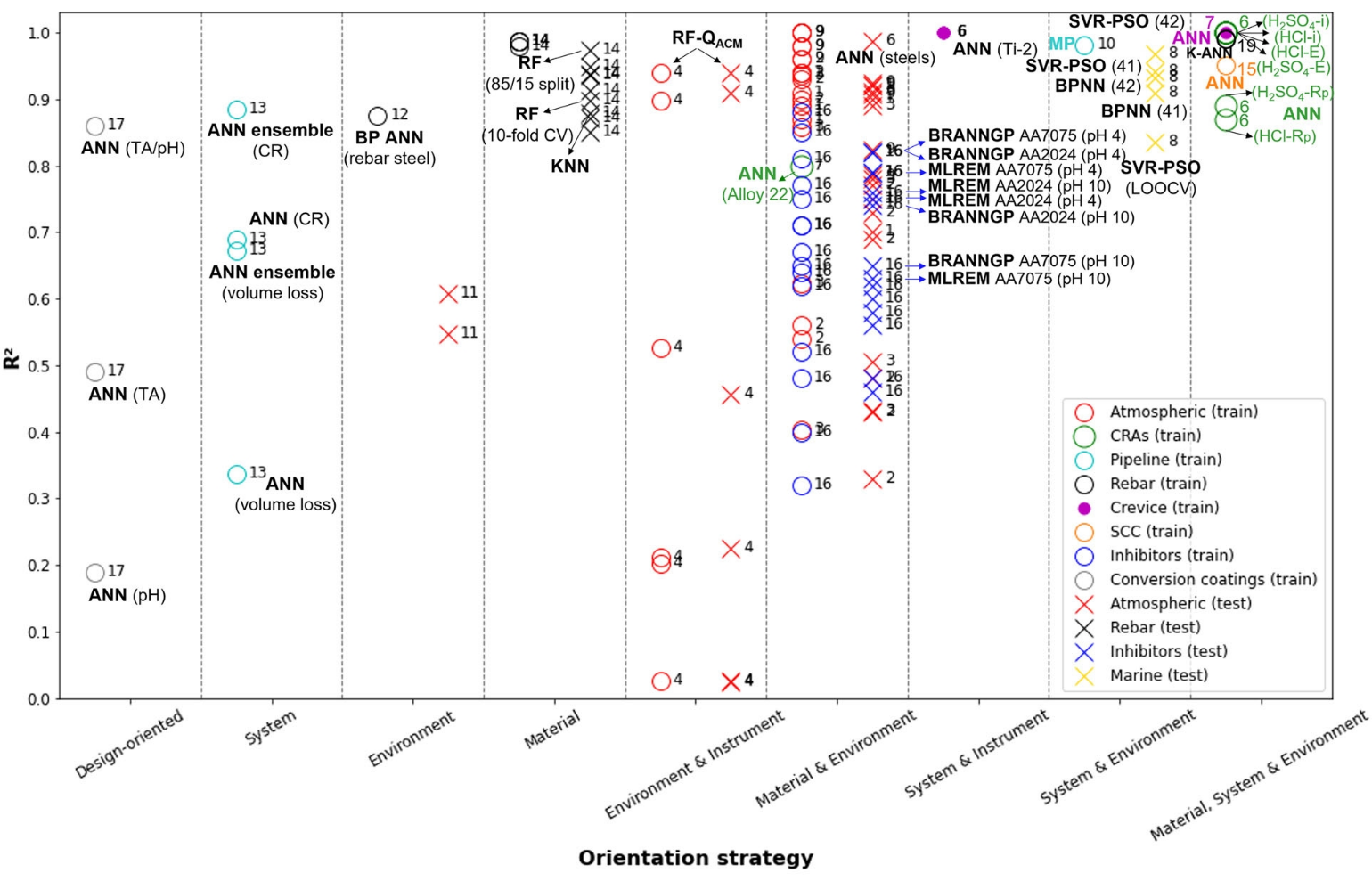
接下来介绍的模型的拟合优度量化值（即决定系数R2）都会在下方三张图中显示。

图3 用R²表示的ML模型的性能，由“腐蚀主题”分割，并由“定向策略”组织。

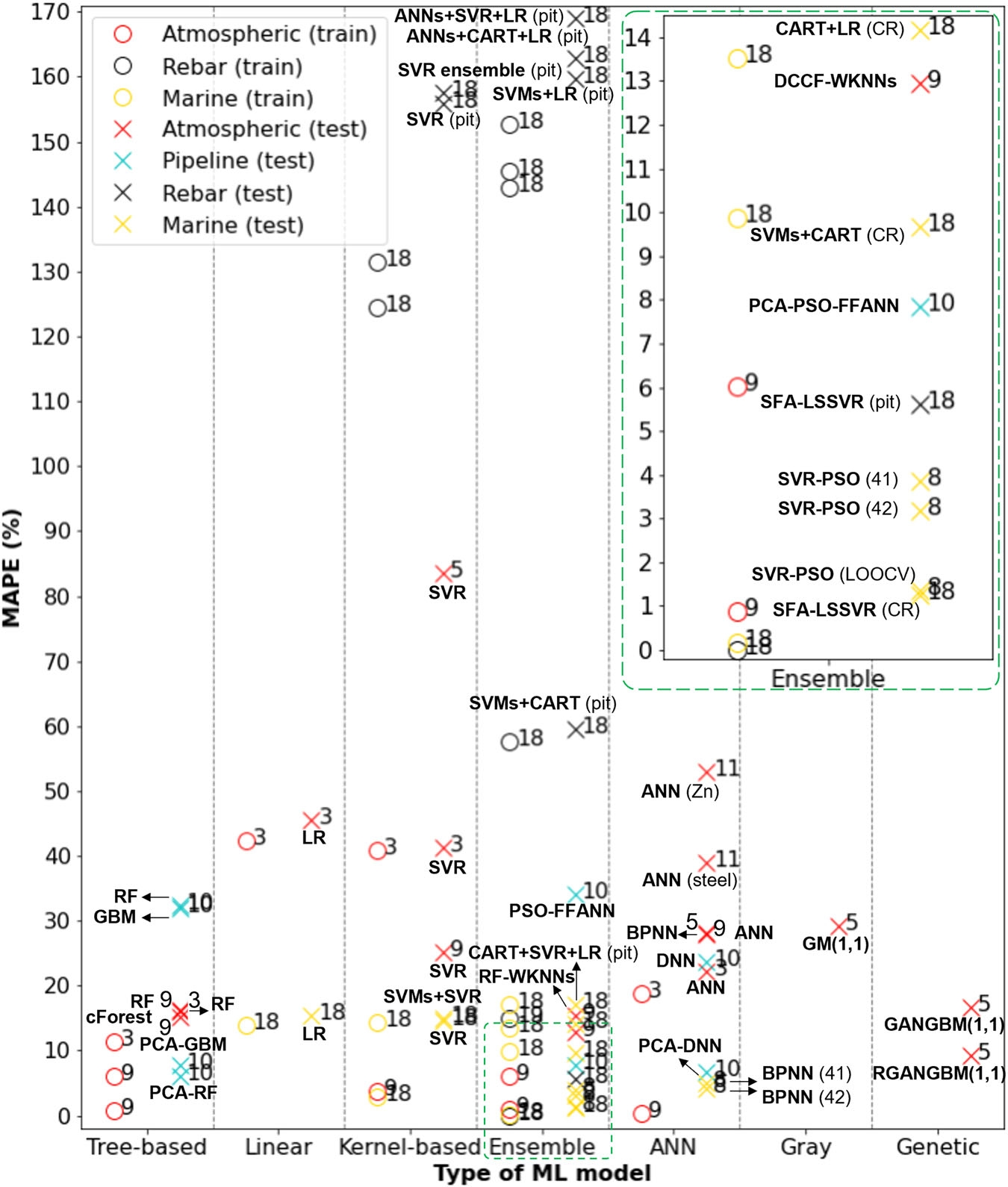
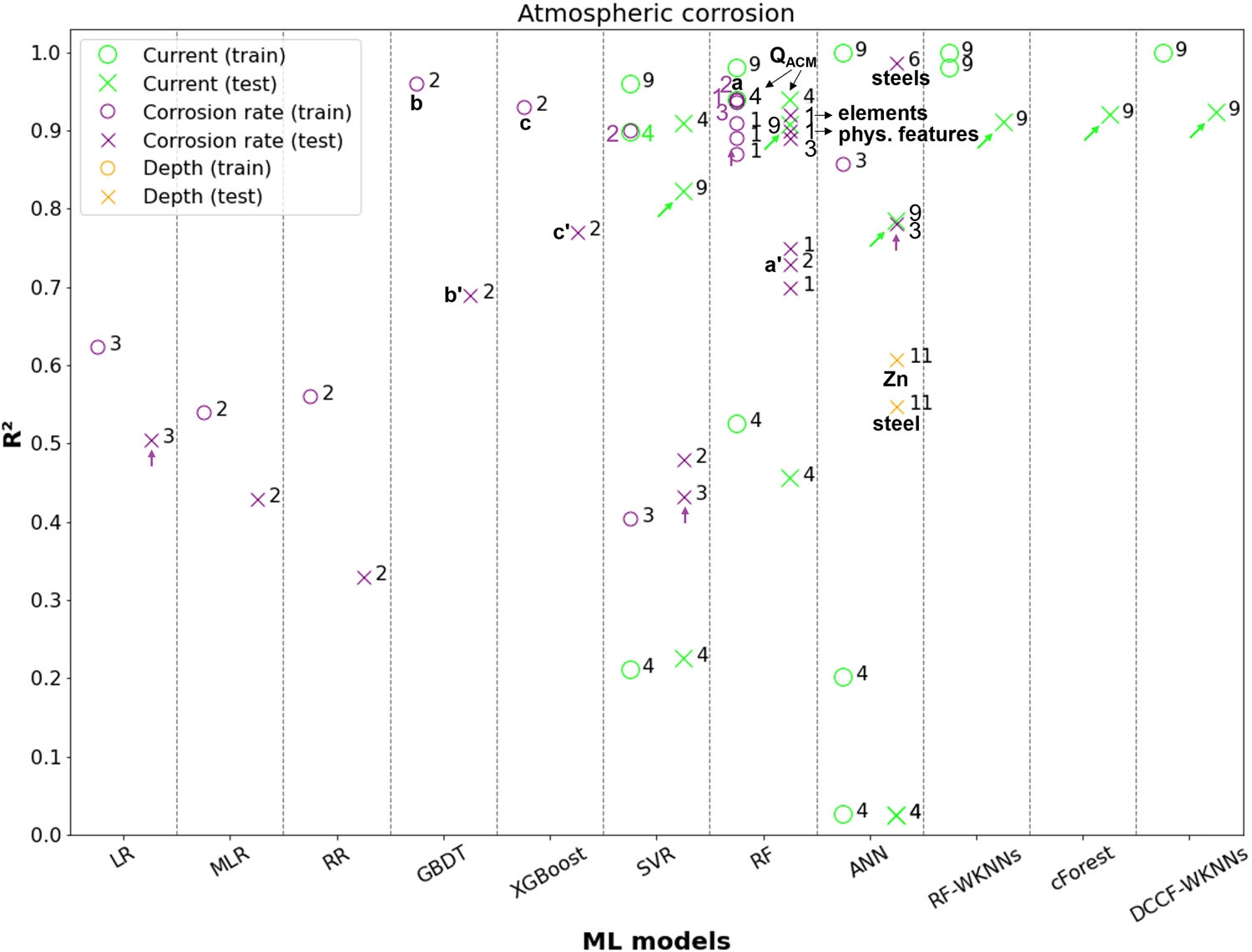


图4 MAPE表示的ML模型的性能，由“腐蚀主题”和“ML模型类型”分割

图5 用R²表示的ML模型应用于大气腐蚀的性能

### 大气腐蚀

大气腐蚀的数据集- 传统上是通过接触试验获得的，通常是小样本量和由多个腐蚀影响因素定义的高维数。而传统的预测模型对于这种高维的数据集是很难做出准确的预测的，目前最受欢迎的深度学习模型，对这种高维数据有着非常强大的预测能力。这篇文章中的作者也是重点讲述了ANN模型对于大气腐蚀数据集的处理效果。

通过灵敏度分析，ANN还可以证明锌的CRs最可能受到腐蚀产物变化的影响。在较高的湿润时间(TOW)下，腐蚀速率减慢，这表明水分抑制了腐蚀，这是由于保护性碳酸盐的形成。同时，在高浓度SO2下，由于少量可溶碱性硫酸盐向可溶硫酸盐转变，腐蚀显著增加。TOW、SO2浓度与大气腐蚀之间几乎呈线性关系。而后对子数据集进行知识挖掘扩展，作者在对应文献中的分析曲线中发现环境因素比材料因素更重要，在两个腐蚀阶段(时间分割在第5年)，pH值是最重要的特征。在初始阶段，“环境”的重要性始终大于80%。其中，SO2和Cl -浓度的重要性指数高于温度(T)和相对湿度(RH)，而降雨仅在第1年为第2重要特征。在固定阶段(5年以上)，腐蚀产物将LAS与环境屏蔽，降低了环境因素的重要性(SO2的重要性远低于Cl−、T和RH)。

深度学习由于其原理性在于通过多个特征值之间插入隐藏层来而后经过笛卡尔积式的交叉计算特征值与隐层之间权重，所以特征数量和隐层数量越多，模型的复杂度更大，且对于样本训练集的数量要求并不高，便能创造出性能更好的预测模型。所以，在大气腐蚀上运用深度学习的方法是非常明智的。但是很可惜本文作者比较偏重于决定系数的比较，也就是预测模型性能的比较，并没有简要说明盖如何选择合适的隐层数量才能达到图中高达0.98的决定系数。

### 海洋腐蚀

深度学习虽然可以构建优异的评估器，但也有一个比较突出的缺点，若是样本数量过小的话，容易过拟合，使得模型的泛化能力很弱，所以在样本数量够多时，应该选择其他的模型获得泛化能力强的评估器。例如这篇文章讲述的海洋腐蚀所用的BPNN、SVR-PSO和分层集成模型就是一个很好的例子。

训练样本数量越高(41或42 / 46)的BPNN (MAPE测试)回归性能更好，分别为5.001%或4.215%)。与BPNN相比，aSVR-PSO具有卓越的泛化能力，结果MAPE(测试)值分别为3.840%、3.180%和1.360%(分别为41、42和45 (LOOCV)训练样本)。尽管如此，从拟合的角度来看，值得注意的是SVR-PSO (LOOCV)产生了最低的R²(测试)值:0.836。与相应模型8相关的所有报告指标(MAPE和R²(测试))在图3和图4中突出显示。

最近，Chou等人使用与参考文献8中相同的数据集，还尝试了碳钢海洋腐蚀的ML建模。预测包括由四个知名机器学习器构建的单一和集成模型，包括SVR/SVMs、分类和回归树(CART)和线性回归(LR)。各种ML模型通常用于构造单个分类器，但是集成系统会补偿由单个分类器所产生的错误。分层集成模型被证明进一步提高SVR(最佳个体模型)的预测精度。

受自然启发的元启发式优化算法，如萤火虫算法(FA)已经有效地解决了困难的优化问题。通过将智能萤火虫算法(SFA)与最小二乘算法SVR49集成，实现了一种混合模型。

更有意义的是，该SFA-LSSVR混合模型被证明对3C钢的CRs建模是有效的，优于所有其他模型。SFA-LSSVR(1.26%)和所有其他测试模型(SVR, CART + LR, LR, CART +SVR + LR, SVMs+ SVR, SVMs+CART)的MAPE(测试)值在图4中突出显示

这说明集成分布模型的“集合每个模型一起预测，而后软投票决定预测结果”原理在海洋腐蚀中很有效果。但是据我了解，由于需要多个模型一起预测，集成分布模型对计算机算力的要求非常高，这也是为什么集成分布模型为何最近才开始饱受关注，和计算机的发展有很大关系。

### 管道腐蚀

了解和预测管道腐蚀风险对于维护石油生产的健康和安全至关重要。油气公司腐蚀风险评估的第一步是在常规测量的许多参数中选择用于估计腐蚀缺陷深度的输入参数。

管道腐蚀方面的模型构造过程有我非常欣赏的一个点，它使用了主成分分析（接下来简称PCA）对数据集进行了一个预处理！这很符合我在学习机器学习中产生的一个观点：“无监督学习不适合用来构建预测模型，而更适合用来构建数据预处理器。”

在应用上述模型之前，他们使用了多元多项式回归来拟合来自60口井的整个数据集(R²(训练)= 0.9819，如图3所示为“MP”)。目的是通过拥有包含不同腐蚀严重程度场景的新数据集来避免数据偏差。

PCA变换强调了在预测腐蚀缺陷深度增长的未来状态方面的改进。对于PCA非转换输入和转换输入，PSO-FFANN、GBM、RF和DNN模型产生的训练MAPE(%)值分别为34.1329、31.9266、32.4267、23.647和7.8588、6.0082、7.7421、6.681310。均方误差大大降低了！所有这些模型/度量关系如图4所示。这正说明了PCA预处理数据的重要性，它能去除数据中的噪声影响，强调第一主成分中的特征影响能力。

通过对于模型的优化和集成，不仅能够腐蚀缺陷深度进行预测，还能确定特定管道段的腐蚀风险。

De Masi等人开发了一种预测模型，用于识别暴露于较高腐蚀风险的管道段(现有方法无法重现测量活动中观察到的行为)。作者的前提是，结合不同性质的数据(管道的几何轮廓、流体动力学多相变量和确定性模型)可以极大地提高预测任务的性能13。

因此，通过考虑多个网络输入，人工神经网络集成模型可以正确识别管道中最可能发生腐蚀的区域。该策略不仅提高了确定性模型的预测性能，而且提高了单神经网络模型的预测性能。例如，为ANN和ANN集合实现的R²(测试)值分别为:0.689和0.884 (CR)或0.336和0.672(体积损失)。这些性能指标都映射在图3中。

### 总结

由于个人精力有限，我只能简单的介绍比较喜欢的三个模型和发表了自己的一些看法。

总而言之，机器学习还是很依赖数据集的，无论何种模型，只要数据集足够大，模型的拟合优度都会相应地提高。但实际中样本数量往往难尽人意，这时就需要进行大量地模型比较以选出最好的评估器、采用较好的数据预处理优化数据集的作用、选出最佳的模型参数等等方法来不断选出最好的模型。很幸运阅览了《Reviewing machine learning of corrosion prediction in a data-oriented perspective》这篇文章，它让我再一次意识到机器学习的上限非常非常高，这决定了机器学习以后在制造业的前景非常广阔。