附件1 项目编号：

**中山大学**

**大学生创新训练计划项目申请书**

**项目名称：****基于机器学习的珠江口东侧**

**水域低氧长期变化规律和驱**

**动因素研究**

**相关学科：\_\_\_\_\_\_\_\_\_环境科学\_\_\_\_\_\_\_\_**

**申 请 人：\_杨涵 周鑫 陈至雪 潘文轩\_**

**所在院系：\_\_\_\_\_\_\_环境科学与工程\_\_\_\_**

**指导教师：\_\_\_\_\_\_\_\_\_\_胡嘉镗\_\_\_\_\_\_\_\_\_**

预期成果类别：

* 社科类社会调查报告及学术论文
* 自然科学类学术论文

教 务 部 制

填表时间 二O二 一 年 3 月 13 日

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 项目名称 | 基于机器学习的珠江口东侧水域低氧长期变化规律与驱动因素研究 | | | | | | | | | | | | | | | | | | |
| 申请人或申请团队 | 姓 名  （负责人请加\*号） | | 学号 | | | 性别 | 身份证号 | | | 年级 | | 专业 | 联系电话 | | | 手机 | | E-mail | |
| \*杨涵 | | 19324073 | | | 男 | 50022620000617153X | | | 2019 | | 环境科学 | 17783031232 | | | 17783031232 | | 931295352@qq.com | |
| 陈至雪 | | 19335030 | | | 女 | H60601223（回乡证） | | | 2019 | | 计算机科学与技术（超级计算） | 13536328014 | | | 13536328014 | | 13536328014@163.com | |
| 周鑫 | | 17338242 | | | 男 | 522527199908140014 | | | 2017 | | 生物技术 | 18996256143 | | | 18996256143 | | 412055371@qq.com | |
| 潘文轩 | | 19335163 | | | 男 | 441202200106161519 | | | 2019 | | 计算机科学与技术 | 13822629003 | | | 13822629003 | | panwx5@mail2.sysu.edu.cn | |
|  | |  | | |  |  | | |  | |  |  | | |  | |  | |
| 指导教师 | 姓名 | | 性别 | 年龄 | 职称 | | | 职务 | | | 所属院系 | | | 联系电话 | | | 手机 | | E-mail |
| 胡嘉镗 | | 男 | 39 | 副教授 | | | 无 | | | 环境科学与工程学院 | | | 020-84114978 | | | 13533324551 | | hujtang@mail.sysu.edu.cn |
| 研究起止时间 | | | 2021年 3月至 2021年 12月 | | | | | | | | | | | | | | | | |
| 一、申请理由（包括自身具备的知识条件、自己的特长、兴趣及开展本项目的基础等）  杨涵：对物联网及人工智能领域十分感兴趣，具有创新精神。除在环境学院学习环境领域的专业知识外，广泛拓展自己的知识面和培养各项思维，修习《机器学习基础》，《物联网技术与应用导论》，《创新创业基础》等公选课程，均取得不错的成绩。喜欢编程，自修C语言，Python，以及在课程学习了MATLAB。在《机器学习基础》公选课的学习后，对机器学习尤其深度学习等方面知识进行了更深入的学习，除理论的学习之外，也积极尝试结合专业知识进行实际应用。  周鑫：对机器学习领域感兴趣，学习过《机器学习基础》、《面向应用的 Mathematica/Matlab 功能化编程》等课程，有一定的机器学习基础及作图基础。对编程解决问题很感兴趣，学习过多种语言，包括JAVA、python、C/C++等。课余时间较多，乐于学习新的知识和尝试使用机器学习的方法解决本专业以外的问题，拓宽自己的知识范围。  陈至雪：专业方向是计科，有一定的编程基础，熟悉C语言、C++。修读过公选课程《机器学习基础》，对机器学习和深度学习相关知识有所了解，在修读此课程期间也对python进行了基础的学习。在专业课程中学习了MATLAB。希望能够利用自身所学，在不同的领域有所应用和创新。  潘文轩：计算机学院计算机科学与技术专业大二学生。平时积极好学，在学校修学《机器学习基础》公选课程，并在课外自学了一些神经网络相关的知识。本人掌握C/C++，python，latex，markdown等编程语言，学习之余参加数学建模比赛，曾获全国大学生数学建模竞赛（广东赛区）二等奖。同时在实验室学习过一年，具有良好的自学能力和文献阅读能力。希望能在本次比赛中利用自身所学的知识和技能，同队友合作，共同解决跨学科难题。 | | | | | | | | | | | | | | | | | | | |
| **二、项目创新特色概述(50字以内)**  基于三种不同机器学习模型，对香港水域溶解氧进行模拟，并通过设置情景案例对溶解氧长期变化，年际波动的驱动因子进行研究，对比三种模型应用的差异。 | | | | | | | | | | | | | | | | | | | |
| **三、计划项目实施思路（3000字以内）** | | | | | | | | | | | | | | | | | | | |
| 1．研究意义与目的，同类研究工作国内外研究现状与存在的问题等  （1）研究背景  近年来由于人类活动与气候变化，全球海域频繁爆发水体缺氧现象，对其沿岸生态系统产生破坏性的影响，河口及其邻近海域的低氧/缺氧问题由此也成为全球关注的重点。  珠江是中国第二大河流，径流量在全球排第13位，约为330×109m3/a。由于过去几十年工业、农业活动和珠三角城市群的快速发展，珠江口接纳了大量的人为碳物质和营养盐负荷。而随着珠三角经济社会的进一步发展，珠江口的入海营养盐负荷将不断升高。(Strokal, Kroeze et al. 2015)沿海低氧系统中水体养分含量与低氧环境之间的强相关关系已有诸多文献记载与说明。(Li, Bianchi et al. 2011)由此，大量营养元素的输入为珠江河口水域低氧环境形成的重要因素。此外，近年来越来越多的研究表明气候变化助长了沿海海洋缺氧的蔓延。气候变化可以改变海洋环流或垂直分层，从而改变氧源和汇之间的平衡。(Levin, Ekau et al.)除此外，珠江口海域低氧形成还与其水体化学生物过程，有机碳颗粒（POC）等相关。(Li, Song et al. 2019)  本项目将针对位于珠江河口东侧的香港海域近年频繁出现的低氧现象进行探究,并根据探究的结果针对性的提出改善措施，为环境保护及生态修复等相关法律法规的制定提供科学依据。  （2）国内外研究现状及进展  海域水体缺氧现象背后的机理极为复杂，涉及到水质，水生态，水动力过程及生物化学过程等多方面的影响及水体的响应机制。  为探究其形成机制，国内外学者对全球不同海域低氧系统的形成也做出了大量的研究，如Li  等采用化学生物标志物及微化石等对长江口富营养化及低氧水域环境的形成原因进行了探究;(Li, Bianchi et al. 2011);Ni等采用耦合的物理-生物化学地球模型对切萨皮克湾长期低氧现象进行了模拟与探究。(Ni, Li et al. 2020)；  对珠江口低氧形成机制，国内外学者对此也进行了不少研究。  基于在珠江口及香港近岸海域开展的大量调查与分析，Yin等对珠江河口富营养化的过程开  展了全方位的研究,得出了以下河口富营养化模型。如图1所示。(Yin, Xu et al. 2010)  C:\Users\杨涵\AppData\Local\Temp\1615737078(1).png  图1：河口富营养化及生态缓冲概念模型  在学者Wang等人对珠江水质的研究中，使用经过充分验证的物理-生物地球化学模型和DO追  踪法研究缺氧和氧气动力学对河流输入的人为源摄动的响应，分析得出影响夏季珠江河口低氧环境是由珠江水域网络的含低溶解氧，高营养元素，高有机碳的大量淡水的输入所致。(Wang, Hu et al. 2018)  （3）主要存在的问题与不足  前人对珠江水域的时空分布特征，长期变化趋势及其主要影响因素已有所探究，但其采用的传统的统计学习方法，无法对珠江低氧与其影响因子非线性、时滞性、时空异质性的复杂响应关系进行准确地探究与预测。(Wang, Hu et al. 2018)  利用复杂的水质模型是预测水质的主要方法，但是，由于诸如因素如数据限制，大量的计算以及严格的边界条件的限制，这些模型不容易采用，需要极高的计算成本。尽管水质模型如EFDC模型非常有能力模拟水质的复杂动态，但是水质预测仍然难以进行,因其需要大量的时间与数据构建模型，且模型的准确率极度依靠边界条件的确定。(Jiang, Li et al. 2018)作为替代方案，使用数据驱动的方法可以有效的建立水质变量之间的关系，并不需要对边界条件进行划分，从而节省了计算成本。(Liang, Zou et al. 2020)  （4）研究意义  机器学习模型是一类新兴的数据分析技术，该类方法直接从数据中挖掘特征，而无需依赖于预定的方程，其主要优势在于在足够的数据训练后，它能对复杂非线性关系进行准确地拟合。如被誉为最好的机器学习算法之一的随机森林算法（Random forest, RF），它具有如极高的准确率，极强的数据分析挖掘能力，分析复杂相互作用分类特征，以及能给出变量重要性评估等诸多优点。(Iverson, Prasad et al. 2008)  理论上说，如随机森林此类机器学习算法为分析及量化水体富营养化的影响因素及机制提供了新的思路与方法，但目前使用机器学习模型并将其应用于海域或湖泊低氧水体环境形成机理的探究仍鲜有报道。而据目前的研究，此类应用均取得了不错的成果。如罗晓春等使用随机森林算法计算了太湖富营养化各主导气象因子的贡献率(罗晓春, 杭鑫 et al. 2019)；如吴梅林等人使用自组织神经网络评估气候变化和人类活动对南海大亚湾水质的影响。(Wu, Wang et al. 2015)；  在机器学习模型中，最活跃的分支莫过于源于人工神经网络的深度学习，自提出以来就受到学术界和工业界的高度关注。BP（Back Propagation）神经网络是目前应用最广泛的神经网络，也有不少学者将其应用于水质方向的探究，如Lee等将其应用于香港沿海水域的藻华动态模拟模拟(Lee, Huang et al. 2003)除此外，学者潘佳敏等还将BP神经网络与小波分析相耦合后用于云南星云湖富营养化气象因子的分析。小波分析是多分辨率分析方法并且具有良好的局部检测功能，常用于噪声去除、多尺度趋势分析。将其与BP神经网络耦合后，一方面能优化模型的输入数据，另一方面能对构建的多个神经网络进行评价筛选。(潘佳敏，冯禹昊，谢平，方精云, Jiamin et al. 2021)  近些年深度学习理论的突破，不断有更新更有效的神经网络模型被开发及应用。其中长短期记忆神经网络(Long short-term memory, LSTM)常用于时间序列变量的模拟，将尤其适用于对水质的模拟与预测，不过其目前在水质预测的应用极少。(Liang, Zou et al. 2020)  因此，本研究项目于新兴机器学习模型在水质预测上的应用具有探索性意义。  参考文献：  Iverson, L. R., A. M. Prasad, S. N. Matthews and M. Peters (2008). "Estimating potential habitat for 134 eastern US tree species under six climate scenarios." Forest Ecology and Management **254**(3): 390-406.  Jiang, L., Y. Li, X. Zhao, M. R. Tillotson, W. Wang, S. Zhang, L. Sarpong, Q. Asmaa and B. Pan (2018). "Parameter uncertainty and sensitivity analysis of water quality model in Lake Taihu, China." Ecological Modelling **375**: 1-12.  Lee, J. H. W., Y. Huang, M. Dickman and A. W. Jayawardena (2003). "Neural network modelling of coastal algal blooms." Ecological Modelling **159**(2): 179-201.  Levin, L. A., W. Ekau, A. J. Gooday, F. Jorissen, J. J. Middelburg, W. Naqvi, C. Neira, N. N. Rabalais and J. Zhang.  Li, X., T. S. Bianchi, Z. Yang, L. E. Osterman, M. A. Allison, S. F. DiMarco and G. Yang (2011). "Historical trends of hypoxia in Changjiang River estuary: Applications of chemical biomarkers and microfossils." Journal of Marine Systems **86**(3-4): 57-68.  Li, Y., G. Song, P. Massicotte, F. Yang, R. Li and H. Xie (2019). "Distribution, seasonality, and fluxes of dissolved organic matter in the Pearl River (Zhujiang) estuary, China." Biogeosciences **16**(13): 2751-2770.  Liang, Z., R. Zou, X. Chen, T. Ren, H. Su and Y. Liu (2020). "Simulate the forecast capacity of a complicated water quality model using the long short-term memory approach." Journal of Hydrology **581**.  Ni, W., M. Li and J. M. Testa (2020). "Discerning effects of warming, sea level rise and nutrient management on long-term hypoxia trends in Chesapeake Bay." Sci Total Environ **737**: 139717.  Strokal, M., C. Kroeze, L. Li, S. Luan, H. Wang, S. Yang and Y. Zhang (2015). "Increasing dissolved nitrogen and phosphorus export by the Pearl River (Zhujiang): a modeling approach at the sub-basin scale to assess effective nutrient management." Biogeochemistry **125**(2): 221-242.  Wang, B., J. Hu, S. Li, L. Yu and J. Huang (2018). "Impacts of anthropogenic inputs on hypoxia and oxygen dynamics in the Pearl River estuary." Biogeosciences **15**(20): 6105-6125.  Wu, M. L., Y. S. Wang and J. D. Gu (2015). "Assessment for water quality by artificial neural network in Daya Bay, South China Sea." Ecotoxicology **24**(7-8): 1632-1642.  Yin, K., J. Xu and P. Harrison (2010). A Comparison of Eutrophication Processes in Three Chinese Subtropical Semi-Enclosed Embayments with Different Buffering Capacities. Coastal Lagoons**:** 367-397.  罗晓春, 杭鑫, 曹云, 杭蓉蓉, 李亚春, X. LUO, X. HANG, Y. CAO, R. HANG and Y. LI (2019). "太湖富营养化条件下影响蓝藻水华的主导气象因子  Dominant meteorological factors affecting cyanobacterial blooms under eutrophication in Lake Taihu." 湖泊科学 **31**(5): 1248-1258.  潘佳敏，冯禹昊，谢平，方精云, P. Jiamin, F. Yuhao, X. Ping and F. Jingyun (2021). "基于小波-神经网络耦合模型对云南星云湖富营养化气象驱动因子的分析  Meteorological driving factor of eutrophication in Lake Xingyun, Yunnan based on the Coupling Model of Wavelet analysis and Neural Network." 湖泊科学 **33**(2): 428-438. | | | | | | | | | | | | | | | | | | | |
| 2． 研究内容及工作方案  （1）研究内容  本项目拟利用香港水域1986-2018于七十六个香港环保署水质监测站点的现场观测数据。  该数据源存在以下问题。   1. 数据不均匀：不同站点总采样的时间点数与时间分布不同。 2. 数据不连续：并非所有站点都具有三十年来每个月的水质信息，数据缺失现象严重。而对LSTM模型而言，不连续的数据将很大程度上影响输出结果。因此在训练数据之前，需采取合适方法填补缺失值。 3. 数据异常值现象：部分站点的采样数据出现异常，所测值为非数值，需要在训练数据之前使用月平均值或中值等进行替换。 4. 数据分散，共含76个站点的数据，而76个站点又对水域分顶部，中部，底部三层取样。76个站点其不同取样层的数据无法同时作为训练数据送入神经网络。而根据机器学习中“没有免费的午餐”原则，不同训练数据的神经网络模型的最优参数不同。一方面训练上百个模型的计算成本及耗费精力过于巨大，另一方面并非所有站点的数据都适合进行训练。如DM2站点32年间中层和底层仅有一次观测记录。故在训练数据之前，需对站点进行筛选。   以上内容为数据的预处理部分研究工作。需研究如何采用合适的方法填补缺失值，替换异常  值；如何有效筛选站点数据，确保其能代表珠江口东侧海域溶解氧的实际情况等问题。  数据清洗后的研究内容是模型框架的建立。  首先基于新兴的机器学习模型长短期记忆网络LSTM，对香港水域溶解氧进行模拟。LSTM模型的优势在于它能对时间序列变量进行十分准确的预测与模拟，但它无法如随机森林模型一般得出各变量因子的贡献率。由此需要设置情景案例单独探究这些变量对溶解氧长期变化，年纪波动的驱动作用。  神经网络的平均影响值法（简称MIV算法）被认为是在神经网络模型中进行敏感变量筛选和评价变量重要性的最好方法之一。具体过程是，将训练数据的每个变量因子等比例增大或缩小相同倍数，得到两组新训练数据的自变量，代入用原始数据训练好的神经网络，分别得到两组数据的估计结果，而这两组估计结果的差值被称为IV，对结果取平均即得到平均影响值（MIV）,依次算出各个自变量的MIV，就能确定各个输入变量对网络结果的影响程度。此即情景案例的具体含义及其过程，等比例改变训练数据，计算其对结果的影响程度。  使用小波-神经网络耦合模型时，同样需要设置情景案例对驱动因子进行探究。  对于随机森林模型，则可通过模型自身的RI指数(Relative Importance)直接判定变量的驱动作用。不过，在进行模型之间的比较时，可能也需要计算随机森林模型的MIV值，即也需要对随机森林模型设置情景案例。  总得来说，需要处理的工作包括，模型训练之前的数据清洗，不同模型框架的建立，情景案例的设定与分析，绘制图表及分析，以及最后论文的撰写与答辩准备。  （2）技术路线  C:\Users\杨涵\AppData\Local\Temp\1615768874(1).png  图2 本项目的技术路线图 | | | | | | | | | | | | | | | | | | | |
| 3．拟解决主要问题  （1）预测珠江口东侧溶解氧变化趋势，并在不同情景条件下进行模拟。  （2）探究珠江口东侧溶解氧长期变化，年际波动的驱动因子。  （3）探究不同模型在珠江溶解氧模拟预测的准确度差异，以及依据情境条件模拟探究驱动因素影响程度的差异**。** | | | | | | | | | | | | | | | | | | | |
| 4．研究支持条件（使用仪器或设备）  无特殊使用设备。  使用软件包括MATLAB, Python, SPSS等。 | | | | | | | | | | | | | | | | | | | |
| 5．预期研究成果（研究论文、设计、专利、产品、鉴定、推广应用等）    《基于机器学习的珠江口东侧水域低氧长期变化规律与驱动因素研究》论文一篇 | | | | | | | | | | | | | | | | | | | |
| 6．项目创新之处（原始创新：重大科学发现、技术发明；集成创新：融合多种相关技术，形成新产品、新产业；引进消化吸收再创新：在引进国内外先进技术的基础上，学习、分析、借鉴，形成具有自主知识产权的新技术）    创新性地使用了新型模型进行水质预测。如长短期记忆网络LSTM为新兴的机器学习模型，具有很强的模拟预测能力，在各行各业都大放异彩，但在环境领域的应用还鲜有探究，目前在水质预测方面的应用更是极少。另一方面，珠江东侧河口低溶解氧的原因尚未得到定量说明，水质预测具有极强的时空异质性，不同的地域需要定制不同的模型。无论是随机森林，小波-神经网络耦合模型，还是更新的LSTM，将其应用于香港水域进行溶解氧模拟都是创新性研究。 | | | | | | | | | | | | | | | | | | | |
| **四、计划项目管理（800 字以内）** | | | | | | | | | | | | | | | | | | | |
| 1、计划项目人员分工  该项目的工作划分为四个模块。  一是模型把控，由组长杨涵进行。工作包括对整个项目实验框架的构建，情景案例的设定，与使用机器学习完成实验任务的具体思路等。  二是数据处理，由组员陈至雪进行。主要工作为数据的预处理工作，对数据中异常值与缺失值进行清洗。  三是图表绘制，由组员潘文轩及周鑫负责。主要工作为模型所得数据可视化，便于展示与分析。  四是数据分析，由团队讨论分析，揭示模型背后的环境学意义。 | | | | | | | | | | | | | | | | | | | |
| 2、计划项目研究时间安排（查阅资料 、选题、项目研究方案、开题报告、实验研究、数据统计、处理与分析、研制开发、结题、撰写研究论文和总结报告、参加结题答辩和成果推广等）     |  |  | | --- | --- | | **时间** | **目标** | | 2021.3-2021.4 | 查阅资料，确定研究方案 | | 2021.4-2021.5 | 实现不同情境条件下LSTM模型对珠江口东侧水域溶解氧的模拟预测并探究其驱动因素 | | 2021.5-2021.7 | 实现不同情境条件随机森林模型对珠江口东侧水域溶解氧的模拟预测并探究其驱动因素 | | 2021.7-2021.9 | 实现不同情境条件小波-神经网络耦合模型对珠江口东侧水域溶解氧的模拟预测并探究其驱动因素 | | 2021.9-2021.10 | 对比三种模型差异 | | 2021.10-2021.11 | 完成论文撰写 | | 2021.11-2021.12 | 论文润色及准备答辩 | | | | | | | | | | | | | | | | | | | | |
| 1. **项目成员是否有申请延期的大学生创新创业训练计划项目（**如有，请填写项目成员姓名、立项时间、立项等级、项目名称和指导教师姓名   无 | | | | | | | | | | | | | | | | | | | |
| 1. **经费预算及使用计划**   总计： 15000 元人民币 | | | | | | | | | | | | | | | | | | | |
| 序号 | | 支出项目 | | | | | | | 经费测算依据 | | | | | | 金额（元） | | | | |
| 1 | | 复印与打印费 | | | | | | | 资料复印和打印费 | | | | | | 1000 | | | | |
| 2 | | 测试费 | | | | | | | 机器学习模型运行的测试费 | | | | | | 10000 | | | | |
| 3 | | 软件权限购买费 | | | | | | | 非开源网络资源获取所需的费用 | | | | | | 4000 | | | | |
| 经费使用计划：  申请团队签名： | | | | | | | | | | | | | | | | | | | |
| **指导教师意见** | 签名：  年 月 日 | | | | | | | | | | | | | | | | | | |
| **院系意见** | 签名（公章）：  年 月 日 | | | | | | | | | | | | | | | | | | |
| **学校审核意见** | 签名（公章）：  年 月 日 | | | | | | | | | | | | | | | | | | |
| **备注** |  | | | | | | | | | | | | | | | | | | |