1st Workshop of the Institute of Artificial Intelligence

Title e Abstract

Lecture

Speaker

Altigran Soares e Juan Colonna - UFAM

Title: *Laboratory for Artificial Intelligence Research in the Amazon* (Laboratório de Pesquisas para Inteligência Artificial na Amazônia)

Abstract: In this lecture, we will address the main initiatives of the Laboratory for Artificial Intelligence Research in the Amazon (LAPIAM) at the Institute of Computing, UFAM (IComp/UFAM). We will discuss artificial intelligence techniques applied to ecological restoration and environmental monitoring of Amazonian ecosystems. The presentation will include projects related to the use of bioacoustic indices for biodiversity estimation, the detection and classification of animal species from images and vocalizations, as well as water quality monitoring and the use of IoT sensors.

A key highlight will be the development of large-scale multimodal models (MLLMs), which integrate data from various sources, such as audio, images, and sensors, to identify patterns in primary and secondary forest areas. We will also discuss the impact of anthropogenic noise on local fauna and the use of explainable AI (XAI) approaches to decode these changes. The lecture will explore how these advances contribute to environmental conservation and offer new solutions for ecological monitoring and management in the region.

Nesta palestra, abordaremos as principais iniciativas do Laboratório de Pesquisas para Inteligência Artificial na Amazônia (LAPIAM), do Instituto de Computação da UFAM (IComp/UFAM). Discutiremos técnicas de inteligência artificial aplicadas à restauração ecológica e ao monitoramento ambiental de ecossistemas amazônicos. Serão apresentados projetos relacionados ao uso de índices bioacústicos para estimativa de biodiversidade, detecção e classificação de espécies animais a partir de imagens e vocalizações, além do monitoramento da qualidade da água e do uso de sensores IoT. Um dos destaques será o desenvolvimento de modelos multimodais de larga escala (MLLMs), que integram dados de diferentes fontes, como áudio, imagens e sensores, para identificar padrões em áreas de florestas primárias e secundárias. Também discutiremos o impacto de ruídos antropogênicos sobre a fauna local e o uso de abordagens de IA explicável (XAI) para decodificar essas alterações. A palestra explora como esses avanços contribuem para a preservação ambiental e oferecem novas soluções para o monitoramento e a gestão ecológica na região.

Speaker

Antonio Tadeu Gomes

Title: Physics-Informed Machine Learning: Methods and Applications

Abstract: In this talk we present our research group IPES (Innovative Parallel numErical Solvers) at LNCC and some of the ongoing developments and results obtained so far by our group in the area of physics-informed machine learning, covering topics such as: Physics-Informed Neural Networks (PINNs), Deep Networks for Operator Learning (DeepONets), and their interaction with classical numerical methods, in particular multiscale finite element methods. We cover both methodological aspects and applications in areas such as oil & gas and urban mobility.

Speaker

Álvaro Coutinho

Title: Recent Advances in Scientific Machine Learning for Computational Mechanics

Abstract: In recent years, there has been significant interest in using data-driven methods to solve problems in science and engineering. Numerical simulations for these problems can be costly, making data-driven methods valuable for understanding and improving efficiency in quantifying and predicting states. This talk will review recent advancements in Scientific Machine Learning for Computational Mechanics, such as dynamic mode decomposition, physics-informed neural networks, manifold learning, and neural operators, as applied to relevant problems. These problems are of interest in sustainable resource exploration, geophysics, and various industrial applications. The talk will show how data-driven information can improve predictions, help explore parametric manifolds for unseen scenarios, and reconstruct high-dimensional simulations with lower-dimensional structures in a feasible time.

Speaker

Antonio Henrique Carlan Junior (UFPel) e Gabriela Mores (LNCC)

Title: ONIA Brazil - Digital literacy and talent screening through education

Abstract: The Brazilian National AI Olympiad (ONIA-Brazil) is an educational initiative organized jointly with the IIA/LNCC that aims to use a scientific Olympiad to reach a large number of young people, digitally teach those who have a less solid foundation and select the best participants for the International Olympiad in AI (IOAI).

In collaboration with the researcher from IIA involved in this initiative, the coordinator of ONIA and board member of IOAI will report on the results of the initial phase of the 1st ONIA, the

educational impacts of the initiative, the participation profile of the Brazilian states and how the Olympiad places Brazil in the international spotlight for digital literacy in basic education.

Speaker

Carla Osthoff

Title: The Santos Dumont Supercomputer in the National and International Landscape of HPC and AI Research (O Supercomputador Santos Dumont no cenário nacional e internacional das pesquisas de HPC e IA)

Abstract: The SDumont supercomputer, hosted at LNCC/MCTIC in Petrópolis, RJ, is funded by MCTIC to serve the Brazilian academic community. This lecture will present the features of the new architecture of the supercomputer, which delivers approximately 23 PetaFlops, and the future AI Supercomputer, as part of the National Artificial Intelligence Plan (PBIA). Additionally, it will highlight the research and collaborations developed by the High-Performance Computing (HPC) sector at LNCC.

O supercomputador SDumont, hospedado no LNCC /MCTIC, na cidade de Petrópolis-RJ, financiado pelo MCTIC para atender a comunidade acadêmica brasileira. Nesta palestra serão apresentadas as características da nova arquitetura do supercomputador de cerca de 23 PetaFlops e do futuro Supercomputador para IA, como parte do Plano Nacional de Inteligencia Artificial (PBIA) e das pesquisas e colaborações desenvolvidas pelo setor de Processamento de Alto Desempenho do LNCC.

Speaker

Carmen Bonifácio

Title: The impact of Generative AI on the Brazilian labor market

Abstract: Generative AI is already changing the labour process.Thinking about the Future Of Work and the actions needed for the country to be able to handle the changes in the best way, we developed a sample survey aimed at reflecting the real and current scenario of the impacts of generative AI on the labor market in Brazil. This initial study highlights the vision of the main players involved in the production chain in the face of the new technologies available.

Speaker

Dennis Shasha

Title: DietNerd: Algorithms, Architecture, and Experiments of a Large Language

Abstract: DIETNERD is a large language model-based system designed to enhance public health education in diet and nutrition. The system responds to user questions with concise, evidence-based summaries and assesses the quality and potential biases of cited research. This talk

describes the system's workflow, back end implementation, and the prompts used. Accuracy and quality-of-response results are presented based on an automated comparison against systematic surveys and against the responses of similar state-of-the-art systems through human feedback from registered dieticians. Thus, DIETNERD could be a tool to bridge the gap between complex scientific literature and public understanding. DIETNERD can be accessed at https://dietnerd.org/

Speaker

Eduardo Bezerra

Title: RioNowcast: AI for Precipitation Nowcasting

Abstract: Nowcasting, the short-term forecasting of weather phenomena, is crucial for various applications such as flood prediction, agriculture, and transportation planning. In this presentation, we provide an overview of the RioNoscat consortium, a multi-institutional research project that aims at using AI techniques to build predictive models for precipitation nowcasting. Concretely, in this project, we investigate approaches to leverage multiple data sources to train Machine Learning models for precipitation nowcasting. We present some preliminary approaches and corresponding experimental results in the context of this project. Additionally, we address challenges and considerations specific to the integration of heterogeneous data sources to build machine learning models in a spatio-temporal context.

Speaker

Fábio Cozman

Title: Research at the USP Center for Artificial Intelligence **(**Pesquisas no Centro de Inteligência Artificial da USP)

Abstract: Este seminário apresenta um pouco da história do C4AI, o Centro de Inteligência Artificial USP/IBM/FAPESP, e discute alguns dos seus principais resultados em quatro anos de operação. O centro foi criado para avançar a pesquisa em inteligência artificial, investigando, entre outros temas, o processamento de línguas do Brasil, os agentes conversacionais com capacidade de raciocínio, o aprendizado de máquina conectado a modelos físicos, e o impacto da tecnologia de IA na sociedade. O C4AI tem procurado não só desenvolver essa tecnologia, mas também debatê-la e disseminá-la. Após uma apresentação geral do C4AI e das lições aprendidas na sua construção, o seminário se deterá por alguns tópicos específicos, em particular argumentação automatizada e inferência causal.

This seminar provides an overview of the history of C4AI, the USP/IBM/FAPESP Center for Artificial Intelligence, and discusses some of its main achievements over four years of operation. The center was established to advance AI research, exploring topics such as the processing of Brazilian languages, conversational agents with reasoning capabilities, machine learning

integrated with physical models, and the societal impact of AI technology. C4AI aims not only to develop this technology but also to debate and disseminate it. Following a general introduction to C4AI and the lessons learned during its development, the seminar will focus on specific topics, particularly automated argumentation and causal inference.

Speaker

Fernando Schapachnik

Title: Some myths about the labor market and AI.

Abstract: In this talk we will walk through some ideas about AI, technology and the labor market. The idea is to go beyond the simple binary question of whether AI will replace workers or not, and visit some of the historical data to better understand de challenges involved.

Speaker

Gilson Giraldi

Title: Cell Image Segmentation: Past, Present and Challenges

Resumo: Despite the scientific advances since the beginning of cell theory in the early 19th century, attaining a complete understanding of cellular mechanisms remains an open issue. Today, biologists have several microscopic imaging techniques to visualize cellular phenomena by acquiring high-resolution image volumes. Consequently, computational techniques for image analysis have become fundamental for subsequent progress in cell biology. In this line, image segmentation is, in general, a central problem considering the importance of cellular morphology and the role of cell boundary for the analysis of intra and inter-cellular processes. This avenue starts with traditional approaches based on intensity thresholding and deformable models, among others. More recently, owing to the success of deep learning techniques in data analysis, deep neural networks have been extensively applied for cell image segmentation with outstanding results. However, the huge amount of 3D data and variability of observed phenomena pose new challenges for cell image segmentation and analysis at present.

Speaker

Isabella Guedes

Title: Insights from WAIC2024 and Meetings with Leading Chinese AI Institutions

Abstract: This talk will share insights from Brazil's mission to the World Artificial Intelligence Conference 2024 (WAIC2024) and formal visits to Chinese institutions such as the Shanghai Data Exchange, Shanghai AI Institute, Kuaishou, and BAAI. We explored approaches to data exchange, AI infrastructure, and innovative research practices that could enrich Brazil’s AI ecosystem. The presentation will discuss actionable insights and strategies from these interactions, highlighting pathways for enhanced international cooperation in the AI field.

Speaker

José Antonio Macedo

Title: A.I. for Health

Resumo: The promotion of health through the use of Artificial Intelligence (AI) presents significant potential for the remote monitoring of patients with chronic diseases, alleviating the burden on hospitals and improving overall well-being. The advancement of technologies such as smartphones, cloud computing, and the Internet of Things (IoT) has facilitated access to data and empowered individuals in self-care. While AI has already proven effective in medical diagnostics, its application in health monitoring through wearable devices is still in its early stages. In this presentation, we aim to discuss research findings related to improving quality of life through AI in the areas of health and well-being, focusing on four key research areas: machine learning model management, anomaly detection in biomedical data, health monitoring via wearable devices, edge AI computing, and AI-assisted anamnesis. We will highlight how multidisciplinary collaboration between researchers and institutions is crucial for the success of these initiatives.

Speaker

Laurent Dardenne

Title: New Drug Prospecting: Merging Computational Intelligence and AI Strategies through DockThor-VS

Abstract: This presentation will highlight the development of the DockThor-VS web server (https://www.dockthor.lncc.br), a platform created by researchers at LNCC for identifying and designing novel drug candidates. Some recent and upcoming methodological developments in DockThor-VS will be discussed, with a particular focus on the integration of advanced computational and artificial intelligence techniques, especially for de novo drug design. The DockThor-VS web server is connected to the Brazilian supercomputer, Santos Dumont, providing a dedicated and freely accessible service to the scientific community.

Speaker

Marcel Pedroso

Title: The Data Science Platform for Health)PCDas-Fiocruz

Abstract: Relato de experiência sobre a construção e evolução da Plataforma de Ciência de Dados aplicada à Saúde (PCDaS) do Laboratório de Informação em Saúde (Lis) do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (Icict) da Fundação Oswaldo Cruz (Fiocruz). A Plataforma é um projeto de pesquisa e desenvolvimento tecnológico, em parceria com o Laboratório Nacional de Computação Científica (LNCC), lançada em 2016, a PCDaS tem

como objetivo principal disponibilizar serviços tecnológicos e computação científica para armazenamento, gestão, análise, visualização e disseminação de grandes quantidades de dados de saúde e seus determinantes socioambientais para pesquisadores, docentes e discentes de instituições de ensino e pesquisa, bem como gestores governamentais. Extra: Missão Xangai, China IIA-LNCC - experiências e aprendizados.

Speaker

Patrick Valduriez

Title: Inria Brasil Strategic Partnership: spotlight on AI

Abstract: Inria-Brasil is a strategic partnership between Inria and LNCC, and Brazilian universities. It has been formally signed on April 14, 2023 during the Inria-Brasil workshop at LNCC, by representatives of Inria and MCTI. The goal is to foster collaboration and research excellence on topics of shared interest, in particular, high-performance computing, artificial intelligence and data science. Innovation is also important, in particular, technology transfer to industry and impact on society. The partnership is supported through Inria's international relation programs (associated team, Inria international chair, postdoc, etc.) and Brazilian funding agencies (CAPES, CNPq, FAPs, ...).

Speaker

Renato Portugal

Title: Implementation of Quantum Neural Networks with 1-Qubit Neurons

Abstract: Neural networks, inspired by biological systems, have revolutionized machine learning and artificial intelligence. While classical neural networks have achieved remarkable success, quantum neural networks (QNNs) offer the potential for substantial advancements through quantum superposition and entanglement. In this work, we propose a novel QNN architecture based on 1-qubit neurons, each represented by a parameterized quantum circuit designed to maximize the degrees of freedom allowed by quantum mechanics. This streamlined yet powerful design enables the construction of QNNs with minimal hardware requirements by employing a sequence of these neurons, achievable with just a single qubit. We demonstrate the effectiveness of our approach across various tasks, including prediction and classification on both synthetic and real-world datasets. Our results indicate that increasing the number of 1-qubit neurons in the network enhances performance and enables the learning of complex, nonlinear patterns.

Speaker

Roberto Souto

Title: Data Assimilation by Machine Learning for Use in Weather Forecasting Models

Abstract: In computational modeling of physical phenomena, it is inherent to have errors in the representation of the natural state of the processes through mathematical models solved by numerical methods. Techniques that incorporate information from observational data of the phenomenon can be applied to reduce the uncertainty of this error. Combining observation data with a previous forecast is obtained with data assimilation techniques, which estimate the initial condition for the following forecast cycle. Data assimilation is a critical problem in several operational forecasting systems. The model must simulate meteorological phenomena on different time scales to verify whether the model with assimilation reproduces the weather systems. Forecasts at grid points and observation points must be evaluated compared to the predefined ground truth. Calculate differences between observations and analyses and between observations and the background to verify whether the analysis submitted to the assimilation process approaches the ground truth. In numerical weather forecasting, data assimilation is the process that demands the most computational effort of the entire forecasting system. One or more supervised neural networks can be configured/trained to emulate a data assimilation method. In this talk, we will show that results with supervised neural networks, such as other machine learning techniques, indicate a substantially reduced computation time for data assimilation with good precision.

Poster

Douglas Terra

Authors: Douglas Terra Machado, Otávio José Bernardes Brustolini, Yasmmin Côrtes Martins, Marco Antonio Grivet Mattoso Maia, Ana Tereza Ribeiro de Vasconcelos

Title: DEGRE: A GENERALIZED LINEAR MIXED MODEL TOOL FOR PAIRWISE DIFFERENTIAL GENE EXPRESSION ANALYSIS

Abstract: Technological advances in RNA-Seq and Bioinformatics have improved the quantification of gene transcriptional levels in cells, tissues, and cell lines, allowing for the identification of Differentially Expressed Genes (DEGs), which are important for understanding disease mechanisms, biological responses, and therapeutic targets. DESeq2 and edgeR are widely used tools for DEG inference that employ generalized linear models (GLMs), focusing on fixed effects within the experimental design. However, including random effects reduces the risk of missing potential DEGs that may be essential in the context of the studied biological phenomenon. By incorporating generalized linear mixed models (GLMMs), both fixed and

random effects are considered, resulting in more accurate and biologically relevant identification of DEGs. Methods: We present DEGRE (Differentially Expressed Genes with Random Effects), a user-friendly and unsupervised tool that enables researchers to include fixed and random effects in DEG analysis for RNA-Seq data. DEGRE preprocesses raw count matrices before applying GLMMs to gene-specific counts, thus preparing data for more reliable DEG detection under complex experimental designs. The tool utilizes the Wald test to assess the statistical significance of derived regression coefficients, with options for P-value adjustment through the Benjamini-Hochberg or Bonferroni methods, enhancing accuracy in DEG inference. Results: DEGRE was evaluated on simulated datasets with known identification of DEGs, where fixed effects were present, and random effects were introduced to measure the impact of experimental designs with high biological variability. DEGRE's preprocessing pipeline effectively removes technical variation from the count matrices and successfully retains overdispersed genes while eliminating technical noise, thus enhancing the stability of DEG inference. The biological coefficient of variation is inferred from the counting matrices to assess variability before and after the preprocessing. The DEGRE is computationally validated through its performance by simulating counting matrices, which have biological variability related to fixed and random effects. In a case study on transcriptomic data from patients with bipolar disorder, DEGRE identified novel candidate DEGs, underscoring its potential to reveal genes relevant to disease pathology and therapeutic interventions. Conclusions: DEGRE performs data preprocessing and applies GLMMs for DEGs' inference in complex biological variability studies. By efficiently filtering out genes that may affect inference accuracy, DEGRE improves the robustness of biological analyses. Also, the computational and biological validation of DEGRE has shown to be promising in identifying possible DEGs in experiments derived from complex experimental designs. This tool may help handle random effects on individuals in the inference of DEGs and presents a potential for discovering new interesting DEGs for further biological investigation.

Felipe Curcio

Authors: Felipe Curcio, Mariza Ferro, Fabio Porto, Eduardo Bezerra

Title: Data Fusion for Real-Time Precipitation Forecasting in a Spatiotemporal Context (Fusão de Dados para Previsão Imediata de Precipitação em um Contexto Espaço-Temporal)

Abstract: With the increasing availability of meteorological data from satellites, radars, and surface stations, along with advancements in numerical and reanalysis models, there is a growing demand to integrate and process this data in order to enhance forecasting and meteorological studies. In this work, we propose a data fusion approach for real-time precipitation forecasting by integrating data from meteorological and rain gauges in Rio de Janeiro with high-resolution reanalysis data from the ERA5Land model. We performed real-time precipitation forecasting using the STConvS2S model, a deep learning architecture for processing spatiotemporal data using only convolutional layers. The data is structured for a rectangular region of interest, with dimensions $Lat \times Lon$, where Lat and Lon correspond to the number of grid points in latitude and longitude, enabling an increase in the resolution and spatial coverage of the forecast. We present preliminary results of the proposed methodology with

integrated data from the Websirenes system, which consists of a network of 83 rain gauges in Rio de Janeiro, from April 2011 to April 2012. Furthermore, we present the coverage of stations from the National Institute of Meteorology (INMET), AlertaRio, and Websirenes in Rio de Janeiro, and conduct analyses on the formation of convective rainfall, particularly for days with extreme events.

Com o aumento da disponibilidade de dados meteorológicos de satélites, radares e estações de superfície, juntamente com os avanços em modelos numéricos e de reanálise, há uma demanda crescente para integrar e processar esses dados a fim de aprimorar previsões e estudos meteorológicos. Neste trabalho propomos uma abordagem de fusão de dados para previsão de precipitação imediata integrando dados de estações meteorológicas e pluviométricas no Rio de Janeiro com dados de reanálise de alta resolução do modelo ERA5Land. Realizamos a previsão de precipitação imediata utilizando o modelo STConvS2S, uma arquitetura de Aprendizado Profundo para processamento de dados espaço-temporais utilizando apenas camadas convolucionais. Os dados são estruturados para uma região retangular de interesse, com dimensões $Lat \times Lon$, onde Lat e Lon correspondem ao número de pontos de uma grade retangular em latitude e longitude, isso viabiliza um aumento na resolução e na cobertura espacial da previsão. Apresentamos os resultados preliminares da metodologia proposta com dados integrados do sistema Websirenes, que consiste em uma rede de 83 pluviômetros no Rio de Janeiro, de abril de 2011 até abril de 2012. Além disso, apresentamos a cobertura das estações do Instituto Nacional de Meteorologia (INMET), AlertaRio e Websirenes no Rio de Janeiro, e realizamos análises na formação de chuvas convectivas, em especial para dias com eventos extremos.

Frank Quispe

Title: QAOA Algorithm with Qiskit: A Variational Approach for Optimization Problems

Abstract: Considering the current legislation on data protection, in particular Law 13.709/2018 (General Law for the Protection of Personal Data), the signatories and recipients of this e-mail must keep confidentiality and guarantee confidentiality in relation to the personal data to which they have access through this electronic communication

Gabriela Moraes

Authors: Gabriela Moraes, Fábio Porto

Title: Integrating observations and predictions in ontologically grounded knowledge graphs

Abstract: Weather forecasting services face an increasingly challenging task of alerting the population about extreme weather events. Precipitation forecasting is a relevant research topic with significant impact on decision-making for monitoring urban areas. Motivated by the opportunities to apply machine learning models in predicting weather events, many researchers aim to build ML predictive models for extreme precipitation forecasting in urban areas. The construction of these models requires observational data for training, validation, and testing, sourced from different origins, typically collected and stored separately in files using a tabular data format that lacks explicit semantics. Through training these models and using observational

data representing the current atmospheric conditions, we can make inferences that need to be stored in a manner that facilitates access and reuse. This scenario has motivated the need for an integrated view of these data to support decision-making applications. Integrating predicted data with the observational data that contributed to those predictions offers several significant advantages. Among these advantages is model explainability. In this sense, the association between prediction results and the context in which the forecast was computed highlights the aspects that influenced the prediction. This work demonstrates that knowledge graphs are a useful tool for integrating observational data, predicted data, and the predictive models that relate them. We first propose the construction of the ObsML ontology, built upon a foundational ontology, from which alignment techniques between classes are extended to include those necessary to represent the entire study domain of the research. ObsML describes the entire scenario of weather predictions and provides a solid foundation for the creation and maintenance of the knowledge graph, assisting in information management. From a use case, we observe that the created ontology proved effective in supporting the construction of a knowledge graph intended for storing data involved in the training of weather prediction models and the inferences made by these models. With these results, we hope to establish a graph with a common semantic description, generating a higher level of abstraction that does not depend on the physical infrastructure or data format. Thus, the construction of decision-making applications in the meteorological context can benefit from the existence of a representation that integrates domain data.

Helano Jorge da Rocha Andrade

Authors: Helano Andrade, Rocío Zorrilla D.Sc., Tiziano Labruzzo,Prof. Roberto S. Pinto D.Sc.[1], Prof. Fábio Porto, D.Sc.

Title: Data Augmentation for Deep Learning on Tabular Data in Geophysics Analysis of Ocean subsurface using mCSEM: Evaluating HAT and Spline Techniques

Abstract: This work investigates the use of data augmentation techniques to address data limitations in training deep learning models for geophysical applications, specifically focusing on synthetic marine Controlled-Source Electromagnetic (mCSEM) data formatted as tabular datasets. Such geophysical data are often scarce and costly to acquire, presenting challenges for models to generalize effectively. In this context, we explore the feasibility of two data augmentation approaches: the Histogram Augmentation Technique (HAT), as discussed by Sathianarayanan et al. (2022), which is applied to both continuous and discrete columns and preserves the original data distribution. The effectiveness of the augmented data produced with HAT will be evaluated using the XGBoost model (Chen & Guestrin, 2016) to assess improvements in accuracy and robustness. Additionally, we employ splines as an independent technique for generating synthetic data. Splines provide smoothed representations of variables, which can enhance the fidelity of CSEM data (de Boor, 1978; Hastie et al., 2009). These approaches aim to improve model performance by increasing data diversity and fidelity, optimizing resources for high-resolution modeling of marine geophysical features.

Leandro C. Souza

Authors: Leandro C. Souza (UFPB), Bruno Guingo (doutorando PPG-LNCC), Gilson Giraldi (LNCC), Renato Portugal (LNCC)

Title: Implementation of Quantum Neural Networks with 1-Qubit Neurons

Abstract: Neural networks, inspired by biological systems, have revolutionized machine learning and artificial intelligence. While classical neural networks have achieved remarkable success, quantum neural networks (QNNs) offer the potential for substantial advancements through quantum superposition and entanglement. In this work, we propose a novel QNN architecture based on 1-qubit neurons, each represented by a parameterized quantum circuit designed to maximize the degrees of freedom allowed by quantum mechanics. This streamlined yet powerful design enables the construction of QNNs with minimal hardware requirements by employing a sequence of these neurons, achievable with just a single qubit. We demonstrate the effectiveness of our approach across various tasks, including prediction and classification on both synthetic and real-world datasets. Our results indicate that increasing the number of 1-qubit neurons in the network enhances performance and enables the learning of complex, nonlinear patterns.

Mauro Sérgio dos Santos Moura

Authors:

Title: Comparative Study of Transformers in Spatiotemporal Precipitation Forecasting (Estudo Comparativo de Transformers na Previsão Espaço Temporal de Precipitação)

Abstract: Precipitation is a meteorological phenomenon that plays a key role in various human activities, highlighting the need for the development of models for its forecasting. Deep Learning Models (DLMs) have shown excellent performance in various time series forecasting tasks. Among these DLMs, Transformer-based models stand out due to their attention mechanisms. This study aims to investigate the applicability of Transformer models, such as Autoformer, in comparison with classical models, like DLinear, for spatiotemporal precipitation forecasting, emphasizing how attention mechanisms identify spatiotemporal patterns.

A precipitação é um fenômeno meteorológico que desempenha um papel fundamental em diversas atividades humanas, o que ressalta a necessidade do desenvolvimento de modelos para sua previsão. Modelos de Aprendizado Profundo (Deep Learning Models, DLMs) têm demonstrado excelente desempenho em várias tarefas de previsão de séries temporais. Entre esses DLMs, destacam-se os modelos baseados em Transformers, que se sobressaem devido aos seus mecanismos de atenção. Este estudo tem como objetivo investigar a aplicabilidade de modelos Transformers, como o Autoformer, em comparação com modelos clássicos, como o DLinear, para a previsão espaço-temporal de precipitação, enfatizando como os mecanismos de atenção identificam padrões espaço-temporais.

Otávio J. B. Brustolini

Authors: Otávio J. B. Brustolini

Title: (m, n)-mer—a simple statistical feature for sequence classification

Abstract: The complexity of genomic sequence classification drives the development of computational methodologies that integrate biological knowledge with statistical and machine learning techniques, such as alignment-free feature extraction techniques. These techniques are well-established in comparing genomes that do not share an alignable set of common genes, because they assess sequence similarity without resorting to sequence alignment to overcome the limitations of well-established alignment-based methods. Among all existing alignment-free techniques, k-mers —subsequences of nucleotides of length k — have been widely used. However, k-mer approaches face limitations when classifying shorter sequences. To overcome these limitations, we developed the (m,n)-mer feature extraction technique, which is grounded in the concept of conditional probability. This technique defines the probability of a nucleotide subsequence occurring given that another subsequence has already occurred, thus providing a more informative analysis of sequence data. The (m,n)-mer feature addresses several hypotheses: i) conditional probabilities can provide greater informational content than non-conditional counterparts, ii) the order of nucleotides within a genomic sequence is non-random, and iii) there is a dependency relationship among the nucleotide bases in biological sequences. Our comparative analysis of (m,n)-mer and k-mer features utilized 11 distinct datasets for binary, multiclass, and clustering classifications. The (m,n)-mer method demonstrated significantly improved performance in supervised and unsupervised classification strategies, particularly regarding the diversity of viral and bacterial species. These findings highlight the transformative potential of (m,n)-mers in future sequence classification research, suggesting a promising new approach to accurately analyze complex genomic data.

Otávio J. B. Brustolini

Authors: Otávio J. B. Brustolini

Title: Predicting novel mosquito-associated viruses from metatranscriptomic dark matter

Abstract: The exponential growth of metatranscriptomic studies dedicated to arboviral surveillance in mosquitoes has yielded an unprecedented volume of unclassified sequences referred to as the virome dark matter. Mosquito-associated viruses are classified based on their host range into Mosquito-specific viruses (MSV) or Arboviruses. While MSV replication is restricted to mosquito cells, Arboviruses infect both mosquito vectors and vertebrate hosts. The identification of novel arboviruseshas become a priority due to global concerns about mosquito-borne disease outbreaks. To address the challenges in identifying emerging arboviruses amid vast amounts of unknown sequences, we developed the MosViR pipeline. This tool identifies complex genomic discriminatory patterns for predicting novel MSV or Arboviruses from viral contigs as short as 500 bp. The pipeline combines the predicted probability score from multiple predictive models, ensuring a robust classification with Area Under ROC (AUC) values exceeding 0.99 for test datasets. To assess the practical utility of MosViR in actual cases, we conducted a comprehensive analysis of 24 published mosquito metatranscriptomic datasets. By mining this

metatranscriptomic dark matter, we identified 605 novel mosquito-associated viruses, with eight putative novel Arboviruses exhibiting high probability scores. Our findings highlight the limitations of current homology-based identification methods and emphasize the potentially transformative impact of the MosViR pipeline in advancing the classification of mosquito-associated viruses. MosViR offers a powerful and highly accurate tool with transformative potential on future mosquito virome studies, enhancing our ability to capture the diverse genomic landscape of mosquito-associated viruses and repurpose large portions of unclassified sequences.

Samuel R. Torres

Authors: Samuel R. Torres, Raphael Saldanha, Rocio Zorrilla, Victor Ribeiro, Eduardo Pena, Fábio Porto

Title: Beyond Temporal Dimensions: Integrating KMeans-DTW and QuadTree Entropy for Advanced Multivariate Séries Insights

Abstract: The efficacy of machine learning models are contingent on input data quality and model selection itself. In this work we highlight the importance of data quality, particularly in identifying regions within the input space that ex- hibit similar behavior. Clustering is used to group similar data, and is explored for their potential to enhance model performance by identifying these regions. The aim is to provide insights into the effectiveness of using clus- tering to improve machine learning model performance.

Sergio Luque Mamani

Authors: Sergio Luque Mamani

Title: Forecasting dengue cases based on climatic variables using machine learning and deep learning models

Abstract: Dengue fever virus (DENV) is a moquisto-borne virus mostly spread by Aedes aegypti. According to the World Health Organization, about 100 – 400 million cases of the disease occur annually. In 2023 alone, there were over five million cases of dengue fever reported in more than 80 countries and territories worldwide. Brazil, Peru and Bolivia have reported the highest number of dengue cases in 2023. (Akinsulie and Idris, 2024). Hence, an early prediction of dengue continues to be a major concern for public health in countries with high prevalence of dengue. Creating a robust forecast model for the accurate prediction of dengue is a complex task and can be done through various data modelling approaches (Kakarla, 2023), (Nguyen, 2022), (Jain, 2019). In this work, we have applied Machine Learning and Deep Learning models to predict the dengue prevalence in districts of the department of Piura in Peru, considered priority regions. We consider the number of dengue cases as the target variable and weather variables such as mean temperature, minimum temperature, maximum temperature, relative humidity, precipitation and diurnal temperature variation. To understand the lagged impact of climate factors on dengue transmission, a cross-correlation analysis was conducted and the appropriate lag time for each meteorological factor was determined between 0 and 12 months.

Similarly, autocorrelation analysis was performed to assess the impact of past dengue cases on current dengue cases. The model performance and forecast accuracy were evaluated using MAE, MSE, RMSE and AIC. The Random Forest and LSTM models showed superior predictive performance in detecting dengue outbreaks and the peak of cases compared to other models (Figure 1), which were able to identify trends in dengue cases but did not successfully predict the timing of outbreaks. Also, some variables were shown to have a minor influence on the increase in dengue cases. Thus, it was observed that climatic variables such as temperature, humidity and precipitation are closely related to the occurrence of dengue cases (Figure 2). These predictions will provide valuable insights for public health officials to manage dengue outbreaks more effectively.

Thiago Moeda

Authors: Moeda, T.; Benyosef, L. and Camacho, E.

Title: Prediction of magnetic activities for solar cycle 25 using neural networks

Abstract: Magnetic storms are events related to the interaction of the solar wind with the Earth's magnetosphere. The effects caused by these phenomena are perceived on the Earth's surface. In practice, the measurement used to evaluate magnetic storms is obtained through the average values of the horizontal component of four observatories close to the Earth's magnetic equator. This method is known as the disturbed storm time index (Dst index). Since the 1990s, research in Artificial Intelligence has been improving as a viable solution to practical operational problems related to Space Weather. The long historical series and the diversity of observational data of the solar wind from satellites located at the first Earth-Sun Lagrangian point provides a good opportunity for the development of modern Data-driven computational methodologies. Thus, we conducted a study for the prediction of geomagnetic storms, specifically the prediction of the Dst index, using recurrent neural networks, suitable for time-dependent multivariate data sets. For this purpose, data from NASA's OMNI satellite, with hourly temporal resolution, the provisional Kyoto Dst index, magnetic, electrical and plasma variables were used. The modeling process for this work was carried out in two stages. Initially, we trained and tested the models for cycles 23 and 24. Then, for cycle 25, we evaluated their performance in periods containing moderate and intense magnetic storms. Predictions were made for the sequence of moderate events that occurred in the first half of 2023, in which the magnetic storms induced disturbances in the atmospheric density, causing the destruction of dozens of low-orbit satellites due to atmospheric drag. Additionally, we analyzed and compared the prediction of the intense magnetic storm of May 2024. The results obtained are promising and focus on detecting the amplitude of the sudden impulse that, if it exists, occurs before the main phase of a magnetic storm, whose amplitude variation of the Dst index exceeds 400 nT.

Vinicius Kreischer

Authors:

Title: Gypscie: ML Lifecycle Management with Artifact Tracking and Multi-Environment Execution

Abstract: Managing the machine learning (ML) lifecycle is increasingly important for organizations utilizing data-driven models to guide critical decisions. Key aspects include traceability (keeping detailed records of model versions and performance over time), provenance (maintaining clear histories of data sources and model evolution), and sharing (facilitating collaboration and knowledge transfer across teams and projects). Moreover, easy access to diverse execution environments is essential to efficiently develop, test, and deploy models across different stages. Gypscie is a ML management platform that provides access to artifacts and their related metadata, as well as ML services, like model training, prediction, and data transformation. It also includes a dataflow language that allows users to define ML tasks as sequences of independent activities, streamlining the orchestration and customization of workflows. Additionally, Gypscie supports execution across heterogeneous environments, including dedicated servers, supercomputers, clusters, and cloud infrastructure, providing flexible and scalable options for managing complex ML tasks. Gypscie has been applied across different domains, demonstrating its versatility. Notable examples include Metereology, where it has been used to predict and analyze extreme rainfall events, and Mooring Line Integrity Monitoring, where it predicts potential structural failures in offshore platforms, decreasing repairing costs.

Victor de Paula Dornellas Ribeiro

Authors:

Title: FastCORE: computational-efficient coreset selection

Abstract: The machine learning literature traditionally assumes that the accuracy of a model increases as more information is provided for its training. The intuition is that a dataset contains samples of a given phenomenon, and that we are able to better detail its distribution as we accumulate measurements. The literature also presents evidence linking the increase in training time to the volume of data. To overcome hardware and time limitations, we can apply training example selection techniques – called coresets. Originally, a coreset is the smallest set of points that best approximates a surface, preserving the maximum amount of information possible. In the context of machine learning, a coreset refers to the set of techniques aimed at selecting a subset of data without compromising the quality of the model's training.

A literatura especializada em aprendizado de máquina tradicionalmente assume que a precisão de um modelo aumenta na medida que fornecemos mais informação para seu treinamento. A intuição é que um conjunto de dados contém amostras sobre um determinado fenômeno, e que somos capazes de detalhar melhor sua distribuição na medida que acumulamos medições. A literatura apresenta ainda evidências que relacionam o aumento do tempo de treinamento de maneira proporcional ao volume de dados. Para contornar as limitações de hardware e tempo podemos aplicar técnicas de seleção de exemplos de treinamento – *coresets*. Originalmente um *coreset* é o menor conjunto de pontos que melhor aproxima uma superfície, preservando a maior quantidade de informação possível. No contexto de aprendizado de máquina, *coreset* é o

conjunto de técnicas que busca selecionar um subconjunto de dados, sem que a qualidade do treinamento de um modelo seja comprometida.