MINISTRY OF SCIENCE AND HIGHER EDUCATION OF THE REPUBLIC OF KAZAKHSTAN

ҚАЗАҚСТАН РЕСПУБЛИКАСЫНЫҢ ҒЫЛЫМ ЖӘНЕ ЖОҒАРЫ БІЛІМ МИНИСТРЛІГІ

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РЕСПУБЛИКИ КАЗАХСТАН

KAZAKHSTAN

ХАЛЫҚАРАЛЫҚ УНИВЕРСИТЕТІ

INTERNATIONAL UNIVERSITY SINCE 2009

МЕЖДУНАРОДНЫЙ УНИВЕРСИТЕТ

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

Published since 2020.
Volume 7. 3 (27). 2026
July–September

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

2020 жылдан бері шығарылады
Том 7. 3 (27). 2026
Шілде-Қыркүйек

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

Издается с 2020 г.
Том 7. 3 (27). 2026
Июль-Сентябрь

Свидетельство о постановке на учет периодического печатного издания в Министерство информации и общественного развития Республики Казахстан № KZ82VPY00020475, выданное от 20.02.2020 г.

Зарегистрировано в Международном центре регистрации серийных изданий ISSN (ЮНЕСКО, Париж, Франция). ISSN 2708–2032 (print), ISSN 2708–2040 (online)

Журнал входит в Перечень научных изданий, рекомендуемых КОКНВО МНВО РК для публикации основных результатов научной деятельности.

EDITOR-IN-CHIEF:
Kateryna Kolesnikova — Doctor of Technical Sciences, professor, Vice-Rector for Research, International Information Technology University (Kazakhstan)

DEPUTY EDITOR-IN-CHIEF:
Madina Ipalakova — Candidate of Technical Sciences, associate professor, Director of the Research Department, International Information Technology University (Kazakhstan)

EDITORIAL BOARD:
Abdul Razak — PhD, professor, Department of Cybersecurity, International Information Technology University (Kazakhstan)
Lucio Tommaso De Paolis — Director of the R&D Department of the AVR Laboratory, Department of Engineering for Innovation, University of Salento (Italy)
Liz Bacon — Professor, Deputy Vice-Chancellor, Abertay University (United Kingdom)
Michele Pagano — PhD, Professor, University of Pisa (Italy)
Mukhtarbay Otelbayev — Doctor of Physical and Mathematical Sciences, professor, academician of the National Academy of Sciences of the Republic of Kazakhstan, professor of the Department of Mathematical and Computer Modeling, International Information Technology University (Kazakhstan)
Bolatbek Rysbaiuly — Doctor of Physical and Mathematical Sciences, professor, professor of the Department of Computing and Data Science, Astana IT University (Kazakhstan)
Yevgeniya Daineko — PhD, research professor, Department of Information Systems, International Information Technology University (Kazakhstan)
Nurzhan Duzbayev — PhD, associate professor, Vice-Rector for Digitalization and Innovation, International Information Technology University (Kazakhstan)
Bakhtgerei Sinchev — Doctor of Technical Sciences, professor, Department of Information Systems, International Information Technology University (Kazakhstan)
Nurgul Seilova — Candidate of Technical Sciences, Dean of the Faculty of Computer Technologies and Cybersecurity, International Information Technology University (Kazakhstan)
Ardak Mukhamediyeva — Candidate of Economic Sciences, Dean of the Faculty of Business, Media and Management, International Information Technology University (Kazakhstan)
Zamira Abdikalikova — PhD, associate professor, Head of the Department of Mathematical and Computer Modeling, International Information Technology University (Kazakhstan)
Yerlan Shildibekov — PhD, associate professor, Head of the Department of Economics and Business, International Information Technology University (Kazakhstan)
Damilya Yeskendirova — Candidate of Technical Sciences, associate professor, Head of the Department of Cybersecurity, International Information Technology University (Kazakhstan)
Aigul Niyazgulova — Candidate of Philological Sciences, Professor, Head of the Department of Media Communications and History of Kazakhstan, International Information Technology University (Kazakhstan)
Altai Aitmagambetov — Candidate of Technical Sciences, Professor, Department of Radio Engineering, Electronics and Telecommunications, International Information Technology University (Kazakhstan)
Yelena Bakhtiyarova — Candidate of Technical Sciences, associate professor, Head of the Department of Radio Engineering, Electronics and Telecommunications, International Information Technology University (Kazakhstan)
Kanibek Sansyzbay — PhD, research professor, Department of Cybersecurity, International Information Technology University (Kazakhstan)
Sakhybay Tynymbayev — Candidate of Technical Sciences, Professor, Research Professor, Department of Computer Engineering, International Information Technology University (Kazakhstan)
Ali Abd Almisreb — PhD, associate professor, Department of Cybersecurity, International Information Technology University (Kazakhstan)
Mohamed Ahmed Hamada — PhD, associate professor, Department of Information Systems, International Information Technology University (Kazakhstan)
Yang Im Chu — PhD, Professor, Gachon University (South Korea)
Tadeusz Wallas — PhD, Vice-Rector, Adam Mickiewicz University (Poland)
Orken Mamyrbayev — PhD, Deputy Director for Science, RSE Institute of Information and Computational Technologies, Committee for Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Kazakhstan)
Sergey Bushuyev — Doctor of Technical Sciences, professor, Director of the Ukrainian Project Management Association “UKRNET,” Head of the Department of Project Management, Kyiv National University of Construction and Architecture (Ukraine)
Svetlana Beloshitskaya — Doctor of Technical Sciences, professor, Department of Computing and Data Science, Astana IT University (Kazakhstan)

MANAGING EDITOR
Raushan Mrzabayeva — Master of Science, editor, International Information Technology University (Kazakhstan)

International Journal of Information and Communication Technologies
Periodicity: 4 times a year.
Languages: Kazakh, Russian, English
DOI prefix: 10.54309
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Thematic focus: "Information technology"; "Digital technologies in the development of socio-economic systems"; "Information security and communication technologies".
Distribution: Materials are distributed under the Creative Commons Attribution 4.0
Journal website: https://journal.iitu.edu.kz
Owner: International Information Technology University JSC (Almaty).
Copyright: © International Journal of Information and Communication Technologies, 2026

РЕДАКЦИЯ

БАС РЕДАКТОР:

Колесникова Катерина Викторовна — техника ғылымдарының докторы, профессор, Халықаралық ақпараттық технологиялар университетінің ғылыми-зерттеу қызметі жөніндегі проректор (Қазақстан)

БАС РЕДАКТОРДЫҢ ОРЫНБАСАРЫ:

Ипалакова Мадина Тулегеновна — техника ғылымдарының кандидаты, қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университетінің ғылыми-зерттеу қызметі жөніндегі департамент директоры (Қазақстан)

РЕДАКЦИЯЛЫҚ АЛҚА:

Разак Абдул — PhD, Халықаралық ақпараттық технологиялар университеті киберқауіпсіздік кафедрасының профессоры (Қазақстан)

Лучио Томмазо де Паолис — Саленто Университеті (Италия) инновация және технологиялық инжиниринг департаменті AVR зертханасының зерттеу және әзірлеу бөлімінің директоры

Лиз Бэкон — профессор, Абертей Университеті (Ұлыбритания) вице-канцлерінің орынбасары

Микеле Пагано — PhD, Пиза Университетінің (Италия) профессоры

Өтелбаев Мухтарбай Өтелбайұлы — физика-математика ғылымдарының докторы, профессор, ҚР ҰҒА академигі, Халықаралық ақпараттық технологиялар университеті математика және компьютерлік модельдеу кафедрасының профессоры (Қазақстан)

Рысбайулы Болатбек — физика-математика ғылымдарының докторы, профессор, Есептеу және деректер ғылымдары департаментінің профессоры, Astana IT University (Казахстан)

Дайнеко Евгения Александровна — PhD, Халықаралық ақпараттық технологиялар университеті ақпараттық жүйелер кафедрасының профессор-зерттеушісі (Қазақстан)

Дузбаев Нуржан Токкужаевич — PhD, қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университеті цифрландыру және инновациялар жөніндегі проректор (Қазақстан)

Синчев Бахтгерей Куспанович — техника ғылымдарының докторы, профессор, Халықаралық ақпараттық технологиялар университеті ақпараттық жүйелер кафедрасының профессоры (Қазақстан)

Сейлова Нургуль Абадуллаевна — техника ғылымдарының докторы, Халықаралық ақпараттық технологиялар университеті компьютерлік технологиялар және киберқауіпсіздік факультетінің деканы (Қазақстан)

Мухамедиева Ардак Габитовна — экономика ғылымдарының кандидаты, Халықаралық ақпараттық технологиялар университеті бизнеса медиа және басқару факультетінің деканы (Қазақстан)

Абдикаликова Замира Турсынбаевна — PhD, қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университеті математика және компьютерлік модельдеу кафедрасының меңгерушісі (Қазақстан)

Шильдибеков Ерлан Жаржанович — PhD, қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университеті экономика және бизнес кафедрасының меңгерушісі (Қазақстан)

Дамеля Максутовна Ескендирова — техника ғылымдарының кандидаты, қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университеті киберқауіпсіздік кафедрасының меңгерушісі (Қазақстан)

Ниязгулова Айгуль Аскарбековна — филология ғылымдарының кандидаты, доцент, профессор, Халықаралық ақпараттық технологиялар университеті медиакоммуникация және Қазақстан тарихы кафедрасының меңгерушісі (Қазақстан)

Айтмагамбетов Алтай Зуфарович — техника ғылымдарының кандидаты, Халықаралық ақпараттық технологиялар университеті радиотехника, электроника және телекоммуникация кафедрасының профессоры (Қазақстан)

Бахтиярова Елена Ажибековна — техника ғылымдарының кандидаты, қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университеті радиотехника, электроника және телекоммуникация кафедрасының меңгерушісі (Қазақстан)

Канибек Сансызбай — PhD, қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университеті киберқауіпсіздік кафедрасының профессор-зерттеушісі (Қазақстан)

Тынымбаев Сахибай — техника ғылымдарының кандидаты, профессор, Халықаралық ақпараттық технологиялар университеті компьютерлік инженерия кафедрасының профессор-зерттеушісі (Қазақстан)

Алмисреб Али Абд — PhD, Халықаралық ақпараттық технологиялар университеті киберқауіпсіздік кафедрасының қауымдастырылған профессоры (Қазақстан)

Мохамед Ахмед Хамада — PhD, Халықаралық ақпараттық технологиялар университеті ақпараттық жүйелер кафедрасының қауымдастырылған профессоры (Қазақстан)

Янг Им Чу — PhD, Гачон университетінің профессоры (Оңтүстік Корея)

Тадеуш Валлас – PhD, Адам Мицкевич атындағы (Польша) университеттің проректоры

Мамырбаев Оркен Жумажанович — PhD, ҚР ҒЖБМ Ғылым комитеті ақпараттық және есептеу технологиялары институты ӨМК директорының ғылым жөніндегі орынбасары (Қазақстан)

Бушуев Сергей Дмитриевич — техника ғылымдарының докторы, профессор, Украинаның "УКРНЕТ" жобаларды басқару қауымдастығының директоры, Киев ұлттық құрылыс және сәулет университеті жобаларды басқару кафедрасының меңгерушісі (Украина)

Белощицкая Светлана Васильевна — техника ғылымдарының докторы, доцент, Astana IT University есептеу және деректер ғылымы кафедрасының профессоры (Қазақстан)

ЖАУАПТЫ РЕДАКТОР:

Мрзабаева Раушан Жалиевна — магистр, Халықаралық ақпараттық технологиялар университетінің редакторы (Қазақстан)

Халықаралық ақпараттық және коммуникациялық технологиялар журналы
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Префикс DOI: 10.54309
Мерзімділігі: жылына 4 рет.
Басылым тілі: қазақ, орыс, ағылшын.
Тақырып бағыты: "Ақпараттық технологиялар"; "Ақпараттық қауіпсіздік және коммуникациялық технологиялар"; "Әлеуметтік-экономикалық жүйелерді дамытудағы цифрлық технология".
Журнал сайты: https://journal.iitu.edu.kz
Тарату: материалдар Creative Commons Attribution 4.0 лицензиясы бойынша таратылады
Меншік иесі: АҚ «Халықаралық ақпараттық технологиялар университеті» (Алматы қ.).
Авторлық құқық: © Халықаралық ақпараттық және коммуникациялық технологиялар журналы, 2026

РЕДАКЦИЯ

ГЛАВНЫЙ РЕДАКТОР:

Колесникова Катерина Викторовна — доктор технических наук, профессор, проректор по научно-исследовательской деятельности Международного университета информационных технологий (Казахстан)

ЗАМЕСТИТЕЛЬ ГЛАВНОГО РЕДАКТОРА:

Ипалакова Мадина Тулегеновна — кандидат технических наук, ассоциированный профессор, директор департамента по научно-исследовательской деятельности Международного университета информационных технологий (Казахстан)

РЕДАКЦИОННАЯ КОЛЛЕГИЯ:

Разак Абдул — PhD, профессор кафедры кибербезопасности Международного университета информационных технологий (Казахстан)

Лучио Томмазо де Паолис — директор отдела исследований и разработок лаборатории AVR департамента инноваций и технологического инжиниринга Университета Саленто (Италия)

Лиз Бэкон — профессор, заместитель вице-канцлера Университета Абертей (Великобритания)

Микеле Пагано — PhD, профессор Университета Пизы (Италия)

Отелбаев Мухтарбай Отелбайулы — доктор физико-математических наук, профессор, академик НАН РК, профессор кафедры математического и компьютерного моделирования Международного университета информационных технологий (Казахстан)

Рысбайулы Болатбек — доктор физико-математических наук, профессор, профессор Astana IT University (Казахстан)

Дайнеко Евгения Александровна — PhD, профессор-исследователь кафедры информационных систем Международного университета информационных технологий (Казахстан)

Дузбаев Нуржан Токкужаевич — PhD, ассоциированный профессор, проректор по цифровизации и инновациям Международного университета информационных технологий (Казахстан)

Синчев Бахтгерей Куспанович — доктор технических наук, профессор, профессор кафедры информационных систем Международного университета информационных технологий (Казахстан)

Сейлова Нургуль Абадуллаевна — кандидат технических наук, декан факультета компьютерных технологий и кибербезопасности Международного университета информационных технологий (Казахстан)

Мухамедиева Ардак Габитовна — кандидат экономических наук, декан факультета бизнеса медиа и управления Международного университета информационных технологий (Казахстан)

Абдикаликова Замира Турсынбаевна — PhD, ассоциированный профессор, заведующая кафедрой математического и компьютерного моделирования Международного университета информационных технологий (Казахстан)

Шильдибеков Ерлан Жаржанович — PhD, ассоциированный профессор, заведующий кафедрой экономики и бизнеса Международного университета информационных технологий (Казахстан)

Дамеля Максутовна Ескендирова — кандидат технических наук, ассоциированный профессор, заведующая кафедрой кибербезопасности Международного университета информационных технологий (Казахстан)

Ниязгулова Айгуль Аскарбековна — кандидат филологических наук, доцент, профессор, заведующая кафедрой медиакоммуникаций и истории Казахстана Международного университета информационных технологий (Казахстан)

Айтмагамбетов Алтай Зуфарович — кандидат технических наук, профессор кафедры радиотехники, электроники и телекоммуникаций Международного университета информационных технологий (Казахстан)

Бахтиярова Елена Ажибековна — кандидат технических наук, ассоциированный профессор, заведующая кафедрой радиотехники, электроники и телекоммуникаций Международного университета информационных технологий (Казахстан)

Канибек Сансызбай – PhD, ассоциированный профессор, профессор-исследователь кафедры кибербезопасности, Международного университета информационных технологий (Казахстан)

Тынымбаев Сахибай – кандидат технических наук, профессор, профессор-исследователь кафедры компьютерной инженерии, Международного университета информационных технологий (Казахстан)

Алмисреб Али Абд — PhD, ассоциированный профессор кафедры кибербезопасности Международного университета информационных технологий (Казахстан)

Мохамед Ахмед Хамада — PhD, ассоциированный профессор кафедры информационных систем Международного университета информационных технологий (Казахстан)

Янг Им Чу — PhD, профессор университета Гачон (Южная Корея)

Тадеуш Валлас – PhD, проректор университета имен Адама Мицкевича (Польша)

Мамырбаев Оркен Жумажанович — PhD, заместитель директора по науке РГП Института информационных и вычислительных технологий Комитета науки МНВО РК (Казахстан)

Бушуев Сергей Дмитриевич — доктор технических наук, профессор, директор Украинской ассоциации управления проектами «УКРНЕТ», заведующий кафедрой управления проектами Киевского национального университета строительства и архитектуры (Украина)

Белощицкая Светлана Васильевна — доктор технических наук, доцент, профессор кафедры вычислений и науки о данных Astana IT University (Казахстан)

ОТВЕТСТВЕННЫЙ РЕДАКТОР:

Мрзабаева Раушан Жалиевна — магистр, редактор Международного университета информационных технологий (Казахстан)

Международный журнал информационных и коммуникационных технологий
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Префикс DOI: 10.54309
Периодичность: 4 выпусков в год.
Язык издания: казахский, русский, английский.
Тематическая направленность: "Информационные технологии"; "Информационная безопасность и коммуникационные технологии"; "Цифровые технологии в развитии социо-экономических систем".
Сайт журнала: https://journal.iitu.edu.kz
Распространение: материалы распространяются по лицензии Creative Commons Attribution 4.0
Собственник: АО «Международный университет информационных технологий» (г. Алматы).
Авторские права: © Международный журнал информационных и коммуникационных технологий, 2026

CONTENTS

DIGITAL TECHNOLOGIES IN THE DEVELOPMENT OF SOCIO-ECONOMIC SYSTEMS

A Adalbek, S.Ye. Kerimkhulle, K.D. Kuliev, O. Mekesh, Zh.Zh. Khamitova
DEVELOPMENT OF RESULT-ORIENTED DATA SMOOTHING MODELS AND ANALYSIS OF THEIR EFFECTIVENESS 11
A.T. Iskakova, G.E. Kurmankulova, K. Gromaszek, K.O. Togzhanova, R.S. Algozhayeva
APPLICATION OF MACHINE LEARNING AND DEEP LEARNING FOR PREDICTIVE PLANNING AND RESOURCE EFFICIENCY IN PRODUCTION SYSTEMS 30
A.B. Kassymova, R.K. Uskenbayeva, Zh.A. Sydykov, D. Omirzak, S. Kalibekov
ARCHITECTURE OF AN INTELLIGENT CHATBOT FOR SMART CITY: AI-BASED RECEIVING, ANALYSIS AND ROUTING OF CITIZEN REQUESTS 46
O.M. Olzhayev, B.S. Omarov, A. Suliman
YOLO-CBAM: A LIGHTWEIGHT NEURAL NETWORK MODEL WITH INTEGRATED ATTENTION MECHANISM FOR ROAD DAMAGE ANALYSIS 66
K. Yermagambetova, A. Sarsekeyeva, Y. Akhlassov
OVERCOMING FRAGMENTATION IN THE CLASSIFICATION OF DIGITAL CULTURAL HERITAGE: AN INTEGRATIVE MULTIDIMENSIONAL MODEL 78

INFORMATION TECHNOLOGY

D. Abzhanova, S. Sharipova
ARCHITECTURE AND PRACTICAL IMPLEMENTATION OF A DIGITAL PLATFORM FOR MONITORING AND DATA EXCHANGE ON TRANSBOUNDARY WATER RESOURCES IN THE ZHAMBYL REGION 101
A. Aitanuly, O.A. Ussatova
ENHANCING TWO-FACTOR AUTHENTICATION FOR CLOUD HEALTHCARE SECURITY: THE ROLE OF DEEP LEARNING 122
O. Alseitov, O. Kuchanskyi
PHYSICS-INFORMED HYBRID LEARNING FOR HOURLY POWER-PLANT EMISSIONS: A MULTI-POLLUTANT BENCHMARK WITH UNCERTAINTY AND TRANSFER 141
A.M. Akhmetova, D.O. Baizhanova, Zh.N. Alisheva, B.A. Shaimov
ADAPTIVE OUTPUT ROUTER FOR NATURAL LANGUAGE MODELS CONSIDERING KEY PERFORMANCE INDICATORS 160
A.K. Aitim
AUTOMATIC AND SEMI-AUTOMATIC ANNOTATION OF KAZAKH TEXT CORPORA 180
Z.E. Balabayev, S.V. Biloshchytska
HANDWRITTEN TEXT RECOGNITION FOR CHAGATAI MANUSCRIPTS: COMPARING CRNN TRAINING FROM SCRATCH AND TRANSFORMER-BASED TRANSFER LEARNING 198
D.B. Baumuratova, T.K. Zhukabayeva, K.A. Abdrakhmanov, Z. Ahmad, A.A. Seitov
A NEURO-DIGITAL LEARNING MODEL FOR AI DRIVEN DIGITAL TRANSFORMATION IN HIGHER EDUCATION 219
T. Daoyun, G.A. Amirkhanova, Fu Yanwei, Y. Faruzkyzy
MULTIMODAL SEMANTIC MANIPULATION IN FAKE VIDEOS AND ITS INFLUENCE MECHANISM ON USERS' PERCEIVED TRUST 241
Zh. Doumchariyev, J. Tussupov, A. Akynbekova, M. Khassanova, M. Kassim
ENSEMBLE AND HYBRID LEARNING FOR DIGITAL TWIN MILK CLASSIFICATION 258
K.Z. Igibayev, A.Kh. Kozbakova, O. Auelbekov , T. Merembayev, K. Yessentayev
STUDY OF A DYNAMIC MIXING DEVICE IN A BIOREACTOR USING IOT TECHNOLOGIES 279
A.N. Ismukanova, G. Kochshanova, D.V. Pleskachev, Zh. Sarsenbayeva, P. Schmidt
AUTOMATED CLASSIFICATION OF SCIENTIFIC DATA USING ARTIFICIAL INTELLIGENCE: A COMPARATIVE ANALYSIS OF TRANSFORMER ARCHITECTURES AND HYBRID APPROACHES 306
T. Zh. Kamalov, Ali Abd. Almisreb, Mohammed A. Saleh
INVESTIGATION OF HYBRID QUANTUM-CLASSICAL ALGORITHMS FOR MACHINE LEARNING 323
B. Karimkyzy, K.G. Karasenko, N.A. Alisherov, N.M. Muratbek, A. Karim

KEY PUBLICATIONS OF 2021–2025 ON ARTIFICIAL INTELLIGENCE METHODS IN THE MANUFACTURING AND DIAGNOSTICS OF MEDICAL DEVICES FOR GLYCEMIC MONITORING: A THEMATIC ANALYTICAL REVIEW 338
S.K. Kumyzbaeva, K.E. Baglanbaev, K.S. Sarkyt, M.Zh. Konyrova
THREE-PHASE TRAINING OF INTELLIGENT ENERGY MANAGEMENT FOR A HYBRID ENERGY SYSTEM WITH A CLIMATE-AWARE STATE REPRESENTATION 357
N.E. Karabayev, S.K. Serikbayeva, E. Benkhelif, B. Tassuov
RESEARCH ON THE ROBUSTNESS OF ENSEMBLE MACHINE LEARNING MODELS TO ADVERSARIAL ATTACKS IN INTRUSION DETECTION SYSTEMS FOR IOT NETWORKS 369
M. Nurtas, G. Kapyshev
A METHODOLOGICAL FRAMEWORK FOR THE PREPROCESSING AND INTEGRATION OF GEOSPATIAL AND GEOPHYSICAL DATA FOR MINERAL PROSPECTIVITY PREDICTION USING MACHINE LEARNING 392
A. Neftissov, A. Biloshchytskyi, I. Kazambayev, A. Medetkhan, L. Kirichenko
ENERGY EFFICIENT MICROCLIMATE STABILIZATION FOR SEALED ARCHIVAL STORAGE ENCLOSURES 405
A.Zh. Omurtayeva, U.T. Makhazhanova, A.K. Adilova, Sh.K. Yelezhanova, J. Kultan
AN ADAPTIVE NEURAL-FUZZY MODEL FOR ASSESSING INVESTMENT RISKS IN THE AGRI-INDUSTRIAL COMPLEX: ODERNISATION OF THE DECISION-SUPPORT ARCHITECTURE THROUGH THE INTEGRATION OF MACHINE LEARNING 421
N. Shazhdekeyeva, A. Mukhanbetkaliyeva, A. Abdigaliyeva, A. Zhanekeshova, L. Nurgaliyeva
COMPARATIVE ANALYSIS OF MATHEMATICAL MODELS OF COLLABORATIVE FILTERING IN THE DEVELOPMENT OF A MOVIE RECOMMENDER SYSTEM 445
S. Sharmukhanbet, G. Turmukhanova, O. Findik, V. Makhatova, M. Rakhmetov
AN INTELLIGENT PIPELINE DEFECT DETECTION SYSTEM USING AN IN-PIPE ROBOT 461

INFORMATION SECURITY AND COMMUNICATIONTECHNOLOGIES

S.Z. Daruish, A.U. Utegenova, A.U. Kalizhanova, Zh.Zh. Zhumabekova
MACHINE LEARNING FOR POLLUTION CLASSIFICATION IN ROCKET BOOSTER IMPACT AREAS 480
S.N. Issabayeva, A.K. Rysbayeva, Zh.B. Ibrayeva, N.R. Mussabekov, N.Sh. Akimzhan
MACHINE LEARNING-BASED CYBERATTACK DETECTION IN OT-NETWORKS: A COMPARATIVE ANALYSIS OF NETWORK ANOMALY CLASSIFICATION METHODS 495
G. Shangytbayeva, A.N. Amanov, M.K. Issenova, G.E. Berikkhanova, K.O. Shakerkhan
INFORMATION SECURITY IN GOVERNMENT CLOUD SERVICES: RISK ASSESSMENT AND PROTECTION STRATEGIES 512
Ye. N. Tursynbek, N. Albanbay, D. Djenouri
OPTIMISATION OF COMMUNICATION COSTS IN TWO-TIER FEDERATED LEARNING FOR CYBERATTACK DETECTION IN IOT NETWORKS 529
A. Tastemirov, S.B. Mukhanov, S.T. Amanzholova, B. Aldosh
A COMPARATIVE EVALUATION OF MACHINE LEARNING MODELS FOR ACTIVE DIRECTORY THREAT DETECTION: FEATURE-LEAKAGE CONTROL AND CROSS-SCENARIO VALIDATION ON WINDOWS SECURITY EVENT LOGS 545

МАЗМҰНЫ
ӘЛЕУМЕТТІК-ЭКОНОМИКАЛЫҚ ЖҮЙЕЛЕРДІ ДАМЫТУДАҒЫ ЦИФРЛЫҚ ТЕХНОЛОГИЯЛАР

А Адалбек, С.Е. Керімқұл, К.Д. Кулиев, О. Мекеш, Ж.Ж. Хамитова
ДЕРЕКТЕРДІҢ НӘТИЖЕ БАҒДАРЛАНҒАН ТЕГІСТЕУ МОДЕЛЬДЕРІН ӘЗІРЛЕУ ЖӘНЕ ОЛАРДЫҢ ТИІМДІЛІГІН ТАЛДАУ 11
А.Т. Искакова, Г.Е. Курманкулова, Конрад Громашек, К.О. Тогжанова, Р.С. Алгожаева
ӨНДІРІС ЖҮЙЕЛЕРІНДЕ БОЛЖАМДЫ ЖОСПАРЛАУ ЖӘНЕ РЕСУРСТАРДЫ ҮНЕМДЕУ ҮШІН МАШИНАЛЫҚ ЖӘНЕ ТЕРЕҢ ОҚЫТУДЫ ҚОЛДАНУ 30
А.Б. Касымова, Р.К. Ускенбаева, Ж.А. Сыдыков, Д. Өмірзақ, С. Калибеков
SMART CITY ҮШІН ИНТЕЛЛЕКТУАЛДЫ ЧАТ-БОТ АРХИТЕКТУРАСЫ: АЗАМАТТАРДЫҢ

ӨТІНІШТЕРІН ЖАСАНДЫ ИНТЕЛЛЕКТ НЕГІЗІНДЕ ҚАБЫЛДАУ, ТАЛДАУ ЖӘНЕ БАҒЫТТАУ 46
О.М. Олжаев, Б.С. Омаров, A.Suliman
YOLO-CBAM: ЖОЛДЫҢ ЗАҚЫМДАНУЫН ТАЛДАУҒА АРНАЛҒАН ИНТЕГРАЦИЯЛАНҒАН ЗЕЙІН МЕХАНИЗМІ БАР ЖЕҢІЛ САЛМАҚТЫ НЕЙРОНДЫҚ ЖЕЛІ МОДЕЛІ 66
К. Ермагамбетова, А. Сарсекеева, Е. Ахласов
ЦИФРЛЫҚ МӘДЕНИ МҰРАНЫ ЖІКТЕУДЕГІ ФРАГМЕНТАЦИЯНЫ ЕҢСЕРУ: ИНТЕГРАТИВТІ КӨПӨЛШЕМДІ МОДЕЛЬ 78

АҚПАРАТТЫҚ ТЕХНОЛОГИЯЛАР

Д.Е. Абжанова, С.Е. Шарипова
ЖАМБЫЛ ӨҢІРІНІҢ ТРАНСШЕКАРАЛЫҚ СУ РЕСУРСТАРЫН МОНИТОРИНГТЕУ ЖӘНЕ ДЕРЕКТЕР АЛМАСУҒА АРНАЛҒАН ЦИФРЛЫҚ ПЛАТФОРМАНЫҢ АРХИТЕКТУРАСЫ МЕН ПРАКТИКАЛЫҚ ІСКЕ АСЫРЫЛУЫ 101
А.Айтанұлы, О.А. Усатова
БҰЛТТЫ МЕДИЦИНАЛЫҚ ЖҮЙЕЛЕРДЕ ЕКІ ФАКТОРЛЫ АУТЕНТИФИКАЦИЯНЫ ЖАҚСАРТУ: ТЕРЕҢ ОҚЫТУ ӘДІСТЕРІНІҢ ҚОЛДАНУ МҮМКІНДІКТЕРІ 122
О. Альсеитов, О. Кучанский
ЭЛЕКТР СТАНЦИЯЛАРЫНЫҢ САҒАТТЫҚ ШЫҒАРЫНДЫЛАРЫН ФИЗИКАЛЫҚ ЗАҢДЫЛЫҚТАРДЫ ЕСКЕРЕТІН ГИБРИДТІ ОҚЫТУ: БЕЛГІСІЗДІК ПЕН ТАСЫМАЛДАУДЫ ЕСКЕРЕТІН КӨПЛАСТАУЫШТЫ БЕНЧМАРК 141
А.М. Ахметова, Д.О. Байжанова, Ж.Н. Алишева, Б.А. Шаймов
КРІ КӨРСЕТКІШТЕРІН ЕСКЕРЕ ОТЫРЫП, ТАБИҒИ ТІЛ МОДЕЛЬДЕРІНЕ АРНАЛҒАН ШЫҒЫС МАРШРУТТАУЫНЫҢ АДАПТИВТІ ЖҮЙЕСІ 160
Ә.Қ. Әйтім
ҚАЗАҚ МӘТІНДІК КОРПУСТАРЫН АВТОМАТТЫ ЖӘНЕ ЖАРТЫЛАЙ АВТОМАТТЫ АННОТАЦИЯЛАУ 180
Ж.Е. Балабаев, С.В. Белощицкая
ШАҒАТАЙ ҚОЛЖАЗБАЛАРЫНЫҢ МӘТІНІН ТАНУ: CRNN-ДІ НӨЛДЕН БАСТАП ОҚЫТУ МЕН ТРАНСФОРМЕР НЕГІЗІНДЕГІ ТРАНСФЕРЛІК ОҚЫТУДЫ САЛЫСТЫРУ 198
Д.Б. Баумуратова, Т.К. Жукабаева, К.А. Абдрахманов, З. Ахмад, А.А. Сеитов
ЖОҒАРЫ БІЛІМ БЕРУДЕ AI НЕГІЗДЕЛГЕН ЦИФРЛЫҚ ТРАНСФОРМАЦИЯҒА АРНАЛҒАН НЕЙРО-ЦИФРЛЫҚ ОҚЫТУ МОДЕЛІ 219
Т. Даоюн, Г. Амирханова, Фу Яньвэй, Е. Фарузқызы
ФЕЙК ВИДЕОЛАРДАҒЫ МУЛЬТИМОДАЛЬДЫ СЕМАНТИКАЛЫҚ МАНИПУЛЯЦИЯНЫҢ ПАЙДАЛАНУШЫЛАРДЫҢ ҚАБЫЛДАЙТЫН СЕНІМІНЕ ӘСЕР ЕТУ МЕХАНИЗМІ 241
Ж. Доумчариева, Д. Тусупов, А. Акынбекова, М. Хасанова, М. Кассим
ЦИФРЛЫҚ ЕГІЗДЕГІ СҮТ САПАСЫН АНСАМБЛЬДІК ЖӘНЕ ГИБРИДТІ ОҚЫТУ АРҚЫЛЫ ЖІКТЕУ 258
Қ.З. Игибаев, А.Х. Козбакова, О. Әуелбеков, Т. Мерембаев, Қ. Есентаев
БИОРЕАКТОРДА IoT ТЕХНОЛОГИЯЛАРЫН ПАЙДАЛАНА ОТЫРЫП ДИНАМИКАЛЫҚ АРАЛАСТЫРУ ҚҰРЫЛҒЫСЫН ЗЕРТТЕУ 279
А.Н. Исмуканова, Г.Кощанова, Д.В. Плескачев, Ж.Сарсенбаева, П. Шмидт
ЖАСАНДЫ ИНТЕЛЛЕКТ НЕГІЗІНДЕ ҒЫЛЫМИ ДЕРЕКТЕРДІ АВТОМАТТЫ ТҮРДЕ ЖІКТЕУ: ТРАНСФОРМЕР АРХИТЕКТУРАЛАРЫ МЕН ГИБРИДТІ ТӘСІЛДЕРДІҢ САЛЫСТЫРМАЛЫ ТАЛДАУЫ 306
Т.Ж. Камалов, Али Абд. Альмисреб, Мохаммед А. Салех
МАШИНАЛЫҚ ОҚЫТУҒА АРНАЛҒАН ГИБРИДТІ КВАНТТЫҚ-КЛАССИКАЛЫҚ АЛГОРИТМДЕРІН ЗЕРТТЕУ 323
Б. Кәрімқызы, К.Г. Карасенко, Н.А. Әлішеров, Н.М. Мұратбек, А. Кәрім
2021–2025 ЖЫЛДАР АРАЛЫҒЫНДАҒЫ ГЛИКЕМИЯЛЫҚ МОНИТОРИНГКЕ АРНАЛҒАН МЕДИЦИНАЛЫҚ ҚҰРЫЛҒЫЛАРДЫ ӨНДІРУ ЖӘНЕ ДИАГНОСТИКАЛАУДАҒЫ ЖАСАНДЫ ИНТЕЛЛЕКТ ӘДІСТЕРІ БОЙЫНША НЕГІЗГІ ЖАРИЯЛАНЫМДАР 338
С. К. Кумызбаева, К.Е. Багланбаев, К.С. Саркыт, М. Конырова
КҮЙ-ЖАЙДЫ КЛИМАТҚА СЕЗІМТАЛ ҰСЫНУМЕН ГИБРИДТІ ЭНЕРГОЖҮЙЕНІҢ ИНТЕЛЛЕКТУАЛДЫ ЭНЕРГОМЕНЕДЖМЕНТІН ҮШФАЗАЛЫ ОҚЫТУ 357
Н.Е. Қарабаев, С.К. Серикбаева, Е. Бенкхелифа, Б. Тасуов

IOT-ЖЕЛІЛЕРІНДЕГІ БАСЫП КІРУДІ АНЫҚТАУЖҮЙЕЛЕРІНДЕ МАШИНАЛЫҚ ОҚЫТУДЫҢ АНСАМБЛЬДІ КҮЛГІЛЕРІНІҢ БӘСЕКЕЛЕС ШАБУЫЛДАРҒА ТӨЗІМДІЛІГІН ЗЕРТТЕУ 369
М. Нұртас, Ғ. Капышев
МАШИНАЛЫҚ ОҚЫТУДЫ ҚОЛДАНА ОТЫРЫП, ПАЙДАЛЫ ҚАЗБАЛАРДЫҢ ПЕРСПЕКТИВАЛЫЛЫҒЫН БОЛЖАУ ҮШІН ГЕОКЕҢІСТІКТІК ЖӘНЕ ГЕОФИЗИКАЛЫҚ ДЕРЕКТЕРДІ АЛДЫН АЛА ӨҢДЕУ ЖӘНЕ БІРІКТІРУДІҢ ӘДІСТЕМЕЛІК НЕГІЗІ 392
А. Нефтиссов, А. Билощицкий, И. Казамбаев, А. Медетхан, Л. Кириченко
ГЕРМЕТИКАЛЫҚ АРХИВТІК САҚТАУ ҚАПТАМАЛАРЫНДА МИКРОКЛИМАТТЫ ЭНЕРГИЯ ТИІМДІЛІГІН ЕСКЕРЕ ОТЫРЫП ТҰРАҚТАНДЫРУ 405
Ә.Ж. Омуртаева, У.Т. Махажанова, А.К. Адилова, Ш.К. Ележанова, Я. Култан
АГРОӨНЕРКӘСІПТІК КЕШЕНДЕГІ ИНВЕСТИЦИЯЛЫҚ ТӘУЕКЕЛДЕРДІ БАҒАЛАУҒА АРНАЛҒАН БЕЙІМДЕЛГІШ НЕЙРОНДЫҚ-БҰЛДЫР МОДЕЛЬ: МАШИНАЛЫҚ ОҚЫТУДЫ ИНТЕГРАЦИЯЛАУ АРҚЫЛЫ ШЕШІМ ҚАБЫЛДАУДЫ ҚОЛДАУ СӘУЛЕТІН ЖАҢҒЫРТУ 421
Н. Шаждекеева, А. Мұханбетқалиева, А. Әбдіғалиева, А. Жанекешова, Л. Нұрғалиева
ФИЛЬМДЕРДІ ҰСЫНУ ЖҮЙЕСІН ҚҰРУ КЕЗІНДЕГІ КОЛЛАБОРАТИВТІК СҮЗГІЛЕУДІҢ МАТЕМАТИКАЛЫҚ МОДЕЛЬДЕРІН САЛЫСТЫРМАЛЫ ТАЛДАУ 445
С. Шармұханбет, Г. Тұрмұханова, О. Финдик, В. Махатова, М. Рахметов
ҚҰБЫРІШІЛІК РОБОТТЫ ПАЙДАЛАНУ АРҚЫЛЫ ҚҰБЫР АҚАУЛАРЫН АНЫҚТАУДЫҢ ЗИЯТКЕРЛІК ЖҮЙЕСІ 461

АҚПАРАТТЫҚ ҚАУІПСІЗДІК ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАРҒА АРНАЛҒАН

С.З. Дәруіш, А.У. Утегенова, А.У. Калижанова, Ж.Ж. Жумабекова
ЗЫМЫРАН ТАСЫМАЛДАУШЫЛАРДЫҢ ҚҰЛАУ АЙМАҚТАРЫНЫҢ ЛАСТАНУЫН МАШИНАЛЫҚ ОҚЫТУМЕН ЖІКТЕУ 480
С.Н. Исабаева, А.К. Рысбаева, Ж.Б. Ибраева, Н.Р. Муссабеков, Н.Ш. Акимжан
МАШИНАЛЫҚ ОҚЫТУ НЕГІЗІНДЕ ОТ-ЖЕЛІЛЕРІНДЕГІ КИБЕРШАБУЫЛДАРДЫ АНЫҚТАУ: ЖЕЛІЛІК АНОМАЛИЯЛАРДЫ ЖІКТЕУ ӘДІСТЕРІНІҢ САЛЫСТЫРМАЛЫ ТАЛДАУЫ 495
Г.А. Шангытбаева, А.Н. Аманов, М.К. Исенова, Г.Е. Берикханова, Қ.О. Шакерхан
МЕМЛЕКЕТТІК ОРГАНДАРДЫҢ БҰЛТТЫ СЕРВИСТЕРІНДЕГІ АҚПАРАТТЫҚ ҚАУІПСІЗДІК: ТӘУЕКЕЛДЕРДІ БАҒАЛАУ ЖӘНЕ ҚОРҒАУ СТРАТЕГИЯЛАРЫ 512
Е.Н. Турсынбек, Н. Албанбай, Д. Дженоури
IOT ЖЕЛІЛЕРІНДЕ КИБЕРШАБУЫЛДАРДЫ АНЫҚТАУ ҮШІН ЕКІ ДЕҢГЕЙЛІ ФЕДЕРАЦИЯЛАНҒАН ОҚЫТУДАҒЫ БАЙЛАНЫС ШЫҒЫНДАРЫН ОҢТАЙЛАНДЫРУ 529
А. Тастемиров, С.Т. Аманжолова, Б. Алдош, С.Б. Муханов
ACTIVE DIRECTORY ЖҮЙЕСІНДЕ ҚАУІПТЕРДІ АНЫҚТАУ ҮШІН МАШИНАЛЫҚ ОҚЫТУ МОДЕЛЬДЕРІН САЛЫСТЫРМАЛЫ БАҒАЛАУ: WINDOWS ҚАУІПСІЗДІК ОҚИҒАЛАРЫ ЖУРНАЛДАРЫНДА БЕЛГІ ЖЫЛЫСТАУЫН БАҚЫЛАУ ЖӘНЕ КРОСС-СЦЕНАРЛЫҚ ВАЛИДАЦИЯ 545

СОДЕРЖАНИЕ
ЦИФРОВЫЕ ТЕХНОЛОГИИ В РАЗВИТИИСОЦИО-ЭКОНОМИЧЕСКИХ СИСТЕМ

А Адалбек, С.Е. Керимкулов, К.Д. Кулиев, О. Мекеш, Ж.Ж. Хамитова
РАЗРАБОТКА РЕЗУЛЬТАТИВНО-ОРИЕНТИРОВАННЫХ МОДЕЛЕЙ СГЛАЖИВАНИЯ ДАННЫХ И АНАЛИЗ ИХ ЭФФЕКТИВНОСТИ 11
А.Т. Искакова, Г.Е. Курманкулова, К. Громашек, К.О. Тогжанова, Р.С. Алгожаева
ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО И ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗНОГО ПЛАНИРОВАНИЯ И РЕСУРСОСБЕРЕЖЕНИЯ В ПРОИЗВОДСТВЕННЫХ СИСТЕМАХ 30
А.Б. Касымова, Р.К. Ускенбаева, Ж.А.Сыдыков, Д. Омирзак, С. Калибеков
АРХИТЕКТУРА ИНТЕЛЛЕКТУАЛЬНОГО ЧАТ-БОТА ДЛЯ SMART CITY: ПРИЁМ, АНАЛИЗ И МАРШРУТИЗАЦИЯ ОБРАЩЕНИЙ ГРАЖДАН НА ОСНОВЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА 46
О.М. Олжаев, Б.С. Омаров, A.Suliman
YOLO-CBAM: ЛЕГКОВЕСНАЯ НЕЙРОСЕТЕВАЯ МОДЕЛЬ С ИНТЕГРИРОВАННЫМ МЕХАНИЗМОМ ВНИМАНИЯ ДЛЯ АНАЛИЗА ДОРОЖНЫХ ПОВРЕЖДЕНИЙ 66
К. Ермагамбетова, А. Сарсекеева, Е. Ахласов
ПРЕОДОЛЕНИЕ ФРАГМЕНТАЦИИ В КЛАССИФИКАЦИИ ЦИФРОВОГО КУЛЬТУРНОГО НАСЛЕДИЯ:

ИНТЕГРАТИВНАЯ МНОГОМЕРНАЯ МОДЕЛЬ 78

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

Д.Е. Абжанова, С.Е. Шарипова
АРХИТЕКТУРА И ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ ЦИФРОВОЙ ПЛАТФОРМЫ МОНИТОРИНГА И ОБМЕНА ДАННЫМИ О ТРАНСГРАНИЧНЫХ ВОДНЫХ РЕСУРСАХ ЖАМБЫЛСКОГО РЕГИОНА 101
А. Айтанулы, О.А. Усатова
СОВЕРШЕНСТВОВАНИЕ ДВУХФАКТОРНОЙ АУТЕНТИФИКАЦИИ ДЛЯ ОБЕСПЕЧЕНИЯ БЕЗОПАСНОСТИ ОБЛАЧНЫХ МЕДИЦИНСКИХ СИСТЕМ: ВОЗМОЖНОСТИ ПРИМЕНЕНИЯ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ 122
О. Альсеитов, О. Кучанский
ФИЗИЧЕСКИ ИНФОРМИРОВАННОЕ ГИБРИДНОЕ ОБУЧЕНИЕ ДЛЯ ПОЧАСОВОГО ПРОГНОЗИРОВАНИЯ ВЫБРОСОВ ЭЛЕКТРОСТАНЦИЙ: МНОГОКОМПОНЕНТНЫЙ БЕНЧМАРК С ОЦЕНКОЙ НЕОПРЕДЕЛЁННОСТИ И ПЕРЕНОСИМОСТЬЮ МОДЕЛЕЙ 141
А.М. Ахметова, Д.О. Байжанова, Ж.Н. Алишева, Б.А. Шаймов
АДАПТИВНЫЙ МАРШРУТИЗАТОР ВЫВОДА С УЧЕТОМ КЛЮЧЕВЫХ ПОКАЗАТЕЛЕЙ ЭФФЕКТИВНОСТИ ДЛЯ МОДЕЛЕЙ ЕСТЕСТВЕННОГО ЯЗЫКА 160
Ә.Қ. Әйтім
АВТОМАТИЧЕСКАЯ И ПОЛУАВТОМАТИЧЕСКАЯ АННОТАЦИЯ КАЗАХСКИХ ТЕКСТОВЫХ КОРПУСОВ 180
Ж.Е. Балабаев, С.В. Белощицкая
РАСПОЗНАВАНИЕ РУКОПИСНОГО ТЕКСТА ШАГАТАЙСКИХ РУКОПИСЕЙ: СРАВНЕНИЕ ОБУЧЕНИЯ CRNN С НУЛЯ И ТРАНСФЕРНОГО ОБУЧЕНИЯ НА БАЗЕ ТРАНСФОРМЕРОВ 198
Д.Б. Баумуратова, Т.К. Жукабаева, К.А. Абдрахманов, З. Ахмад, А.А. Сеитов
НЕЙРО-ЦИФРОВАЯ МОДЕЛЬ ОБУЧЕНИЯ ДЛЯ ЦИФРОВОЙ ТРАНСФОРМАЦИИ ВЫСШЕГО ОБРАЗОВАНИЯ С ПОМОЩЬЮ AI 219
Т. Даоюн, Г. Амирханова, Фу Яньвэй, Е. Фарузкызы
МЕХАНИЗМ ВЛИЯНИЯ МУЛЬТИМОДАЛЬНОЙ СЕМАНТИЧЕСКОЙ МАНИПУЛЯЦИИ В ФЕЙКОВЫХ ВИДЕО НА ВОСПРИНИМАЕМОЕ ДОВЕРИЕ ПОЛЬЗОВАТЕЛЕЙ 241
Ж. Доумчариева, Д. Тусупов, А. Акынбекова, М. Хасанова, М. Кассим
КЛАССИФИКАЦИЯ КАЧЕСТВА МОЛОКА В ЦИФРОВОМ ДВОЙНИКЕ АНСАМБЛЕВЫМИ И ГИБРИДНЫМИ МЕТОДАМИ 258
К.З. Игибаев, А.Х. Козбакова, О. Ауелбеков, Т. Мерембаев, К. Есентаев
ИССЛЕДОВАНИЕ УСТРОЙСТВА ДИНАМИЧЕСКОГО ПЕРЕМЕШИВАНИЯ В БИОРЕАКТОРЕ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИЙ IoT 279
А. Н. Исмуканова, Г.Кощанова, Д.В. Плескачев, Ж.Сарсенбаева, П. Шмидт
АВТОМАТИЗИРОВАННАЯ КЛАССИФИКАЦИЯ НАУЧНЫХ ДАННЫХ НА ОСНОВЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: СРАВНИТЕЛЬНЫЙ АНАЛИЗ ТРАНСФОРМЕРНЫХ АРХИТЕКТУР И ГИБРИДНЫХ МЕТОДОВ 306
Т.Ж. Камалов, Али Абд. Альмисреб, Мохаммед А. Салех
ИССЛЕДОВАНИЕ ГИБРИДНЫХ КВАНТОВО-КЛАССИЧЕСКИХ АЛГОРИТМОВ ДЛЯ МАШИННОГО ОБУЧЕНИЯ 323
Б. Каримкызы, К.Г. Карасенко, Н.А. Алишеров, Н.М. Муратбек, А. Карим
КЛЮЧЕВЫЕ ПУБЛИКАЦИИ 2021–2025 ГОДОВ ПО МЕТОДАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ПРОИЗВОДСТВЕ И ДИАГНОСТИКЕ МЕДИЦИНСКИХ ИЗДЕЛИЙ ДЛЯ ГЛИКЕМИЧЕСКОГО МОНИТОРИНГА 338
С.К. Кумызбаева, К.Е. Багланбаев, К.С. Сарқыт, М.Ж. Конырова
ТРЁХФАЗОВОЕ ОБУЧЕНИЕ ИНТЕЛЛЕКТУАЛЬНОГО ЭНЕРГОМЕНЕДЖМЕНТА ГИБРИДНОЙ ЭНЕРГОСИСТЕМЫ С КЛИМАТО-ОСВЕДОМЛЁННЫМ ПРЕДСТАВЛЕНИЕМ СОСТОЯНИЯ 357
Н.Е. Қарабаев, С.К. Серикбаева, Е. Бенкхелифа, Б. Тасуов
ИССЛЕДОВАНИЕ УСТОЙЧИВОСТИ АНСАМБЛЕВЫХ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ К СОСТЯЗАТЕЛЬНЫМ АТАКАМ В СИСТЕМАХ ОБНАРУЖЕНИЯ ВТОРЖЕНИЙ IOT-СЕТЕЙ 369
М. Нуртас, Г. Капышев
МЕТОДОЛОГИЧЕСКАЯ ОСНОВА ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ И ИНТЕГРАЦИИ ГЕОПРОСТРАНСТВЕННЫХ И ГЕОФИЗИЧЕСКИХ ДАННЫХ ДЛЯ ПРОГНОЗИРОВАНИЯ ПЕРСПЕКТИВНО-

СТИ МИНЕРАЛЬНЫХ РЕСУРСОВ С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ 392
А. Нефтиссов, А. Билощицкий, И. Казамбаев, А. Медетхан, Л. Кириченко
ЭНЕРГОЭФФЕКТИВНАЯ СТАБИЛИЗАЦИЯ МИКРОКЛИМАТА В ГЕРМЕТИЧНЫХ АРХИВНЫХ ХРАНИЛИЩАХ 405
А.Ж. Омуртаева, У.Т. Махажанова, А.К. Адилова, Ш.К. Ележанова, Я.Култан
АДАПТИВНАЯ НЕЙРО-НЕЧЕТКАЯ МОДЕЛЬ ДЛЯ ОЦЕНКИ ИНВЕСТИЦИОННЫХ РИСКОВ В АГРОПРОМЫШЛЕННОМ КОМПЛЕКСЕ: МОДЕРНИЗАЦИЯ АРХИТЕКТУРЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ ПОСРЕДСТВОМ ИНТЕГРАЦИИ МАШИННОГО ОБУЧЕНИЯ 421
Н. Шаждекеева, А. Муханбеткалиева, А. Абдигалиева, А. Жанекешова, Л. Нургалиева
СРАВНИТЕЛЬНЫЙ АНАЛИЗ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ КОЛЛАБОРАТИВНОЙ ФИЛЬТРАЦИИ ПРИ ПОСТРОЕНИИ РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫ ФИЛЬМОВ 445
С. Шармуханбет, Г. Турмуханова, О.Финдик, В.Махатова, М. Рахметов
ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА ОБНАРУЖЕНИЯ ДЕФЕКТОВ ТРУБОПРОВОДОВ С ИСПОЛЬЗОВАНИЕМ ВНУТРИТРУБНОГО РОБОТА 461

ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ И КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ

С.З. Дәруіш, А.У. Утегенова, А.У. Калижанова, Ж.Ж. Жумабекова
МАШИННОЕ ОБУЧЕНИЕ ДЛЯ КЛАССИФИКАЦИИ ЗАГРЯЗНЕНИЯ РАЙОНОВ ПАДЕНИЯ РАКЕТ-НОСИТЕЛЕЙ 480
С.Н. Исабаева, А.К. Рысбаева, Ж.Б. Ибраева, Н.Р. Муссабеков, Н.Ш. Акимжан
ОБНАРУЖЕНИЕ КИБЕРАТАК В ОТ-СЕТЯХ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ: СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ КЛАССИФИКАЦИИ СЕТЕВЫХ АНОМАЛИЙ 495
Г.А. Шангытбаева, А.Н. Аманов, М.К. Исенова, Г.Е. Берикханова, Қ.О. Шакерхан
ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ В ОБЛАЧНЫХ СЕРВИСАХ ГОСУДАРСТВЕННЫХ ОРГАНОВ: ОЦЕНКА РИСКОВ И СТРАТЕГИИ ЗАЩИТЫ 512
Е.Н. Турсынбек, Н. Албанбай, Д. Дженоури
ОПТИМИЗАЦИЯ КОММУНИКАЦИОННЫХ ЗАТРАТ В ДВУХУРОВНЕВОМ ФЕДЕРАТИВНОМ ОБУЧЕНИИ ДЛЯ ОБНАРУЖЕНИЯ КИБЕРАТАК В IOT-СЕТЯХ 529
А. Тастемиров, С.Т. Аманжолова, Б. Алдош, С.Б. Муханов
СРАВНИТЕЛЬНАЯ ОЦЕНКА МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБНАРУЖЕНИЯ УГРОЗ В ACTIVE DIRECTORY: КОНТРОЛЬ УТЕЧКИ ПРИЗНАКОВ И КРОСС-СЦЕНАРНАЯ ВАЛИДАЦИЯ НА ЖУРНАЛАХ БЕЗОПАСНОСТИ WINDOWS 545

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.3. Number 27 (2026). Pp. 141–159
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.27.3.008

PHYSICS-INFORMED HYBRID LEARNING FOR HOURLY POWER-PLANT EMISSIONS: A MULTI-POLLUTANT BENCHMARK WITH UNCERTAINTY AND TRANSFER

O. Alseitov, O. Kuchanskyi*
Astana IT University, Astana, Kazakhstan.
E-mail: a.kuchanskyi@astanait.edu.kz

Olzhas Alseitov — PhD student of the School of Artificial Intelligence and Data Science, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0001-9161-7206;
Oleksandr Kuchanskyi — D.Sc., Professor of the School of Artificial Intelligence and Data Science, Astana IT University, Astana, Kazakhstan
E-mail: a.kuchanskyi@astanait.edu.kz, https://orcid.org/0000-0003-1277-8031.

© O. Alseitov, O. Kuchanskyi

Abstract. This study examines whether the same modelling strategy can be used for hourly CO_2, NOx, and SO_2 emission estimation at power plants, or whether the choice of model should depend on the pollutant and deployment setting. We assembled 10,144,680 unit-hours from 583 generating units in Georgia, Kentucky, and Texas for 2022 and 2024 by combining U.S. EPA Clean Air Markets Program Data with EIA-860 plant metadata through the EPA/EIA Power Sector Crosswalk. Emission-factor, statistical, machine-learning, and physics-informed hybrid models were tested using blocked-time, plant-holdout, and seasonal-holdout splits to limit temporal and unit-level leakage. PHASE, the proposed hybrid, uses a scaled emission-factor estimate as a physical prior and learns the remaining residual with a data-driven model while also estimating predictive uncertainty. The results differ sharply by pollutant. For CO_2, the emission-factor baseline already explained most of the variance ($R^2 = 0.992$), whereas its performance was much weaker for NOx ($R^2 = 0.394$) and SO_2 ($R^2 = 0.413$). XGBoost reached $R^2 = 0.999$, 0.973, and 0.980 for CO_2, NOx, and SO_2, respectively. For NOx, removing recent historical features caused the largest deterioration in accuracy, while removing the simple physical prior slightly improved average in-domain MAE. At the same time, PHASE-lite reduced some large-error episodes according to the Diebold-Mariano comparison ($p < 10^{-3}$), achieved 89.2% empirical coverage for a nominal 90% interval, and retained $R^2 = 0.965$ on previously unseen plants. The evidence therefore supports a differentiated strategy: a transparent factor model is already strong for CO_2,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

while NOx and SO_2 require learned representations of changing operating conditions. The hybrid is most useful when uncertainty, large-error robustness, and cross-plant transfer matter more than the lowest average in-domain error. A future corroborative comparison with Sentinel-5P TROPOMI NO_2 observations over Kazakhstan thermal plants is also defined.

Keywords: emission monitoring, predictive emission monitoring system, physics-informed machine learning, uncertainty quantification, plant-level transfer, continuous emission monitoring systems, carbon neutrality

For citation: Alseitov O., Kuchanskyi O. (2026). Physics-informed hybrid learning for hourly power-plant emissions: a multi-pollutant benchmark with uncertainty and transfer // International journal of information and communication technologies. Vol. 7. No. 27. Pp. 141–159. https://doi.org/10.54309/IJICT.2026.27.3.008. (In Eng.).

Conflict of interest: The authors declare that there is no conflict of interest.

ЭЛЕКТР СТАНЦИЯЛАРЫНЫҢ САҒАТТЫҚ ШЫҒАРЫНДЫЛАРЫН ФИЗИКАЛЫҚ ЗАҢДЫЛЫҚТАРДЫ ЕСКЕРЕТІН ГИБРИДТІ ОҚЫТУ: БЕЛГІСІЗДІК ПЕН ТАСЫМАЛДАУДЫ ЕСКЕРЕТІН КӨП ЛАСТАУЫШТЫ БЕНЧМАРК

О. Альсеитов, О. Кучанский*
Astana IT University, Астана, Қазақстан.
E-mail: a.kuchanskyi@astanait.edu.kz

Альсеитов Олжас — Жасанды интеллект және деректер туралы ғылым мектебінің PhD докторанты, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0001-9161-7206;
Кучанский Олександр — техника ғылымдарының докторы, Жасанды интеллект және деректер туралы ғылым мектебінің профессоры, Astana IT University, Астана, Қазақстан
E-mail: a.kuchanskyi@astanait.edu.kz, https://orcid.org/0000-0003-1277-8031.

© О. Альсеитов, О. Кучанский

Аннотация. Бұл зерттеуде электр станцияларындағы CO_2, NOx және SO_2 шығарындыларын сағат сайын бағалау үшін бір модельдеу тәсілі жеткілікті ме, әлде модельді ластаушы зат пен қолдану жағдайына қарай таңдау қажет пе деген сұрақ қарастырылады. АҚШ Қоршаған ортаны қорғау агенттігінің Clean Air Markets Program Data деректері EIA-860 метадеректерімен EPA/EIA Power Sector Crosswalk арқылы біріктіріліп, 2022 және 2024 жылдардағы Джорджия, Кентукки және Техас штаттарының 583 генераторлық қондырғысынан 10 144 680 қондырғы-сағаттық бенчмарк жасалды. Эмиссия коэффициентіне негізделген, статистикалық, машиналық оқыту және физикалық заңдылықтарды ескеретін гибридті модельдер уақыттық және қондырғы деңгейіндегі деректердің ағып ке-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

туін азайту үшін blocked-time, plant-holdout және seasonal-holdout схемаларымен тексерілді. Ұсынылған PHASE тәсілі масштабталған эмиссия коэффициентін физикалық априор ретінде пайдаланып, қалған ауытқуды деректерге негізделген модель арқылы үйренеді және болжам белгісіздігін бағалайды. Нәтижелер ластаушы затқа қарай айқын ерекшеленді. CO_2 үшін эмиссия коэффициенті өзгерістің басым бөлігін түсіндірді ($R^2 = 0,992$), ал NOx ($R^2 = 0,394$) және SO_2 ($R^2 = 0,413$) үшін оның дәлдігі едәуір төмен болды. XGBoost CO_2, NOx және SO_2 бойынша тиісінше $R^2 = 0,999$, 0,973 және 0,980 көрсетті. NOx үшін соңғы уақыттық тарих белгілерін алып тастау дәлдікке ең үлкен зиян келтірді, ал қарапайым физикалық априорды алып тастау домен ішіндегі орташа MAE-ні аздап жақсартты. Сонымен бірге PHASE-lite Diebold-Mariano салыстыруы бойынша ірі қателердің бір бөлігін азайтты ($p < 10^{-3}$), номиналдық 90% аралық үшін 89,2% эмпирикалық қамту берді және бұрын көрмеген станцияларда $R^2 = 0,965$ сақтады. Сондықтан CO_2 үшін түсіндірмелі факторлық модель жеткілікті күшті, ал NOx пен SO_2 үшін өзгермелі пайдалану жағдайларын үйренетін модельдер қажет. Гибрид орташа доменішілік қатені барынша азайтудан гөрі белгісіздік, ірі қателерге орнықтылық және станциялар арасында тасымалдау маңызды болған жағдайда пайдалырақ. Қазақстанның жылу электр станциялары үшін Sentinel-5P TROPOMI NO_2 бақылауларымен болашақ салыстырмалы тексеру хаттамасы да белгіленді.

Түйінді сөздер: шығарындылар мониторингі, шығарындыларды болжамдық мониторингтеу жүйесі, физикалық заңдылықтарды ескеретін машиналық оқыту, белгісіздікті сандық бағалау, электр станциялары деңгейінде тасымалдау, шығарындыларды үздіксіз мониторингтеу жүйелері, көміртегі бейтараптығы

Дәйексөздер үшін: Альсеитов О., Кучанский О. (2026). Электр станцияларының сағаттық шығарындыларын физикалық заңдылықтарды ескеретін гибридті оқыту: белгісіздік пен тасымалдауды ескеретін көпластауышты бенчмарк // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No 27. Б. 141–159. https://doi.org/10.54309/IJICT.2026.27.3.008. (Ағыл. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

ФИЗИЧЕСКИ ИНФОРМИРОВАННОЕ ГИБРИДНОЕ ОБУЧЕНИЕ ДЛЯ ПОЧАСОВОГО ПРОГНОЗИРОВАНИЯ ВЫБРОСОВ ЭЛЕКТРОСТАНЦИЙ: МНОГОКОМПОНЕНТНЫЙ БЕНЧМАРК С ОЦЕНКОЙ НЕОПРЕДЕЛЁННОСТИ И ПЕРЕНОСИМОСТЬЮ МОДЕЛЕЙ

О. Альсеитов, О. Кучанский*
Astana IT University, Астана, Казахстан.
E-mail: a.kuchanskyi@astanait.edu.kz

Альсеитов Олжас — PhD-докторант Школы искусственного интеллекта и науки о данных Astana IT University, Астана, Казахстан
https://orcid.org/0000-0001-9161-7206;

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Кучанский Олександр — доктор технических наук, профессор Школы искусственного интеллекта и науки о данных Astana IT University, Астана, Казахстан
E-mail: a.kuchanskyi@astanait.edu.kz, https://orcid.org/0000-0003-1277-8031.

© О. Альсеитов, О. Кучанский

Аннотация. В работе проверяется, можно ли использовать один и тот же подход для почасовой оценки выбросов CO_2, NOx и SO_2 на электростанциях или выбор модели должен зависеть от загрязняющего вещества и сценария применения. Бенчмарк объёмом 10 144 680 агрегато-часов для 583 генерирующих установок в Джорджии, Кентукки и Техасе за 2022 и 2024 годы сформирован путём объединения U.S. EPA Clean Air Markets Program Data с метаданными EIA-860 через EPA/EIA Power Sector Crosswalk. Модели на основе коэффициентов выбросов, статистические методы, машинное обучение и физически информированные гибриды оценивались на разбиениях blocked-time, plant-holdout и seasonal-holdout, ограничивающих временную утечку данных и пересечение установок между обучением и тестированием. Предложенный PHASE использует масштабированную оценку по коэффициенту выбросов как физический априор, а оставшееся отклонение обучает по данным и дополняет оценкой неопределённости. Результаты существенно различаются по загрязняющим веществам. Для CO_2 факторная модель уже объясняет основную часть вариации (R^2 = 0,992), тогда как для NOx (R^2 = 0,394) и SO_2 (R^2 = 0,413) её точность значительно ниже. XGBoost достиг R^2 = 0,999, 0,973 и 0,980 для CO_2, NOx и SO_2 соответственно. Для NOx наибольшее ухудшение возникало при исключении недавней временной истории, а удаление простого физического априора немного снижало среднюю внутридоменную MAE. При этом PHASE-lite уменьшал часть крупных ошибок по тесту Диболда-Мариано ($p < 10^{-3}$), обеспечивал 89,2% эмпирического покрытия для номинального 90%-го интервала и сохранял R^2 = 0,965 на ранее не встречавшихся установках. Полученные результаты указывают на необходимость дифференцированного подхода: для CO_2 прозрачная факторная модель уже является сильным решением, тогда как NOx и SO_2 требуют обучаемого представления изменяющихся режимов работы. Гибрид наиболее полезен там, где важны неопределённость, устойчивость к крупным ошибкам и перенос между электростанциями, а не минимальная средняя внутридоменная ошибка. Также определён протокол будущего сопоставления с наблюдениями Sentinel-5P TROPOMI NO_2 над тепловыми электростанциями Казахстана.

Ключевые слова: мониторинг выбросов, система прогнозного мониторинга выбросов, физически информированное машинное обучение, количественная оценка неопределённости, перенос на уровне электростанций, системы непрерывного мониторинга выбросов, углеродная нейтральность

Для цитирования: Альсеитов О., Кучанский О. (2026). Физически информированное гибридное обучение для почасового прогнозирования выбросов

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

электростанций: многокомпонентный бенчмарк с оценкой неопределённости и переносимостью моделей // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 27. Стр. 141–159. https://doi.org/10.54309/IJICT.2026.27.3.008. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

Hourly emission data are needed for regulatory control, emissions accounting, and operational assessment, but obtaining a complete continuous record is not always straightforward. Continuous emission monitoring systems (CEMS) provide the reference measurements by combining stack-gas concentration and flow information, yet their coverage differs between power systems and even installed instruments may have gaps caused by calibration, maintenance, or faults. Predictive emission monitoring systems (PEMS) address a different part of the problem: instead of measuring the stack directly, they estimate the hourly emission rate from operating variables that are routinely recorded by the plant, including load, heat input, steam load, operating regime, and unit characteristics. This makes PEMS relevant both as a complementary soft sensor and as a way to reconstruct periods for which direct observations are unavailable.

The difficulty of that estimation is not the same for every pollutant. Emission-factor methods link pollutant mass to fuel throughput and are especially natural for CO_2 because the amount emitted is closely tied to the carbon content and quantity of fuel consumed. The same simplification is much less convincing for NOx and SO_2. Hourly NOx formation changes with flame temperature, excess air, residence time, fuel properties, load transitions, and post-combustion controls; SO_2 also depends on fuel sulfur and the operation of desulfurization equipment. A model that performs well for CO_2 can therefore fail for NOx or SO_2 even when all three targets are observed at the same plant.

Earlier PEMS studies have used artificial neural networks, support-vector methods, ensembles, recurrent models, and physics-informed approaches to estimate boiler emissions from plant operating data (Zhou et al., 2004; Lv et al., 2013; Tan et al., 2019; Adams et al., 2020; Zhu et al., 2024). For structured industrial variables, random forests and gradient boosting are attractive because they capture nonlinear interactions without imposing a fixed functional form (Breiman, 2001; Chen et al., 2016). Sequence models address a different requirement: long short-term memory networks and temporal convolutional networks can represent delayed responses and temporal dependence (Hochreiter et al., 1997; Bai et al., 2018), while attention-based models can combine static attributes, time-varying covariates, and historical observations within one forecasting architecture (Lim et al., 2021). These methods provide strong predictive tools, but they do not by themselves answer when a physical baseline still adds value.

Physics-informed machine learning introduces physical knowledge through

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

model structure, residual formulations, physically motivated features, or modified objectives (Karniadakis et al., 2021). In the present setting, a physical estimate can serve as an interpretable first approximation and the learned component can model the part that changes with combustion state, control operation, and plant-specific behaviour. This division is useful only if the prior captures a meaningful part of the target process. When the prior is too coarse, forcing the learner to correct it may add no benefit and can even increase error. For that reason, we treat the physical component as a hypothesis to be tested rather than as an assumption that a hybrid must outperform a black-box model.

Point accuracy is only one requirement for an emission soft sensor. An operational user also needs to know when a prediction is uncertain. Existing approaches include deep ensembles (Lakshminarayanan et al., 2017), quantile regression (Koenker et al., 1978), and conformal methods (Vovk et al., 2005). Their common purpose is to attach an uncertainty range to the point prediction. In emission monitoring, such intervals are useful because they distinguish a confident estimate from a period in which the model should be treated cautiously, particularly during unusual operating conditions or when a prediction is being used to fill a measurement gap.

The published evidence nevertheless leaves several practical questions open. Many studies report results for a single unit, a short operating period, or one pollutant, making it difficult to separate model capability from the properties of a particular dataset. Random record-level train/test splits are another concern because neighbouring hourly measurements from the same unit are strongly related; putting such observations on both sides of the split can make test performance appear better than it will be in deployment. Statistical comparisons between competing model families are also uncommon, and only a limited number of studies test models on generating units that were completely absent from training. That last setting is particularly important when labelled data are scarce at the target plant.

A second unresolved issue is the role of the physical prior once a flexible learner has access to a large labelled dataset. The prior may improve interpretability, regularize predictions, or help under distribution shift, but it may also become redundant when the data already describe the operating space well. The relevant comparison is therefore broader than average test error alone. It should include pollutant type, operating regime, transfer to unseen equipment, extreme-error behaviour, missing-input robustness, uncertainty calibration, and formal significance testing. This is especially important when two models have similar average scores but behave differently during operationally important episodes (Diebold et al., 1995).

Against this background, the study asks a specific question: under leakage-controlled temporal and cross-plant evaluation, when should hourly power-plant emissions be estimated with a physical factor, a purely data-driven model, or a physics-informed hybrid? We address this question through the following objectives:

1. Build a large-scale benchmark for hourly CO_2, NOx, and SO_2 estimation by linking public CEMS operating records with plant-level technical metadata while preventing temporal and unit-level leakage.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

2. Compare emission-factor, statistical, tree-based, sequence-learning, and physics-informed hybrid models under blocked-time, plant-holdout, and seasonal-holdout evaluation.
3. Measure how the ranking of modelling approaches changes between fuel-bound CO_2 and the more combustion- and control-dependent NOx and SO_2 targets.
4. Use controlled ablations to isolate the contribution of the physical prior, current operating variables, historical features, static plant attributes, and operating-regime indicators.
5. Evaluate uncertainty calibration, large-error behaviour, and transfer to generating units that have limited or no labelled local data.
6. Determine where the proposed hybrid adds practical value relative to a strong black-box model and state the conditions under which it should not be preferred.

The resulting benchmark contains 583 generating units and about 10.1 million valid unit-hours from public U.S. sources. On this benchmark we evaluate PHASE, a residual physics-informed framework in which a scaled emission-factor estimate provides the initial physical component and a learned model predicts the remaining residual. PHASE-lite implements the residual learner with gradient-boosted trees and adds quantile-based uncertainty estimation. All comparisons use blocked-time, plant-holdout, or seasonal-holdout splits rather than random record-level splitting, so the reported scores reflect later time periods, unseen equipment, or held-out seasons rather than neighbouring records from the same episode.

The study contributes five elements. First, it provides a large multi-pollutant benchmark with leakage-controlled splits. Second, it evaluates the PHASE residual formulation in gradient-boosted and neural forms with uncertainty estimation. Third, it compares physical, statistical, machine-learning, and hybrid models separately for CO_2, NOx, and SO_2 rather than assuming one ranking for all pollutants. Fourth, the evaluation includes feature ablations, missing-channel robustness, uncertainty calibration, statistical tests, and transfer to unseen plants. Fifth, it specifies a future Sentinel-5P TROPOMI NO_2 corroboration protocol for Kazakhstan thermal plants while explicitly treating satellite retrievals as contextual evidence rather than supervised stack-level labels.

Data and Problem Formulation

Data sources and integration

We obtained hourly unit-level emissions and operating records from the U.S. Environmental Protection Agency Clean Air Markets Program Data (CAMPD), which contains emissions, operating, compliance, and facility information reported under federal power-sector monitoring programmes (U.S. EPA, 2026a). The study uses the bulk hourly files for generating units in Georgia, Kentucky, and Texas for the calendar years 2022 and 2024.

For each unit-hour, CAMPD provides operating time, gross electrical load, steam load, heat input, and reported pollutant mass. We retained gross load in megawatts

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

and heat input in million British thermal units (mmBtu), while the pollutant quantities were converted to common metric rates: CO_2 from short tons to metric tonnes per hour and NOx and SO_2 from pounds to kilograms per hour. When a record represented less than a full operating hour, we divided the reported pollutant mass by the recorded operating-time fraction so that the target remained an hourly rate.

Plant and generator characteristics came from the 2024 release of Form EIA-860, Annual Electric Generator Report (U.S. EIA, 2025). The variables used in the experiments include nameplate capacity, primary energy source, commissioning year, plant and generator identifiers, and geographic coordinates. We also used the environmental association tables to derive indicators for NOx-control technologies, including selective catalytic and selective non-catalytic reduction, and for SO_2 control through flue-gas desulfurization. When a control attribute was available only at plant level, it was kept as a plant-level indicator rather than treated as a measured unit-level operating state.

Because CAMPD combustion sources and EIA-860 electricity generators use different identifier schemes, the two sources cannot be joined reliably on a single raw identifier. We therefore used the EPA-EIA Power Sector Data Crosswalk to map EPA source identifiers to EIA plant and generator identifiers (U.S. EPA, 2026b). Records were removed when a reliable crosswalk match was unavailable or when variables required by the corresponding experiment were missing. After this integration and the quality-control steps described below, the benchmark contained 583 generating units and 10,144,680 unit-hours.

Data quality control

Quality control was applied before feature construction and before any train/validation/test split. CAMPD hours marked as calibration periods or substituted-data periods were excluded because their reported values do not represent fully observed continuous-monitoring measurements. We also removed invalid timestamps, duplicate unit-hour records, missing targets, negative pollutant masses, and operating values that violated the predefined physical plausibility checks. Performing these filters before modelling prevents the learner from being rewarded for reproducing substituted values or invalid operating states.

Offline observations were handled separately from operating-hour prediction. Hours with zero operating time or confirmed offline status were excluded from supervised training and scoring, so ‘offline’ was not used as an operating-regime class. The remaining valid hours were categorized, where applicable, as start-up, shutdown, ramping, low-load, or baseload. This choice avoids an artificial accuracy gain from large numbers of trivial zero-emission records when the intended task is to estimate emissions during operation.

Some process variables that would be useful for combustion diagnostics are not consistently present in the CAMPD bulk hourly files. In particular, stack oxygen concentration and flue-gas temperature were not available with the coverage required for the benchmark and were therefore excluded. The final feature space relies on routinely available channels such as gross load, heat input, steam load, their temporal

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

changes and historical summaries, together with static plant and generator metadata. Missing inputs were never imputed using information from the validation or test partitions.

Prediction problem

We formulate the task as hourly nowcasting, or soft-sensor estimation. For a generating unit at hour t, the model predicts the CO_2, NOx, or SO_2 mass rate using variables observed during that hour, historical information available no later than t, and static unit attributes. Current inputs include gross load, heat input, steam load, ramp rate, operating-regime indicators, and calendar variables. Historical inputs include lagged values and rolling summaries, while the static block contains capacity, primary fuel, commissioning year, location, and available NOx- and SO_2-control indicators. Every lagged or rolling feature is constructed within a unit and uses only present or past information; no future observation enters feature generation.

The prediction targets are hourly pollutant mass rates: metric tonnes per hour for CO_2 and kilograms per hour for NOx and SO_2. Partial operating hours are normalized by the corresponding operating-time fraction, as described above. This formulation matches a PEMS use case in which plant operating signals are used to reconstruct or independently estimate the emission rate for the same hour. It also makes the physical and learned approaches directly comparable because both receive information that is available at the time of estimation.

The strongest feature configuration includes current operating variables, historical operating features, and, when available, lagged measured emissions. That last feature group changes the deployment interpretation. A model that uses recent measured emissions can check a live CEMS and reconstruct short gaps, but it is not fully independent of the monitoring system that supplies those lags.

We therefore distinguish CEMS-assisted and CEMS-independent settings throughout the analysis. Results from configurations that include lagged measured emissions should be read as evidence for cross-checking and short-gap reconstruction, not as performance of an autonomous replacement during a prolonged CEMS outage.

Leakage-controlled evaluation design

We did not use random record-level splitting. Consecutive observations from the same generating unit are temporally dependent, and a random split can place neighbouring hours from one operating episode in both training and testing. Instead, feature engineering, scaling, emission-factor calibration, hyperparameter selection, and uncertainty calibration were fitted only on training and validation data, and three complementary test schemes were used. The blocked-time split evaluates later observations from units seen during development. The plant-holdout split removes complete generating units from training and validation and assigns all of their hours to the test set, providing a direct test of cross-unit transfer. The seasonal-holdout split separates training and test observations by season to probe changes in demand, ambient conditions, schedules, and plant behaviour. Calibration, substituted-data, offline, and otherwise rejected records were removed before any of these partitions were created.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Materials and methods.

Model formulation and implementation

PHASE starts from a pollutant-specific physical estimate and learns only the part of the observed emission rate that remains unexplained. The physical term is an emission-factor estimate based on heat input and fuel information. Its scale is calibrated on the training partition, after which the residual learner is fitted to the difference between the measured emission and that scaled estimate. Validation and test records are never used to fit the physical scale, so the physical component follows the same leakage restrictions as the learned component.

The expected usefulness of this decomposition differs by pollutant. For CO_2, the factor term should already capture most of the signal because fuel consumption and carbon content dominate the mass balance. For NOx and SO_2, the same term is only a coarse starting point: hourly emissions also respond to combustion state, fuel properties, load transitions, and control-system operation. We therefore do not constrain PHASE to rely on the prior, and the ablation experiments explicitly test whether the physical term improves or worsens predictive accuracy.

The model input is organized into feature groups so that each group can later be removed in the ablation study. Current operating features include gross load, heat input, steam load, operating time, and ramp rate. Historical features include lags at 1, 2, 3, 6, and 24 hours together with rolling means and standard deviations. We add operating-regime and calendar indicators and static attributes such as nameplate capacity, primary fuel, commissioning year, geographic location, and available SCR/FGD indicators. All lagged and rolling features are calculated independently for each generating unit and stop at the prediction hour.

Two deployment configurations are evaluated. When target-derived lags are included, PHASE is CEMS-assisted and is intended for cross-checking or reconstruction of short missing segments. Removing those lags produces a more independent soft sensor. We report the distinction because a high score in the assisted configuration cannot be interpreted as evidence that the same model will remain accurate through a long period without direct emission measurements.

PHASE-lite estimates uncertainty with quantile regression. Separate gradient-boosted models predict the lower, median, and upper conditional residual quantiles; the 5th and 95th percentiles define the nominal 90% interval and the median provides the central residual estimate. The neural PHASE variant instead predicts a conditional mean and variance and is trained with a Gaussian negative log-likelihood objective. We evaluate the resulting intervals by empirical coverage, mean width, and pinball loss. Coverage answers whether observations fall inside the stated interval as often as expected, while width and pinball loss measure how informative and accurate those intervals are.

Comparison models and evaluation metrics

The benchmark compares five modelling choices: persistence, a scaled emission-factor baseline, ridge regression, XGBoost, and PHASE-lite. Persistence uses the most recent available measured emission and serves as a strong sanity check

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

for an autocorrelated hourly series. Ridge regression provides a linear data-driven reference and XGBoost a nonlinear tree-based reference. Point predictions are scored with MAE, RMSE, R^2, and peak-detection F1. MAE summarizes typical absolute deviation, RMSE penalizes large misses more strongly, and R^2 measures the fraction of observed variability explained by the model.

High-emission behaviour is evaluated separately rather than being left to the average-error metrics. We define peaks as observations in the highest emission decile, using a threshold estimated from training data and then fixed for the test set. Precision, recall, and F1 are calculated for these high-emission hours. Every metric is reported separately for CO_2, NOx, and SO_2 because the three targets have different units, scales, and formation mechanisms; values should therefore be compared across models within a pollutant, not across pollutants.

The ablation experiment removes one feature group at a time while keeping the rest of the training and evaluation procedure unchanged. We test lagged history, current operating variables, rolling statistics, operating-regime indicators, static metadata, and the physical component. A separate robustness experiment masks selected channels only at test time to imitate sensor loss, including recent emission history and the main operating signals. For plant transfer, complete generating units are excluded from source training. The zero-shot experiment uses no labelled target-unit observations, and subsequent runs add increasing amounts of local data to measure the gain from adaptation.

We compare PHASE-lite with the strongest purely data-driven model using three complementary statistical procedures. First, a paired bootstrap interval is computed for the difference in MAE. Second, a Wilcoxon signed-rank test is applied to paired per-unit errors to check whether the direction of the difference is consistent across generating units. Third, the Diebold-Mariano test compares the time-indexed squared-error loss differential and therefore places more emphasis on large prediction errors (Diebold et al., 1995). The three tests answer related but not identical questions, which becomes important when interpreting the NOx results.

For the bootstrap and per-unit tests, the generating unit is preserved as the grouping level instead of treating millions of hourly rows as independent samples. The temporal comparison likewise accounts for dependence in successive hourly errors. Data partitions and random seeds are fixed across experiments. The preprocessing, feature-generation, model, and evaluation scripts were archived with the project materials, while the raw CAMPD and EIA-860 datasets remain available from their original public sources.

Results.

Table 1 gives the blocked-time results and immediately shows that pollutant identity matters more than a single model ranking. For CO_2, the scaled emission-factor baseline already reaches $R^2 = 0.992$ with MAE = 2.79 t/h. XGBoost improves that result to $R^2 = 0.999$ and MAE = 0.88 t/h, but it is refining an already accurate physical estimate. The situation changes for NOx: the factor baseline has $R^2 = 0.394$ and MAE = 57.56 kg/h, whereas XGBoost reaches $R^2 = 0.973$ and MAE = 4.39 kg/h, reducing

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

MAE by about 92.4%. SO_2 follows the same pattern, with R^2 increasing from 0.413 to 0.980 and MAE falling from 43.10 to 5.21 kg/h, an improvement of about 87.9%. Figure 1 summarizes the contrast. A constant factor captures most of the predictable CO_2 signal, but it does not represent the operating-state and temporal variation required for hourly NOx and SO_2 estimation.

Table 1. Blocked-time test performance. MAE and RMSE are reported in metric tonnes per hour for CO_2 and kilograms per hour for NOx and SO_2. Compare models only within the same pollutant.

Pollutant / Model	MAE	RMSE	R2	peak F1
CO2 — XGBoost	0.88	3.18	0.999	0.994
CO2 — Emission factor	2.79	11.94	0.992	0.983
CO2 — Persistence	8.41	24.01	0.968	0.963
NOx — XGBoost	4.39	18.60	0.973	0.968
NOx — Ridge	6.55	23.68	0.956	0.956
NOx — Emission factor	57.56	87.96	0.394	0.588
SO2 — XGBoost	5.21	31.12	0.980	0.964
SO2 — Emission factor	43.10	168.98	0.413	0.509

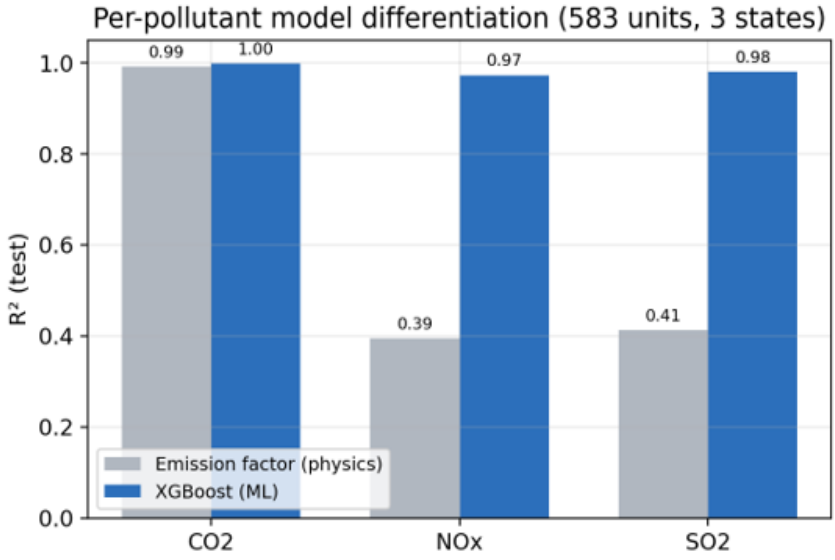


Fig. 1. Per-pollutant model differentiation (test R^2). The emission-factor baseline is strong for CO_2 but weak for NOx and SO_2.

The NOx ranking is stable across the three leakage-controlled schemes in Table 2. XGBoost remains between $R^2 = 0.969$ and 0.981, with MAE from 3.72 to 5.28 kg/h. The largest error occurs in plant-holdout evaluation, which is also the most demanding split because the test equipment is unseen during training. By comparison, the emission-factor baseline stays between $R^2 = 0.374$ and 0.395 and MAE = 53.17–60.66 kg/h. The gap therefore cannot be explained by one particular time partition: the fixed NOx factor remains weak under temporal, seasonal, and cross-unit shifts.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Table 2. NOx performance across blocked-time, plant-holdout, and seasonal-holdout evaluation.

NOx — split	XGBoost MAE	XGBoost R2	Em.factor MAE	Em.factor R2
Blocked-time	4.39	0.973	57.56	0.394
Plant-holdout	5.28	0.969	60.66	0.374
Seasonal-holdout	3.72	0.981	53.17	0.395

The uncertainty results in Fig. 2 are close to the stated nominal probabilities. PHASE-lite covers 79.0% of observations with the nominal 80% interval and 89.2% with the nominal 90% interval; the corresponding pinball loss is 1.46. In an earlier single-state subset, the nominal 90% interval covered 85.8% of observations. The multi-state result is therefore better calibrated in this comparison, although the experiment does not isolate whether the improvement comes from sample size, broader operating diversity, or another dataset difference.

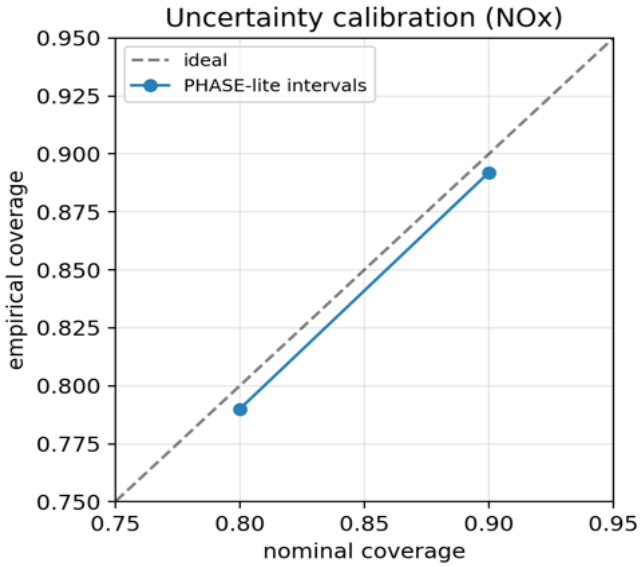


Fig. 2. Calibration of PHASE-lite prediction intervals for NOx: empirical coverage closely follows the nominal level.

Table 3. Point-prediction comparison of XGBoost and PHASE-lite for NOx.

Model	MAE	RMSE	peak F1
XGBoost	4.39	18.60	0.968
PHASE-lite	5.08	18.74	0.965

Table 4 identifies the feature groups that carry the NOx model. Removing lagged history increases MAE from 5.08 to 13.50 kg/h and lowers peak F1 from 0.965 to 0.898, by far the largest deterioration in the ablation. Removing current operating variables has a smaller effect, raising MAE to 5.17 kg/h, while static metadata, regime indicators, and rolling statistics change the score only modestly. The most important

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

negative result is the no-physics row: MAE falls to 4.36 kg/h when the simple prior is removed, compared with 5.08 kg/h for the full hybrid. In-domain NOx accuracy therefore does not benefit from this particular physical prior.

Figure 4 reports the transfer experiment. With no labelled data from the target units, PHASE-lite still reaches $R^2 = 0.965$ and MAE = 5.80 kg/h in plant-holdout evaluation. Adding one week of local observations reduces MAE to 5.59 kg/h, and three months of adaptation raises R^2 to 0.972. These are strong cross-unit results within the analysed U.S. reporting environment. They should not, however, be read as evidence of international transfer, because both source and held-out plants come from the same national data and regulatory system.

Table 4. Feature-group ablation for NOx (PHASE-lite). Recent history is the dominant feature group; the simple physical prior does not improve in-domain MAE.

Ablation (NOx)	MAE	RMSE	peak F1
no_lags	13.50	37.56	0.898
no_operating	5.17	18.76	0.963
full_hybrid	5.08	18.74	0.965
no_static	5.07	19.50	0.964
no_regime	5.06	18.73	0.965
no_rolling	5.02	18.44	0.966
no_physics	4.36	18.61	0.968

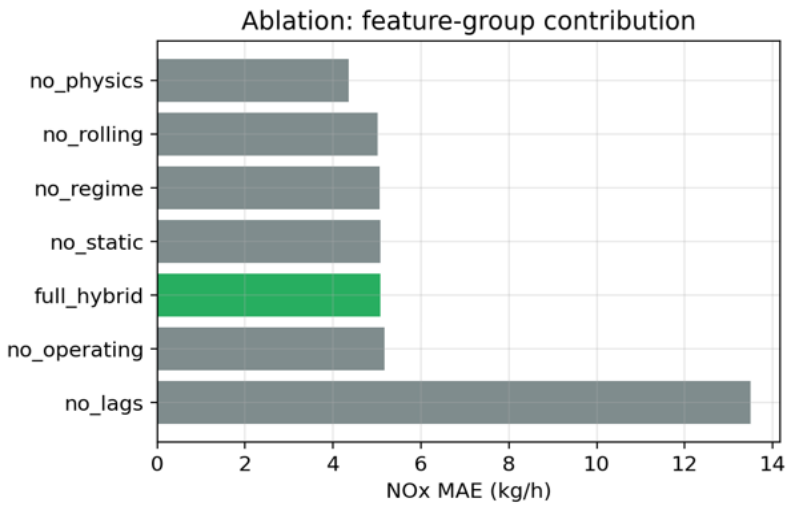


Fig. 3. NOx feature-group ablation; larger MAE increases indicate more influential feature groups.

The aggregate scores hide substantial variation by operating condition. As shown in Fig. 5, ramping hours have MAE = 12.49 kg/h, compared with 4.05 kg/h at baseload, and coal units are more difficult than gas units (13.84 versus 2.99 kg/h). The masking experiment in Fig. 6 points to the same dependence on recent state information. Removing recent emission history at test time drives MAE from 5.08 to 46.52 kg/h, while masking load and heat-input variables raises it to 20.33 kg/h. These results identify the channels that a practical CEMS-assisted deployment must preserve

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

reliably.

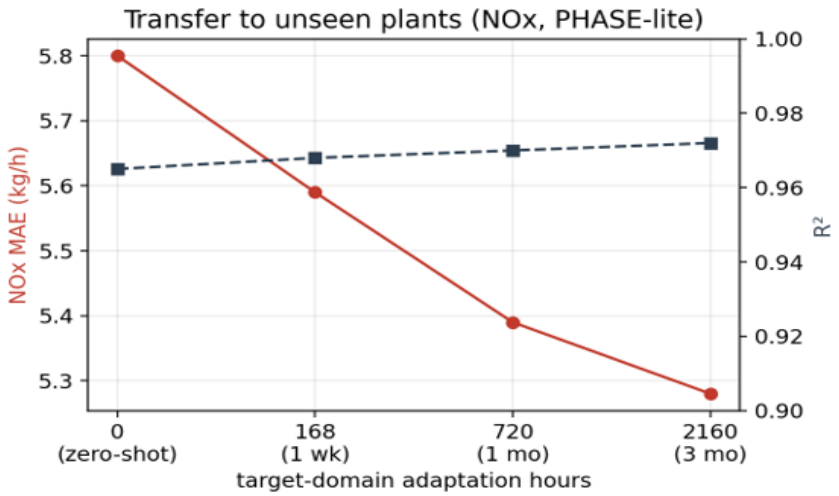


Fig. 4. Transfer to unseen plants: zero-shot performance and the effect of adding local labelled data.

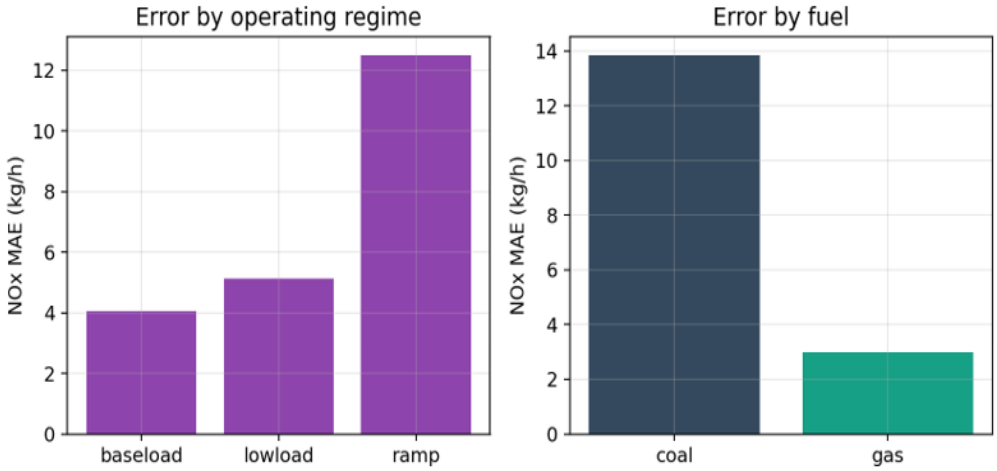


Fig. 5. NOx MAE by operating regime (left) and primary fuel group (right).

The statistical comparison clarifies why no single statement such as 'PHASE is better than XGBoost' is supported by the data. Across 509 test units, the paired bootstrap interval for MAE difference, defined as XGBoost minus PHASE-lite, is [-0.708, -0.684] kg/h; the entire interval is negative, so XGBoost has the lower average absolute error. The Wilcoxon signed-rank test confirms that this direction is consistent across units ($p < 10^{-46}$). The squared-error Diebold-Mariano test points in the opposite direction ($DM = 3.39$, $p = 6.9 \times 10^{-4}$) and favours PHASE-lite. Because squared error weights large misses more heavily, the two results are compatible: XGBoost is better on average MAE, while PHASE-lite avoids some high-magnitude errors. We therefore treat the hybrid as a robustness-oriented alternative rather than a uniformly more accurate model.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Discussion.

Limitations

Taken together, the experiments argue against using one modelling recipe for all three pollutants. CO_2 is the clearest case. In Table 1, a scaled emission factor alone explains almost all observed variability ($R^2 = 0.992$), which makes the physical baseline both accurate and easy to audit. XGBoost reduces the remaining error further, but the practical benefit of that reduction has to be weighed against the additional model complexity. For a fuel-bound pollutant such as CO_2, transparency can therefore be retained without sacrificing much predictive performance.

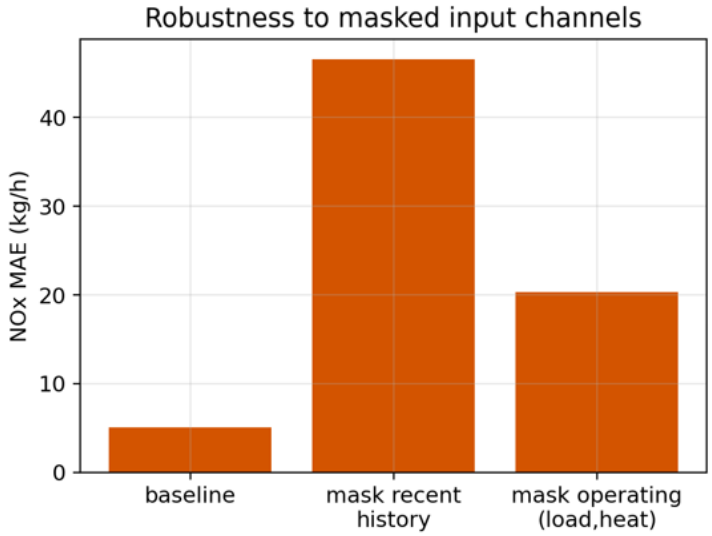


Fig. 6. Sensitivity to masked input channels during NOx inference.

NOx and SO_2 behave differently. Their factor baselines explain less than half of the observed variance, while learned models recover most of it across the tested splits. The reason is consistent with the feature and regime analyses: hourly values change with combustion conditions, load transitions, fuel properties, and control-system operation. The larger NOx errors during ramping and the strong dependence on recent history show that transient state matters. A fleet-wide constant cannot encode that state, so data-driven models are the more appropriate starting point for these pollutants.

The lag ablation also sets an important boundary on how the strongest model can be deployed. Historical features clearly improve NOx estimation, but part of that history consists of recent measured emissions. That makes the best-performing configuration CEMS-assisted. It is suitable for cross-checking a live monitor, detecting anomalous readings, and reconstructing short gaps; it is not a fully independent replacement after a long outage, when measured-emission lags are no longer available. For deployment studies, assisted and independent configurations should therefore be reported separately rather than combined under one PEMS accuracy figure.

The physical prior for NOx produced another useful negative result. Removing it improved average MAE, so with roughly 10 million labelled unit-hours the gradient-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

boosted learner did not need this simple prior to fit the in-domain relationship. At the same time, the full hybrid performed better under the squared-error Diebold-Mariano comparison and delivered calibrated prediction intervals. The contribution of PHASE is therefore not a universal reduction of point error. In this experiment its stronger case is uncertainty-aware prediction and reduced exposure to some large-error episodes.

Plant-holdout performance is encouraging but should be interpreted narrowly. PHASE-lite retains high accuracy on units that were absent from source training, which demonstrates cross-unit transfer inside this dataset. It does not show that the same accuracy will hold in another country, fuel system, or regulatory environment. Nor can the plant-holdout score by itself prove that the physical component caused the transfer benefit. A transfer-specific comparison of the full hybrid with an otherwise identical no-physics model is required before that effect can be attributed to the prior.

The benchmark has several limitations. First, it covers three U.S. states and two calendar years. The blocked-time, seasonal-holdout, and plant-holdout experiments test several forms of distribution shift inside that frame, but they do not cover differences in fuels, boiler technologies, control practices, maintenance conditions, or regulation that may appear in other power systems.

Second, the public data do not contain every variable that would be desirable for combustion and control modelling. Stack oxygen concentration, flue-gas temperature, detailed fuel composition, sulfur content, reagent consumption, and time-resolved control-device status are missing or unavailable with sufficient coverage. This matters most for NOx and SO_2, where those variables are directly connected to formation and removal mechanisms. Access to richer process measurements could improve both prediction and the physical relevance of the prior.

Third, the NOx and SO_2 priors used here are intentionally simple. A more detailed physical term could include boiler type, fuel sulfur, combustion-temperature proxies, excess-air indicators, and control-device efficiency. Such a prior may behave differently from the one evaluated in Table 4, so the present result should be read as evidence against this simple formulation, not against every possible physics-informed design.

Fourth, some control-technology attributes are available only at plant level and are transferred to generating units through the EPA-EIA crosswalk. At multi-unit plants with different equipment configurations, that mapping can introduce assignment uncertainty and weaken the meaning of a nominally unit-level control indicator.

Fifth, the strongest configurations depend on lagged measured emissions. That dependence is acceptable for a CEMS-assisted soft sensor, but it limits fully autonomous use. A separate CEMS-independent evaluation is necessary before recommending the model for long monitoring failures or for plants that have no historical continuous-emission record.

Finally, the study has no supervised external test in Kazakhstan because comparable public hourly CEMS labels were not available. Sentinel-5P TROPOMI NO_2 observations can provide a future corroborative comparison over major thermal plants, but they cannot replace stack-level labels. Satellite columns are affected

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

by atmospheric transport, chemistry, cloud screening, revisit timing, and spatial aggregation, so any Kazakhstan analysis must be presented as supporting evidence rather than direct validation of hourly stack emissions.

The practical conclusion is therefore conditional rather than universal. For CO_2, an emission-factor model is already a strong choice when auditability matters. For NOx and SO_2, learned representations of changing plant operation are necessary for high hourly accuracy. A physics-informed hybrid becomes most relevant when calibrated uncertainty, protection against large errors, or transfer with limited local labels is more valuable than the smallest average in-domain MAE.

Conclusion.

This work establishes a leakage-controlled benchmark for hourly CO_2, NOx, and SO_2 estimation using 10,144,680 unit-hours from 583 generating units and compares physical, statistical, machine-learning, and physics-informed approaches under blocked-time, seasonal-holdout, and plant-holdout evaluation. The main result is not that one model wins everywhere. The pollutant determines the modelling requirement. A scaled emission factor is already highly accurate for CO_2, while NOx and SO_2 need models that learn dynamic operating state and temporal history. For NOx, the simple PHASE prior does not improve average in-domain MAE over a tuned XGBoost baseline, but the hybrid provides calibrated intervals, reduces some large squared-error events, and retains strong accuracy on unseen units. PHASE should therefore be viewed as an uncertainty- and robustness-oriented framework rather than a universally more accurate black-box replacement. The current plant-holdout experiment demonstrates transfer within the analysed U.S. dataset only. The next evaluation steps are to separate CEMS-assisted from fully independent configurations, test stronger pollutant-specific priors, incorporate richer combustion and control variables, and validate on independently labelled non-U.S. data. The planned Sentinel-5P TROPOMI comparison over Kazakhstan thermal plants is a corroborative deployment-oriented step, not a substitute for stack-level validation.

REFERENCES

Adams, D., Oh, D.-H., Kim, D.-W., Lee, C.-H., & Oh, M. (2020). Prediction of SOx–NOx emission from a coal-fired CFB power plant with machine learning: Plant data learned by deep neural network and least square support vector machine // Journal of Cleaner Production. — Vol. 270. — Article 122310. DOI: https://doi.org/10.1016/j.jclepro.2020.122310. [In Eng.].

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv. DOI: https://doi.org/10.48550/arXiv.1803.01271. [In Eng.].

Breiman, L. (2001). Random forests // Machine Learning. — Vol. 45. — No. 1. — Pp. 5–32. DOI: https://doi.org/10.1023/A:1010933404324. [In Eng.].

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — Pp. 785–794. Association for Computing Machinery. DOI: https://doi.org/10.1145/2939672.2939785. [In Eng.].

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy // Journal of Business & Economic Statistics. — Vol. 13. — No. 3. — Pp. 253–263. DOI: https://doi.org/10.1080/07350015.1995.10524599. [In Eng.].

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory // Neural Computation. — Vol. 9. — No. 8. — Pp. 1735–1780. DOI: https://doi.org/10.1162/neco.1997.9.8.1735. [In Eng.].

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning // Nature Reviews Physics. — Vol. 3. — Pp. 422–440. DOI: https://doi.org/10.1038/

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

s42254-021-00314-5. [In Eng.].

Koenker, R., & Bassett, G., Jr. (1978). Regression quantiles // Econometrica. — Vol. 46. — No. 1. — Pp. 33–50. DOI: https://doi.org/10.2307/1913643. [In Eng.].

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles // In Advances in Neural Information Processing Systems. — Vol. 30. — Pp. 6402–6413. [In Eng.].

Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting // International Journal of Forecasting. — Vol. 37. — No. 4. Pp. 1748–1764. DOI: https://doi.org/10.1016/j.ijforecast.2021.03.012. [In Eng.].

Lv, Y., Liu, J., Yang, T., & Zeng, D. (2013). A novel least squares support vector machine ensemble model for NOx emission prediction of a coal-fired boiler // Energy. — Vol. 55. — Pp. 319–329. DOI: https://doi.org/10.1016/j.energy.2013.02.062. [In Eng.].

Tan, P., He, B., Zhang, C., Rao, D., Li, S., Fang, Q., & Chen, G. (2019). Dynamic modeling of NOx emission in a 660 MW coal-fired boiler with long short-term memory // Energy. — Vol. 176. — Pp. 429–436. DOI: https://doi.org/10.1016/j.energy.2019.04.020. [In Eng.].

U.S. Environmental Protection Agency. (n.d.). Clean Air Markets Program Data: Bulk data files. Retrieved August 3, 2026, from DOI: https://campd.epa.gov/data/bulk-data-files. [In Eng.].

U.S. Environmental Protection Agency. (2026b). EPA–EIA Power Sector Data Crosswalk. U.S. Environmental Protection Agency. [In Eng.].

U.S. Energy Information Administration. (n.d.). Form EIA-860 detailed data with previous form data. Retrieved August 3, 2026, from https://www.eia.gov/electricity/data/eia860/. [In Eng.].

Vovk, V., Gammerman, A., & Shafer, G. (2005). Algorithmic learning in a random world. Springer. DOI: https://doi.org/10.1007/b106715. [In Eng.].

Zhou, H., Cen, K., & Fan, J. (2004). Modeling and optimization of the NOx emission characteristics of a tangentially fired boiler with artificial neural networks // Energy. — Vol. 29. — No. 1. — Pp. 167–183. DOI: https://doi.org/10.1016/j.energy.2003.08.004. [In Eng.].

Zhu, B., Ren, S., Weng, Q., & Si, F. (2024). A physics-informed neural network that considers monotonic relationships for predicting NOx emissions from coal-fired boilers // Fuel. — Vol. 364. — Article 131026. DOI: https://doi.org/10.1016/j.fuel.2024.131026. [In Eng.].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

Собственник:
АО «Международный университет информационных технологий» (Казахстан, Алматы)

Главный редактор:
Колесникова Катерина Викторовна

Ответственный редактор:
Мрзабаева Раушан Жалиевна

Компьютерная верстка:
Калабай Замзагуль Ертугановна

Сайт журнала: https://journal.iitu.edu.kz
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)

Подписано в печать 30.09.2026.
050040 г. Алматы, ул. Манаса 34/1, каб. 709, тел: +7 (727) 244-51-09).

Издание АО "Международного университета информационных технологий"