

KAUST ACADEMY

SECTION 2

BIOINFORMATICS.



12

FROM CNA PROFILES TO PROGNOSTIC PREDICTIONS: UTILIZING HILBERT CURVE IMAGES FOR CANCER SUBTYPE CLASSIFICATION AND SURVIVAL ANALYSIS

YAHYA ALHABBOUB

Cancer results from DNA mutations, including Copy Number Alterations (CNAs) where large DNA segments are duplicated or deleted. This study examines the potential of CNAs as biomarkers for cancer subtype classification, tumor stage prediction, and survival analysis, using nearly 10,000 samples across 25 cancer types.We employed Hilbert Curve visualizations of CNA profiles and applied computer vision models, including YOLOv8 and VGG16, for classification tasks. A logistic regression model using patient cytoband data served as a control. YOLOv8 performed best in cancer subtype classification, surpassing VGG16 and the control model, although tumor stage classification yielded poor results across all models.

Survival analysis with DeepSurv showed fluctuating C-index results, indicating issues with overfitting.Despite these challenges, the study successfully explored the utility of CNAs in cancer research. Some subtypes were classified effectively, suggesting CNAs provide valuable insights into tumor characteristics. Key limitations included dataset size and the mismatch between CNA image simplicity and model complexity. Future work will expand the dataset and develop models tailored to CNA image analysis, improving interpretability and performance.

From CNA Profiles to Prognostic Predictions: Utilizing Hilbert Curve Images for Cancer Subtype Classification and Survival Analysis

Yahya Alhabboub 1,4 , Masroor Bayati 1,2, Jüri Reimand 1,2,3

yalhabboub@gmail.com, juri.reimand@utoronto.ca

Abstract

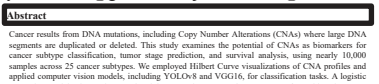
Cancer results from DNA mutations, including Copy Number Alterations (CNAs) where large DNA segments are duplicated or deleted. This study examines the potential of CNAs as biomarkers for cancer subtype classification, tumor stage prediction, and survival analysis, using nearly 10,000 samples across 25 cancer subtypes. We employed Hilbert Curve visualizations of CNA profiles and applied computer vision models, including YOLOv8 and VGG16, for classification tasks. A logistic regression model using patient cytoband data served as a control. YOLOv8 performed best in cancer subtype classification, surpassing VGG16 and the control model, although tumor stage classification yielded poor results across across all models. . Survival analysis with DeepSurv showed fluctuating C-index results, indicating issues with overfitting. Despite these challenges, the study successfully explored the utility of CNAs in cancer research. Some subtypes were classified effectively, suggesting CNAs provide valuable insights into tumor characteristics.

Methodology

Dataset

We utilized the Copy Number Alterations (CNA) dataset from the TCGA - PanCanAtlas Publications collection [1]. This dataset was employed to generate CNA images and calculate cytoband values for each patient.

Preprocessing



Modeling Table

Below is a table that summarizes the key details of each model used:

Model	Task(s)	Dataset	Hyperparameters	Activation Function	Loss Function	Other Notes
YOLOv8n-cls	Cancer subtype classification	CNA Images	Epochs=150, Imgsz=480, Lr=0.01, Batch=16, Optimizer=SGD	N/A	N/A	Focuses on subtype classification
VGG16 Multi-task	Cancer subtype & tumor stage classification	CNA Images	Lr=0.0001, Optimizer=Adam, Epochs=300, Imgsz=224, Batch=32	LeakyReLU, Softmax	CrossEntropyLoss	Multi-task model combining subtype and stage tasks
Logistic Regression	Cancer subtype classification (Control)	Cytoband values	Max_iter=800, Solver='liblinear', Multi_class='ovr', Class_weight='balanced'	N/A	N/A	Control model using cytobands for comparison
DeepSurv	Survival analysis	Cytoband values	Optimizer=Adam, Lr=0.000001, Epochs=200, lr_decay_rate=0.001	SELU, Linear	NegativeLogLikelihood	Uses modified pre-existing code for survival analysis.[2]

Post-processing

Classification tasks :

- Precision.
- Recall.
- F1-score.

Survival analysis tasks:

- C-index.



Sample of a CNA image



Sample of a CNA image

Conclusion

Primary Objective Explored the utility of CNAs in studying tumor characteristics, with some subtypes showing excellent performance, highlighting their potential in cancer research.

Challenges:

- Limited dataset (10,000 samples) made addressing multiple tasks challenging.
- Data augmentation was not feasible due to the reliance on specific image features.
- Pre-trained models may have been too complex for simple CNA images.

Discussion and Future Work

Research Foundation: The study provides a basis for further CNA exploration.

Next Steps:

- Expand the dataset and develop CNA-specific models.
- Include predictions for additional patient characteristics (age, gender).

Acknowledgements & Citations

The From CNA Profiles to Prognostic Predictions: Utilizing Hilbert Curve Images for Cancer Subtype Classification and Survival Analysis is supported by the Data Sciences Institute, University of Toronto. KAUST Academy is financially supporting Yahya Alhabboub.

[1] Zack, T., Schumacher, S., Carter, S. et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet 45, 1134–1140 (2013). https://doi.org/10.1038/ng.2760.

[2] C. Zifan, "DeepSurv.pytorch," GitHub, 10-Apr-2020. [Online]. Available: https://github.com/czifan/DeepSurv.pytorch/tree/master. [Accessed: 10-Aug-2024].

Affiliations

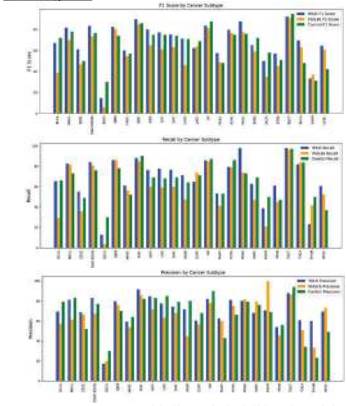
[1] Computational Biology Program, Ontario Institute for Cancer Research. [2] Department of Medical Biophysics, University of Toronto. [3] Department of Molecular Genetics, University of Toronto. [4] KAUST Academy

Objectives

- **Exploring CNAs:** Investigate their role in early cancer diagnosis, treatment, and tumor characterization, including subtype identification, staging, and resistance.
- **AI Model Development:** Develop AI models to distinguish between cancerous and normal mutations.
- **1D to 2D Conversion:** Transform CNA sequences into two-dimensional formats for better analysis, applying computer vision techniques like transfer learning for enhanced interpretation.
- **Model Comparison:** Compare single-task vs. multi-task models in cancer classification tasks.
- **CV vs. ML Models:** Evaluate complex CV models against simpler ML models like logistic regression using patient cytoband data.
- **Survival Analysis:** Use CNAs in survival analysis through the DeepSurv framework.

Results

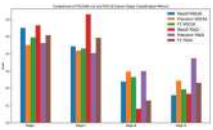
Cancer Subtypes Task



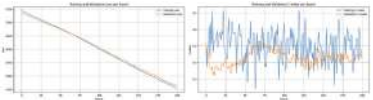
- **Performance Variation:** F1-scores varied, with TGCT showing the highest and ESCA the lowest.
- **Model Comparison:** YOLOv8n-cls performed best, followed by the control model, then VGG16.
- **Metric Consistency:** Precision, recall, and F1-scores were generally aligned.

Tumor Stage Task

Overall Performance:
The classification results for tumor stages are uniformly poor across all stages.



Survival Task



Model Stability: Although the loss function shows ideal convergence across epochs, the fluctuating C-index in both training and validation sets suggests issues with model overfitting or underfitting.

13

DEVELOPMENT OF DEEP LEARNING SOLUTIONS FOR 4D IMAGE SEGMENTATION OF SPATIAL OMICS DATA

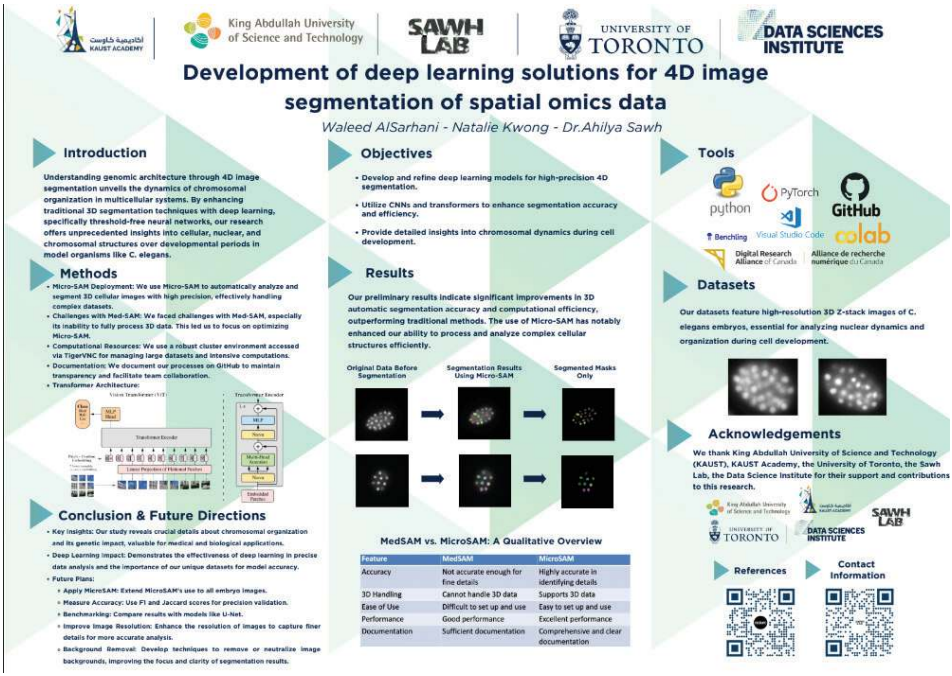


WALEED ALSARHANI

Understanding the intricate dynamics of chromosomal organization within multicellular systems is critical for advancements in medical and biological sciences. Our research focuses on the development and deployment of high-precision deep learning models for 4D image segmentation, aimed at unveiling genomic architecture over developmental periods in model organisms such as C. elegans. By integrating advanced convolutional neural networks (CNNs) and transformers, we significantly improve the accuracy and efficiency of segmentation tasks, particularly through the utilization of Micro-SAM.

In our study, we compare the performance of Micro-SAM with traditional segmentation methods and alternative deep learning models like Med-SAM. Preliminary results demonstrate substantial enhancements in 3D segmentation accuracy and computational efficiency with Micro-SAM, surpassing traditional techniques. These improvements facilitate a deeper understanding of cellular, nuclear, and chromosomal structures, providing valuable insights into their development.

Our findings underscore the impact of deep learning in precise data analysis, highlighting the importance of high-resolution datasets for accurate model performance.


KAUST ACADEMY
King Abdullah University of Science and Technology
SAWH LAB
UNIVERSITY OF TORONTO
DATA SCIENCES INSTITUTE
Development of deep learning solutions for 4D image segmentation of spatial omics data
Waleed AlSarhani - Natalie Kwong - Dr.Ahilya Sawh
Introduction
Understanding genomic architecture through 4D image segmentation unveils the dynamics of chromosomal organization in multicellular systems. By enhancing traditional 3D segmentation techniques with deep learning, specifically threshold-free neural networks, our research offers unprecedented insights into cellular, nuclear, and chromosomal structures over developmental periods in model organisms like C. elegans.
Methods
Micro-SAM Deployment: We use Micro-SAM to automatically analyze and segment 3D cellular images with high precision, effectively handling complex datasets.
Challenges with Med-SAM: We faced challenges with Med-SAM, especially its inability to fully process 3D data. This led us to focus on optimizing Micro-SAM.
Computational Resources: We use a robust cluster environment accessed via TigerVNC for managing large datasets and intensive computations.
Documentation: We document our processes on GitHub to maintain transparency and facilitate team collaboration.
Transformer Architecture:
Conclusion & Future Directions
Key Insights: Our study reveals crucial details about chromosomal organization and its genetic impact, valuable for medical and biological applications.
Deep Learning Impact: Demonstrates the effectiveness of deep learning in precise data analysis and the importance of our unique datasets for model accuracy.
Future Plans:
Apply MicroSAM: Extend MicroSAM's use to all embryo images.
Measure Accuracy: Use F1 and Jaccard scores for precision validation.
Benchmarking: Compare results with models like U-Net.
Improve Image Resolution: Enhance the resolution of images to capture finer details for more accurate analysis.
Background Removal: Develop techniques to remove or neutralize image backgrounds, improving the focus and clarity of segmentation results.
Objectives
Develop and refine deep learning models for high-precision 4D segmentation.
Utilize CNNs and transformers to enhance segmentation accuracy and efficiency.
Provide detailed insights into chromosomal dynamics during cell development.
Results
Our preliminary results indicate significant improvements in 3D automatic segmentation accuracy and computational efficiency, outperforming traditional methods. The use of Micro-SAM has notably enhanced our ability to process and analyze complex cellular structures efficiently.
Original Data Before Segmentation
Segmentation Results Using Micro-SAM
Segmented Masks Only
MedSAM vs. MicroSAM: A Qualitative Overview
Feature | MedSAM | MicroSAM
Accuracy | Not accurate enough for fine details | Highly accurate in identifying details
3D Handling | Cannot handle 3D data | Supports 3D data
Ease of Use | Difficult to set up and use | Easy to set up and use
Performance | Good performance | Excellent performance
Documentation | Sufficient documentation | Comprehensive and clear documentation
Tools
python
PyTorch
GitHub
Benchling
Visual Studio Code
colab
Digital Research Alliance of Canada
Alliance de recherche numérique du Canada
Datasets
Our datasets feature high-resolution 3D Z-stack images of C. elegans embryos, essential for analyzing nuclear dynamics and organization during cell development.
Acknowledgements
We thank King Abdullah University of Science and Technology (KAUST), KAUST Academy, the University of Toronto, the Sawh Lab, the Data Science Institute for their support and contributions to this research.
References
Contact Information

14

YOLO-AR: A CT PROCESS ANALYTICS FRAMEWORK

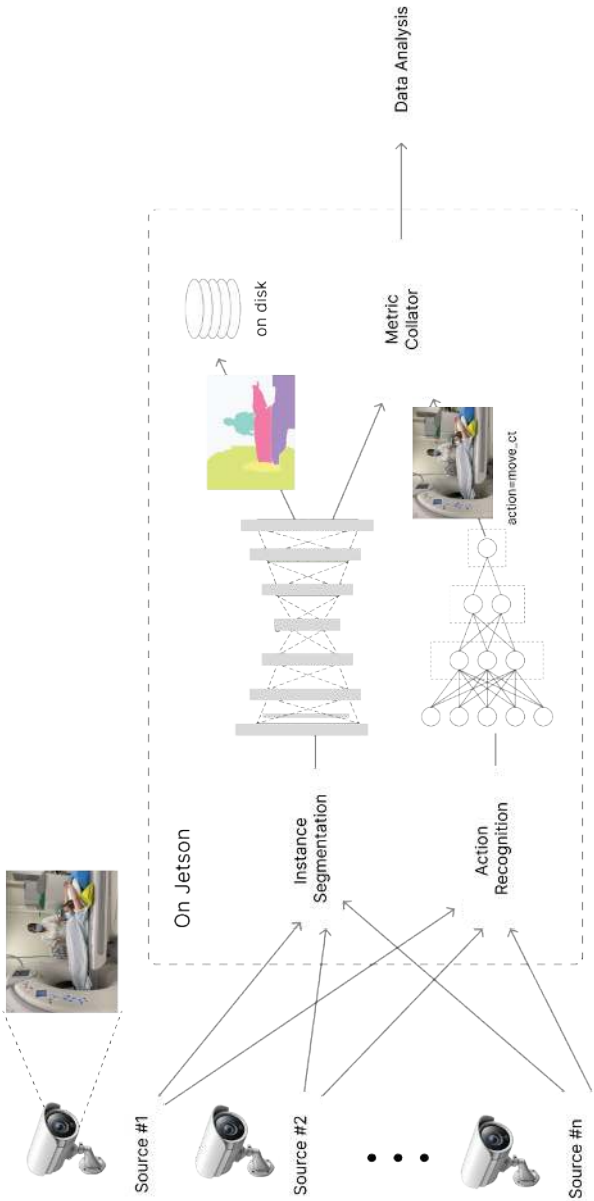


WAEL ALI SULAIS

Diagnostic medical imaging services require significant resources, including equipment, materials and labor. Proper coordination of these resources positively impacts the productivity of the radiology department, which, in turn, influences access to care, costs, and quality. Thus, the availability of near-real-time productivity measurements may improve department operations by providing insight into system bottlenecks. We propose a novel Operations Research (OR) framework for measuring the efficiency of a hospital's medical imaging department. Utilizing a mix of AI methods such as image segmentation and action recognition, we provide a state-of-the-art (SOTA) system for tracking the end-to-end process latency and patient inventory while preserving privacy.

More specifically, we explore YOLOv8 for image segmentation as well as ResNet18 for action recognition. As a part of this endevor, we have collected footage of simulated patient computer tomography (CT) scans for training.

Furthermore, the system has been deployed on a Nvidia Jetson Orin NX in a teaching hospital. Henceforth, We believe that our system can be applied to more hospitals in Greater Toronto Area (GTA), generalized to more types of medical imaging procedures, as well as extended to produce more fine-grained analytics.


Data Analysis
on disk
Metric
Collator
action=move_ct
On Jetson
Instance
Segmentation
Action
Recognition
Source #1
Source #2
Source #n

15

GWAS OF CYSTIC FIBROSIS LUNG DISEASE USING A VARIABLE NUMBER OF TANDEM REPEATS CALLED FROM LONG-READ SEQUENCING



SULIMAN ALI ALRAQEBAH

Cystic fibrosis (CF) is a genetic disease affecting 70,000-100,000 individuals worldwide.1 It is caused by variations in the CFTR gene, impacting multiple organs, particularly the lungs.

A recent genome-wide association study (GWAS) identified possible genetic links to lung function in CF, including a region near the SLC9A3 gene on chromosome 5.2. This region contains repetitive genetic sequences, and whether these repeats contribute to the observed genetic associations is unclear.2 This study aims to use HiFi technology from Pacific Biosciences (PacBio) to explore the association between repeat lengths genome-wide, particularly in the SLC9A3 region, and lung function in CF patients.

Subjects with CF were recruited into the Canadian Cystic Fibrosis Gene Modifier Study (CGMS). High-molecular-weight DNA was sequenced using PacBio whole-genome long-read HiFi sequencing (30X). The Tandem Repeat Genotyping Tool (TRGT)3 and its repeat catalog were used to genotype repeats. The subjects' ancestry was inferred, and significant PCs from PCA were included as covariates.

Figure 3

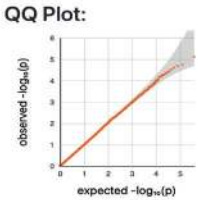
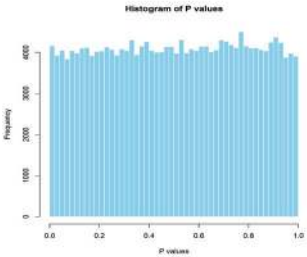


Figure 4



Figures

Figure 1

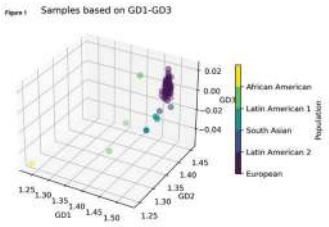
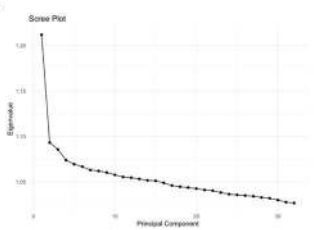
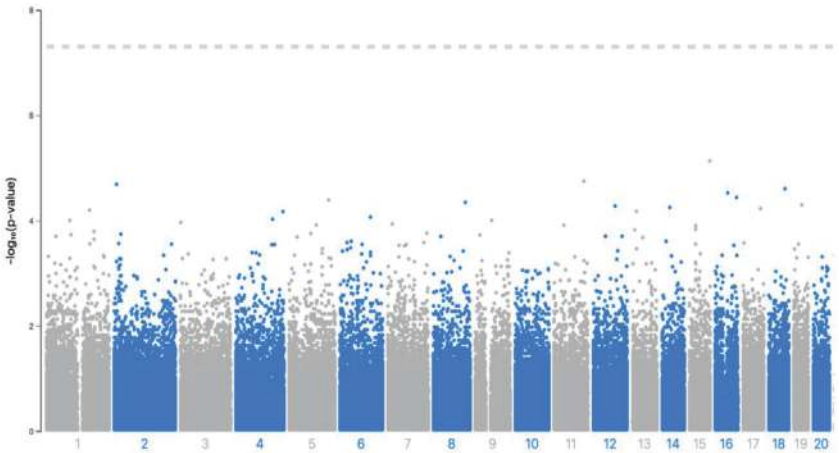


Figure 2





16

DATA SCIENCE TO OPTIMIZE CARDIOVASCULAR DISEASE PREVENTION



SULAIMAN MOUSA ALALWANI

Diabetes and obesity are significant risk factors for cardiovascular disease; however, recent advancements in medications, particularly Glucagon-Like Peptide-1 (GLP-1) agonists and similar drugs, have proven highly effective in protecting patients with these conditions from cardiovascular events.

Our project focuses on investigating the impact of this class of medications on patients with obesity, diabetes, and other comorbidities using electronic health record data.

We apply statistical methodologies, specifically Target Trial Emulation, to simulate randomized clinical trials and analyze real-world data.

Unraveling the Understudied Health Effects of GLP-1 Receptor Agonists: A Nationwide Cohort Study

STUDENTS | SULAIMAN ALALWANI, WENXIN GUO, JIWON SHIN, MAIDOU WANG
SUPERVISORS | DR. JAY LUSK, DR. FAN LI, DR. BRIAN MAC GRORY, AYUSH GUPTA

Background

Glucagon-like peptide-1 (GLP-1) drugs are new and effective treatments for **type II diabetes and obesity**, with additional benefits for cardiovascular health, brain, and kidneys.

Using U.S. national electronic health record (eHR) data provided by Truveta, we compare the safety and efficacy of GLP-1 drugs with traditional medications. We aim to find the best ways to distribute GLP-1 drugs, targeting patients who would benefit the most.

Objectives

Using Truveta's eHR data, we aimed to address the following research questions:

- How do GLP-1 receptor agonists affect - **alleviate or worsen** - obesity, type II diabetes, and cardiovascular and brain health compared to other available medications?
- If there are effects, whether positive or negative, how do they **vary across populations** based on age, race, sex, and baseline health conditions?

Methodology

STEP 1 DATA COLLECTION & FEATURING

Our data come from Truveta, which provides comprehensive health records on **over 100 million patients** across the United States.

From this population, we selected a subset for our study.

Our final study population has the following key features:



STEP 2 TARGET TRIAL EMULATION

A randomized controlled trial randomly assigns patients to the treatment or control group, and collect their health outcomes at the end of a follow-up period. Using patients' existing medical histories, we aim to **emulate** a randomized controlled trial.



As illustrated above, our emulation involves the following steps:

- **Assign an index date** to each patient, defined as the first time they take the drug of interest while meeting all eligibility criteria.
- From the year leading up to the index date, **collect covariates** for each patient
- From the first year after the index date, **collect outcomes** for each patient

* **Outcomes and covariates: patient demographics and relevant health parameters**
* **Eligibility: patients must be non-pregnant, have type II diabetes or obesity, and >1 year of medical data**

STEP 3 STATISTICAL ANALYSIS (ONGOING)

Overlap Weighting

To approximate randomization, we use overlap weighting (Li, Morgan, & Zaslavsky, 2017) to balance baseline covariates between the treatment and control groups. This is achieved through assigning a weight w to each patient's outcome, calculated as follows:

For patients in the treatment group: $w = 1 - e(\text{Covariates})$
For patients in the control group: $w = e(\text{Covariates})$, where

$e(\text{Covariates}) = Pr(\text{Assigned to Treatment group} \mid \text{Covariates}) = logit^{(-1)}(\beta * \text{Covariates})$

Treatment Effect Estimation

For **time-to-event outcomes**, we use a weighted Cox regression model to estimate hazard ratios.

For **other outcomes**, we calculate the Average Treatment Effect for the Overlap Population (ATO).

Additionally, we conduct **subgroup analysis** by age, race, sex, and disease severity to estimate how treatment effects vary by these factors.

Sensitivity Analysis

To reinforce our study's robustness, we conduct sensitivity analyses by testing minor variations to our methodology, such as varying follow-up periods (e.g., 6 months or 2 yrs).

Future Work

Data Processing | Extract covariates, outcomes, and baseline features for each patient. Handle missing data through imputation or removal to ensure dataset completeness.

Descriptive Analysis | Generate tables to visualize data, including summary statistics and pre-adjustment covariate balance. Check for multicollinearity before modeling.

Statistical Analysis | Perform statistical analysis according to our analysis plan.

Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing Covariates via Propensity Score Weighting. Journal of the American Statistical Association, 113(521), 390–400. https://doi.org/10.1080/01621459.2016.1260466

17

THE IMPACT OF CO-OCCURRING CHRONIC PAIN AND DEPRESSION/ANXIETY IN THE US FROM 2019 - 2021



SHOUQ ALSULAMI

Chronic pain often co-occurring with Anxiety/Depression (A/D) and vise versa. This study examines NHIS data from 2019-2021 to assess their prevalence and impact among US adults. We analyzed data from 93,047 adults who self-reported chronic pain and A/D. Participants were grouped by their conditions. We found that 20.4 million US adult experienced both chronic pain and A/D, this group of people shows higher rates of social and functional limitations, poor health reporting, and hospital visits. In conclusion, chronic pain and A/D together are associated with greater health challenges, highlighting the need for integrated treatment strategies.

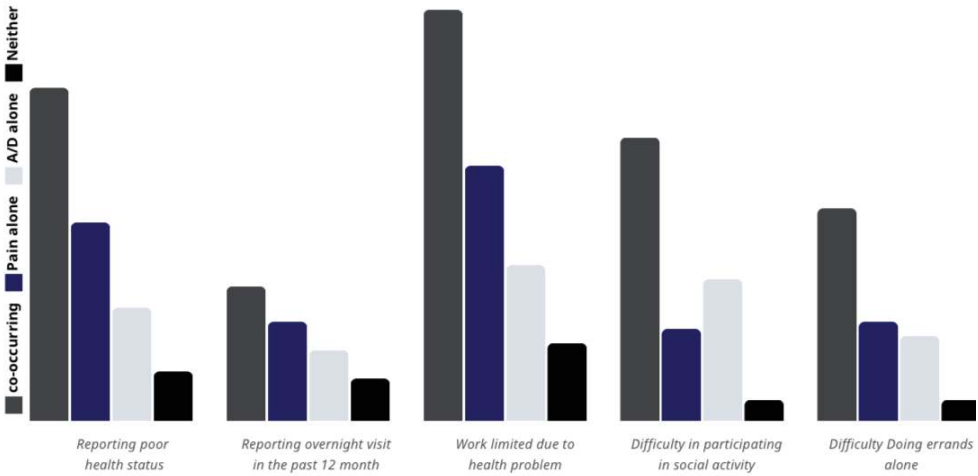


Figure3: Prevalence of Health outcome, Overnight Hospital Visits, and Functional Limitations Among U.S. Adults by Chronic Pain and A/D Status, 2019-2021

18

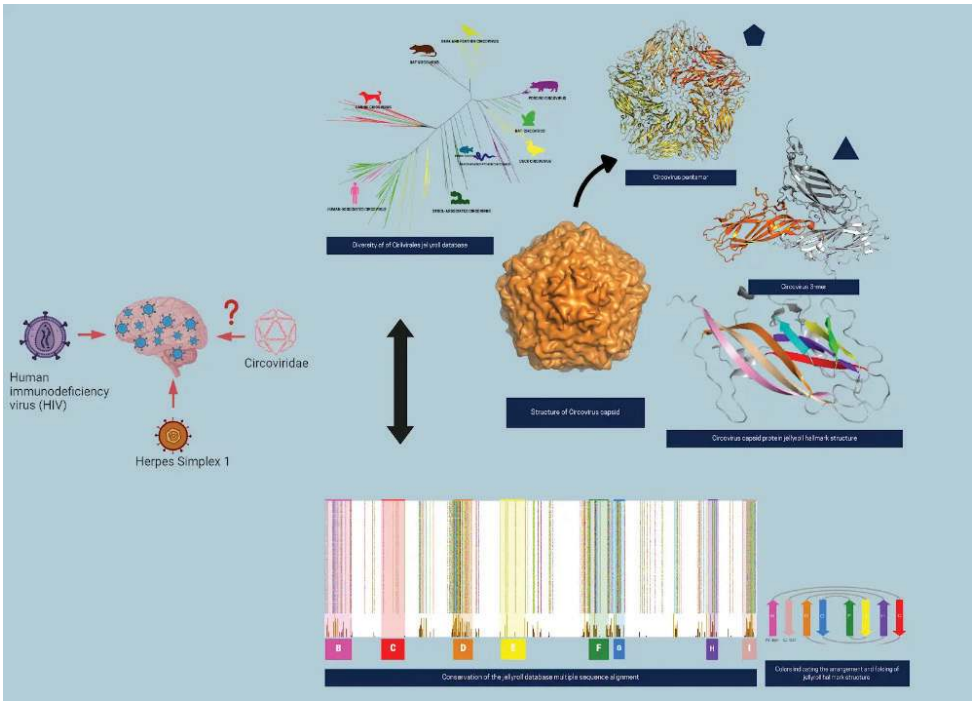
THE SEARCH FOR NOVEL DNA **VIRUSES** IN ALZHEIMER'S DISEASE BRAINS



RAKAN HADI
RIFAN ALSALLUM

This project explores the potential viral origins of Alzheimer's disease, focusing on discovering novel viruses within the Circoviridae family.

By constructing a comprehensive database of a conserved protein sequence known as the Jelly Roll hallmark structure, the research aims to uncover new insights into how these viruses might contribute to Alzheimer's pathology using bioinformatics tools.


Human immunodeficiency virus (HIV)
Herpes Simplex 1
Circoviridae
Diversity of of Cirlivirales jelly roll database
Circovirus pentamer
Circovirus 3-mer
Structure of Circovirus capsid
Circovirus capsid protein jelly roll helmark structure
B
C
D
F
G
H
I
Conservation of the jelly roll database multiple sequence alignment.

19

QUANTIFYING THE EFFECT OF JOINT CALLING

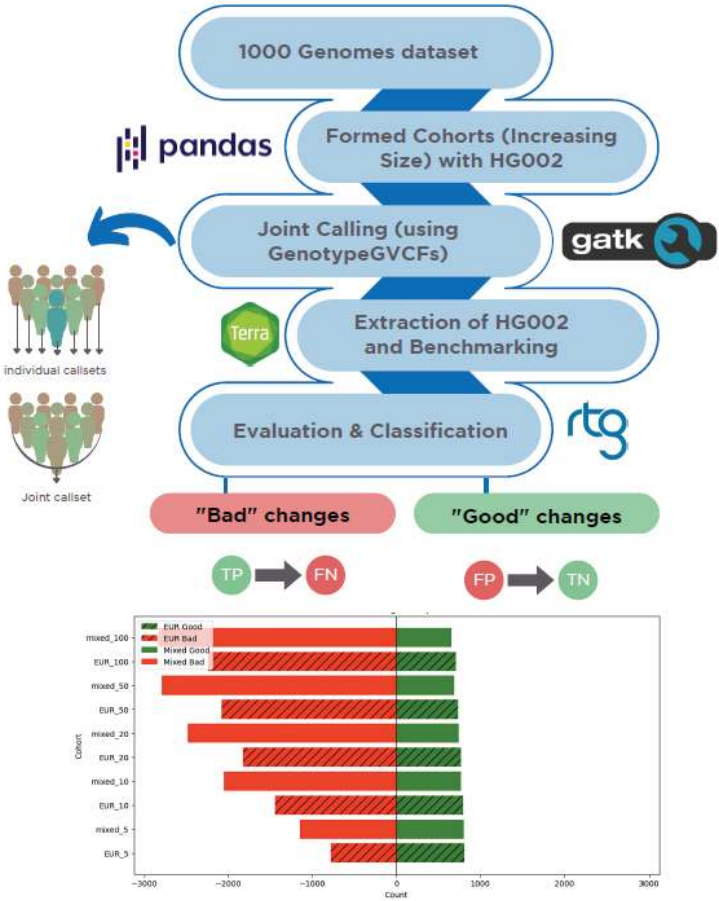


RAGHAD AHMED ALKHAROUBY

Joint variant calling improves the accuracy of variant detection by analyzing genetic data across multiple samples simultaneously. It leverages population statistics and biological assumptions about inheritance to enhance variant detection accuracy. A previous experiment measured the benefits of joint calling and demonstrated that increasing the size of the cohort improves variant sensitivity, with benefits plateauing around 600 samples (Gauthier et al., 2016, ASHG).

The goal of this research was to reproduce this experiment using modern genetic data, measure the benefits of joint calling, and identify the minimum sample size required to achieve optimal performance. Additionally, we explored how varying population structures within the cohort affect the accuracy of joint variant calling.

We analyzed genotypic changes by creating cohorts of increasing size from the 1000 Genomes dataset, ensuring the inclusion of the well-established HG002 sample in each cohort. We performed joint calling using the GenotypeGVCFs tool. The HG002 sample was then extracted from the joint VCF files and benchmarked using VCFeval.


1000 Genomes dataset
pandas
Formed Cohorts (Increasing Size) with HG002
Joint Calling (using GenotypeGVCFs)
gatk
individual callsets
Terra
Extraction of HG002 and Benchmarking
Joint callset
Evaluation & Classification
rtg
"Bad" changes
"Good" changes
TP
FN
FP
TN
EUR Good
EUR Bad
Mixed Good
Mixed Bad
mixed_100
EUR_100
mixed_50
EUR_50
mixed_20
EUR_20
mixed_10
EUR_10
mixed_5
EUR_5
Cohort
−3000
−2000
−1000
0
1000
2000
3000
Count

ALL OF US ANCESTRY
CONCORDANCE ACROSS TOOL



OMAR OSAMA RAWAS

During my internship at the Broad Institute of MIT and Harvard, I worked on evaluating the concordance between two ancestry prediction tools within the All of Us (AoU) dataset, which comprises over 410,000 samples out of a projected 1 million. The project compared categorical ancestry predictions from a Random Forest model with admixture estimates generated by the Rye tool. The primary goal was to assess the reliability of these methods in accurately predicting ancestry, particularly for diverse and underrepresented populations. This analysis is crucial for accurately determining population-specific allele frequencies, which have significant implications for disease diagnosis and understanding genetic risk factors. By validating these methods, we aimed to improve the precision of genetic analyses in diverse populations.

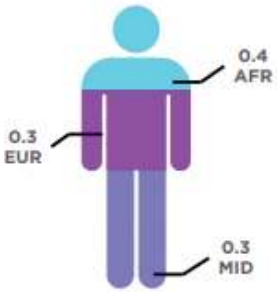
This project checked the reliability of the 2 methods of inferring predicted genetic ancestry

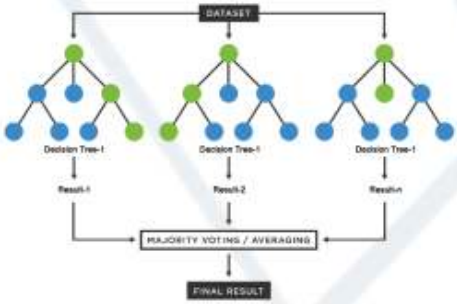
Admixture Estimates

Obtained using *Rye tool*, which provides accurate ancestry estimates by analyzing large genomic datasets, offering percentage-based ancestry proportions for each sample

Categorical Ancestry Prediction

Obtained using a *Random Forest* model trained on 1KG+HGDP data. The model classified samples into 6 ancestry groups (AFR, EUR, MID, EAS, SAS, AMR)





Concordance Evaluation Methods

- Cosine Similarity
- Cohen's Kappa
- Jaccard's Index
- Correlation
- F1 Score

21

INVESTIGATING THE ROLE OF THE X-CHROMOSOME IN AUTOIMMUNE DISEASES: A BIOINFORMATICS APPROACH LEVERAGING SINGLE-CELL GENOMICS

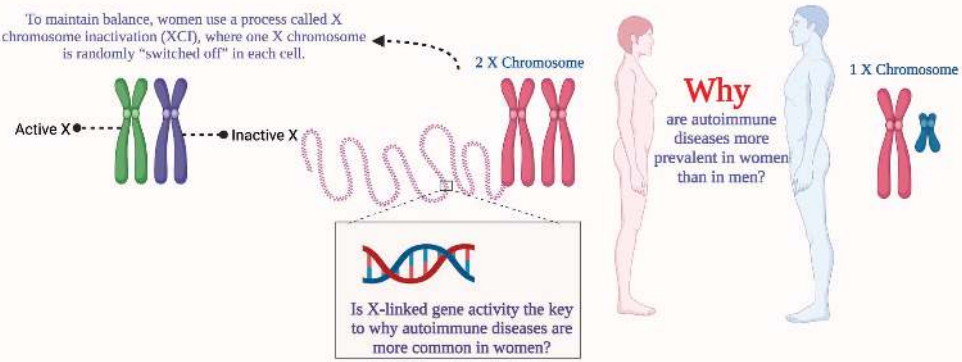


OLA AALI ALYAZIDI

Autoimmune diseases predominantly affect women, suggesting a potential role for the X chromosome in disease susceptibility. This study investigates the involvement of the X chromosome in autoimmune diseases using single-cell RNA sequencing data.

By analyzing gene expression patterns across immune cells from patients with Crohn's disease and type 1 diabetes, we identified differentially expressed genes (DEGs) located on the X chromosome. Our findings reveal a notable proportion of DEGs associated with the X chromosome, underscoring its potential contribution to the pathogenesis of these conditions.

This research provides new insights into the molecular mechanisms underlying autoimmune diseases and highlights the need for further exploration of X-linked genetic factors.



In this study, we aim to investigate the role of the X chromosome in the development and progression of autoimmune diseases. We will begin by analyzing gene expression differences between healthy and diseased samples, with a particular focus on X-linked genes. Additionally, we will compare gene expression patterns across various autoimmune diseases to identify key discrepancies and measure the rate of X chromosome inactivation (XCI) to assess its influence on disease outcomes.

22

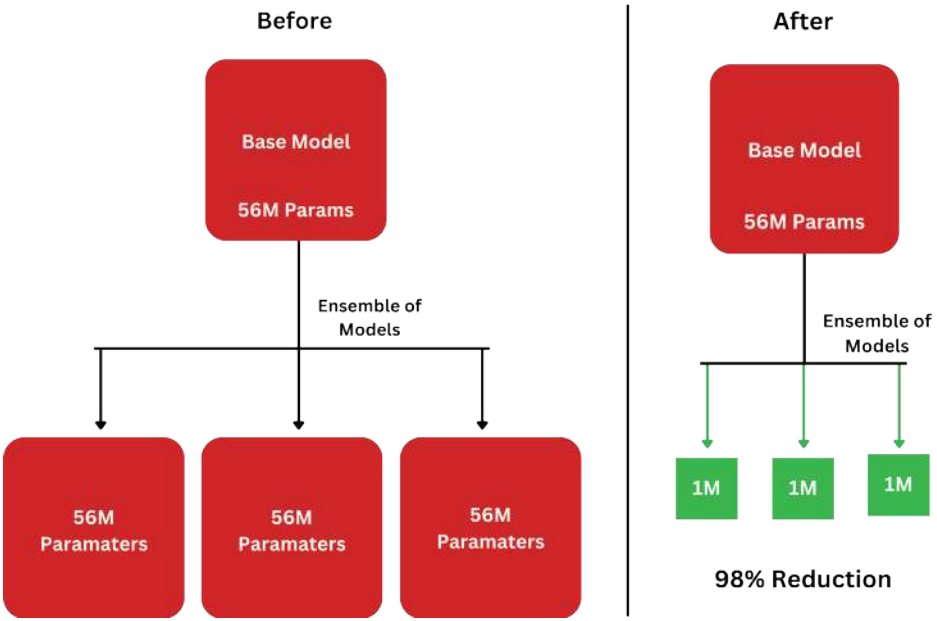
ENHANCING HARMFUL COVARIATE SHIFT DETECTION IN MEDICAL IMAGING WITH LORA



MOHAMMED ALJUDAIBI

Detecting harmful covariate shifts (HCS) is crucial for the safe deployment of machine learning models and to ensure their reliability, especially in high-risk domains like healthcare. HCS occurs when the data distribution in deployment differs from the training data, leading to degraded model performance. Existing methods, such as the Detectron, detect HCS with an ensemble of classifiers trained to agree on training data and disagree on a small sample of deployment data. However, this task becomes more challenging with medical imaging data as it would require significant compute and time. By integrating Low-Rank Adaptation (LoRA) into Detectron, we apply LoRA to the base model architecture, allowing each classifier in the ensemble to benefit from the reduced parameter count and improved efficiency. This significantly reduces the size of all of the classifiers in the ensemble of models, making detecting HCS on medical data more feasible and efficient.

This method makes the detection of HCS in medical imaging using the Detectron more feasible, and also ensures the robustness of deployed models, providing a reliable mechanism to ensure and maintain model performance in healthcare applications.


Before
Base Model
56M Params
Ensemble of Models
56M Paramaters
56M Paramaters
56M Paramaters
After
Base Model
56M Params
Ensemble of Models
1M
1M
1M
98% Reduction

23

EXPLORING THE IMPACT OF ENVIRONMENTAL AND CLIMATE EXPOSURES AND SOCIAL DETERMINANTS OF HEALTH (SDOH) ON MULTIPLE SCLEROSIS (MS)



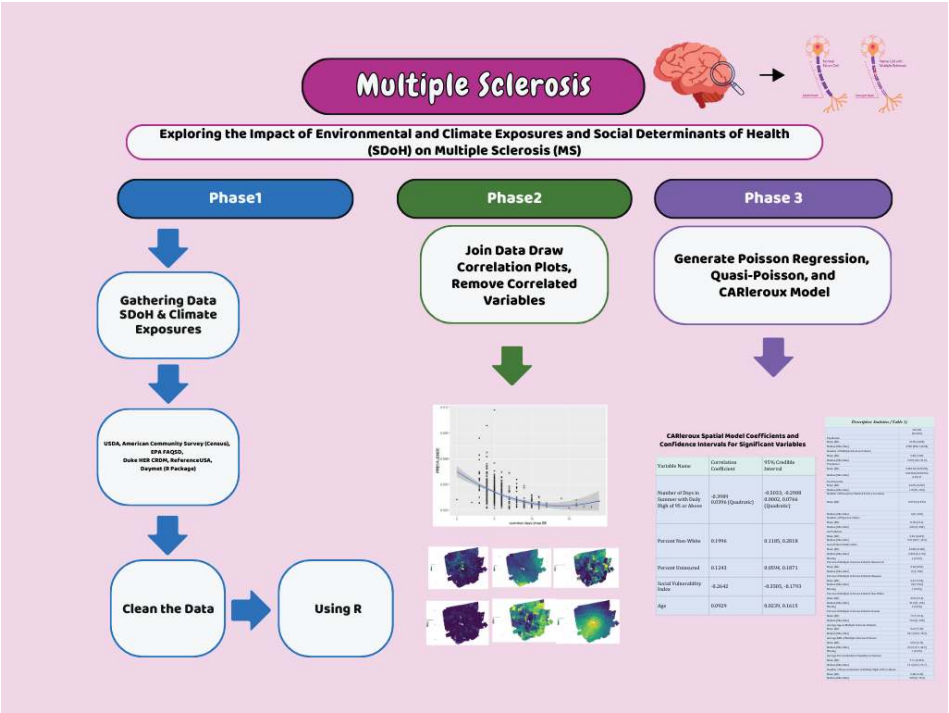
MOHAMMED AL-HUBAIL



ALI ALSALMAN

Multiple Sclerosis (MS) is a serious neurological disorder that affects millions of people around the world, but its causes are not fully understood. While genetics are known to play a role, there is growing evidence that social factors, environmental conditions, and climate stresses may also contribute to the prevalence of MS.

This study aims to explore these connections by analyzing data from Duke University patients living in North Carolina. We will identify associations between social determinants of health (SDoH), environmental factors, and climate variables with the prevalence of MS. This research will provide, for the first time, a detailed picture of how these factors might influence MS in this region, offering insights that could lead to better interventions and outcomes for patients.


Multiple Sclerosis
Exploring the Impact of Environmental and Climate Exposures and Social Determinants of Health (SDoH) on Multiple Sclerosis (MS)
Phase1
Gathering Data SDoH & Climate Exposures
USDA, American Community Survey (Census), EPA FAQSD, Duke HDE CRDH, ReferenceUSA, Daymet (R Package)
Clean the Data
Using R
Phase2
Join Data Draw Correlation Plots, Remove Correlated Variables
Phase 3
Generate Poisson Regression, Quasi-Poisson, and CARleroux Model
CARleroux Spatial Model Coefficients and Confidence Intervals for Significant Variables
Variable Name
Correlation Coefficient
95% Credible Interval
Number of Days in Summer with Daily High of 95 or Above
-0.3989
(-0.5916 (Quadratic)
-0.5933, -0.2048
-0.8602, 0.0766
(Quadratic)
Percent Non-White
0.1996
0.1185, 0.2818
Percent Uninsured
0.1243
0.0934, 0.1571
Social Vulnerability Index
-0.2642
-0.3505, -0.1793
Age
0.0929
0.0239, 0.1619

MISTRAL: DYNAMIC DATA ANALYSIS OF SECURITY THREATS WITHIN BIOMEDICAL RESEARCH ENVIRONMENTS



MANAR ABDULLAH ZAIN ALDEEN

MISTRAL (Massive Internal System Traffic Research Analysis and Logging) is a project funded by NSF and Cisco, aimed at enhancing network monitoring and data collection to protect sensitive patient data in biomedical research environments.

This study establishes a baseline for normal network traffic and enriches datasets with synthetic malicious data to develop automated detection methods for cyber threats. We created two machine learning models that achieved over 96% precision in detecting malicious traffic, particularly port scanning attacks. Our methodologies include traffic visualizations that assist users in monitoring potential threats. The research underscores the importance of safeguarding patient data and ensuring operational security within biomedical research, leveraging advanced techniques in machine learning and data analysis.

MISTRAL: Dynamic Data Analysis of Security Threats within Biomedical Research Environments

Manar Zain Aldeen, Manahil Tariq, Bob Zhang

Manar Zain Aldeen : Taibah University, Medina, Saudi Arabia. Manahil Tariq and Bob Zhang : Duke University, Durham, USA

Introduction

MISTRAL (Massive Internal System Traffic Research Analysis and Logging), funded by NSF and Cisco, enhances Duke University's network monitoring and data collection. The project aims to protect sensitive patient data, including genome and exome sequences, from cyber threats, ensuring research integrity and operational security in biomedical research.

Objective

1. Establish a baseline of normal network traffic.
2. Enrich datasets with synthetic malicious data from port scanning using a Red Team.
3. Develop automated detection methods for malicious activity.
4. Protect sensitive patient data by creating machine learning models that detect breaches and ensure data integrity, preventing unauthorized access

Methodology

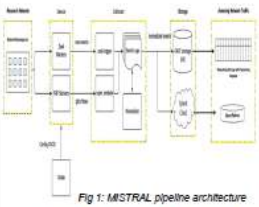
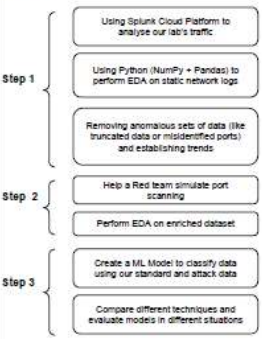


Fig 1: MISTRAL pipeline architecture

We followed the systematic step by step method outlined below to achieve our objectives.



Result

Machine Learning Models

Logistic Regression: Our logistic regression model focuses on detecting port scanning activity. The model was trained using features like Packet-Octet Ratio, Packet-Reverse Packet Ratio, connection counts, and unique destination ports for each IP over 24-hour periods. See results in Fig. 2.

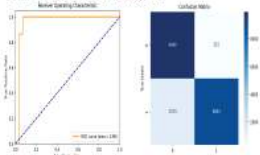


Fig 2: logistic regression model ROC (left) and confusion matrix (right)

Neural Network: Our neural network model monitors data flow to predict port scanning activity. It was trained on flow counts, flow count ratios, flow duration, and destination ports. See results in Fig. 3.

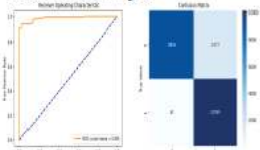


Fig 3: neural network model ROC (left) and confusion matrix (right)

Evaluation of Models

Some other models explored included Random Forest, SVM, and Isolation Forest. However, due to limited synthetic attack data, we primarily focused on two techniques. Our Neural Network model, tested on traffic data (over 60,000 entries) from July 17th, 2024, where no attacks occurred, achieved 79% accuracy with 21% false alarms. See Fig. 4 for model loss and results.

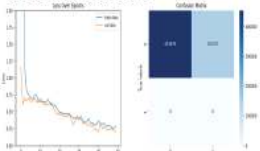


Fig 4: neural network model loss (left) and confusion matrix (right)

Traffic Visualisations

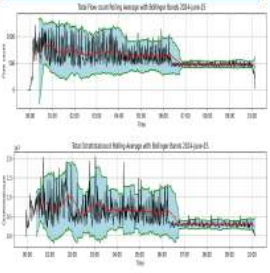


Fig 5: traffic features: flow count (top), octet count (middle) and PCR (bottom)

Traffic pattern variations on June 15, 2024, show key features—flow count, octet count, and PCR—crucial for detecting abnormalities. Data from midnight to 10:12 AM highlight deviations from Bollinger Bands, which should alert users. Significant spikes were cross-checked with alarming IP addresses identified by machine learning models for enhanced network security.

Conclusion

In conclusion, we developed two machine learning models to detect malicious traffic, particularly port scanning attacks. Both models achieved over 96% precision in anomaly detection. Traffic visualizations assist users in monitoring potential attacks. This research is vital for protecting sensitive patient data and ensuring security and privacy in biomedical research environments.

Acknowledgements

We thank the Duke ITSO staff, especially Chloe Poroslay, Alexander Merck, and Eric Hope, for their invaluable assistance. We also acknowledge the KAUST Academy, Code+ MISTRAL team, Honeypots team, and Data+ management for their contributions and support.

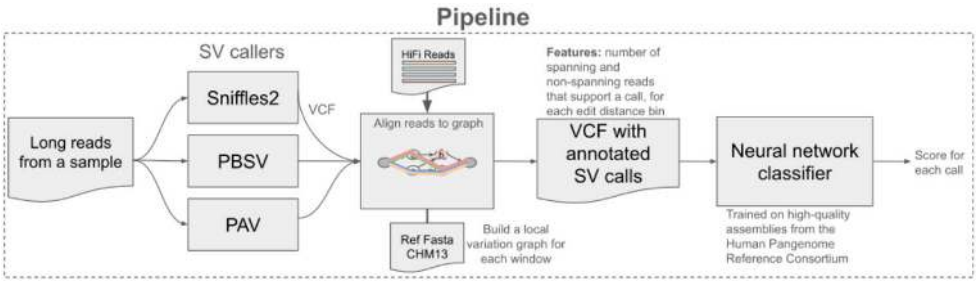
25

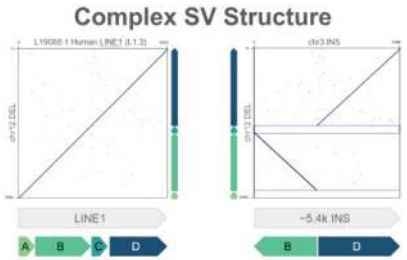
EVALUATING LONG-READ BREAKEND CALLS

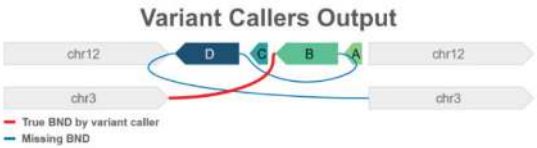


KUMAIL MOAYED ALAWA

Structural Variants (SVs) are variations (like insertions, deletions, inversions, duplications) of length ≥50 base pairs that are found when comparing a sample genome to a reference genome. SVs contribute significantly to human genetic diversity and disease, and they are one of the main goals of long-read sequencing projects. The ongoing All of Us Long Reads project is sequencing ~1000 PacBio CCS samples at 8x coverage. Each sample is analyzed with multiple state-of-the-art SV callers, which, in addition to standard SV calls, emit also a few breakend (BND) calls per sample. BND calls are often discarded because they are believed to be unreliable, but they are potentially useful for representing complex SVs. This work aims at building a machine learning classifier that distinguishes between true and false BNDs by leveraging BND-specific features. We also focus on interpreting the complex variants represented by some true BND calls. We show that true BND calls can be successfully separated from false BND calls (AUC = 0.9). Moreover, we show that, although variant callers are incapable of fully capturing complex variants with BNDs, some of the BND calls they emit are informative.


Pipeline
SV callers
Sniffles2
PBSV
PAV
VCF
Long reads from a sample
HiFi Reads
Align reads to graph
Ref Fasta CHM13
Build a local variation graph for each window
Features: number of spanning and non-spanning reads that support a call, for each edit distance bin
VCF with annotated SV calls
Neural network classifier
Trained on high-quality assemblies from the Human Pangenome Reference Consortium
Score for each call


Complex SV Structure
LINE1
~5.4k INS
A
B
C
D


Variant Callers Output
chr12
chr3
D
C
B
A
True BND by variant caller
Missing BND

MONITORING SPARTINA ALTERNIFLORA USING SELF-SUPERVISED LEARNING

OTHMANE ECHCHABI

ADVAIT BHASKAR PANDIT

MEIXIANG DU



BAYAN HAMEED ALHARBI

The invasive marsh grass Spartina alterniflora poses a significant threat to coastal ecosystems. This project aimed to develop a robust monitoring system using self-supervised learning techniques applied to satellite imagery data. We explored two methods, SimCLR and DINO, to effectively detect and track the spread of Spartina alterniflora in coastal regions. The key challenges included data sparsity and the lack of annotated datasets, which were addressed through the use of data augmentation and alternative data sources. Our findings demonstrate the potential of self-supervised learning in enhancing the accuracy and generalization of Spartina alterniflora detection, paving the way for improved environmental monitoring and conservation efforts.


Satellite imagery
DINO
Spartina alterniflora
SimCLR
0
50
100
150
200
250

27

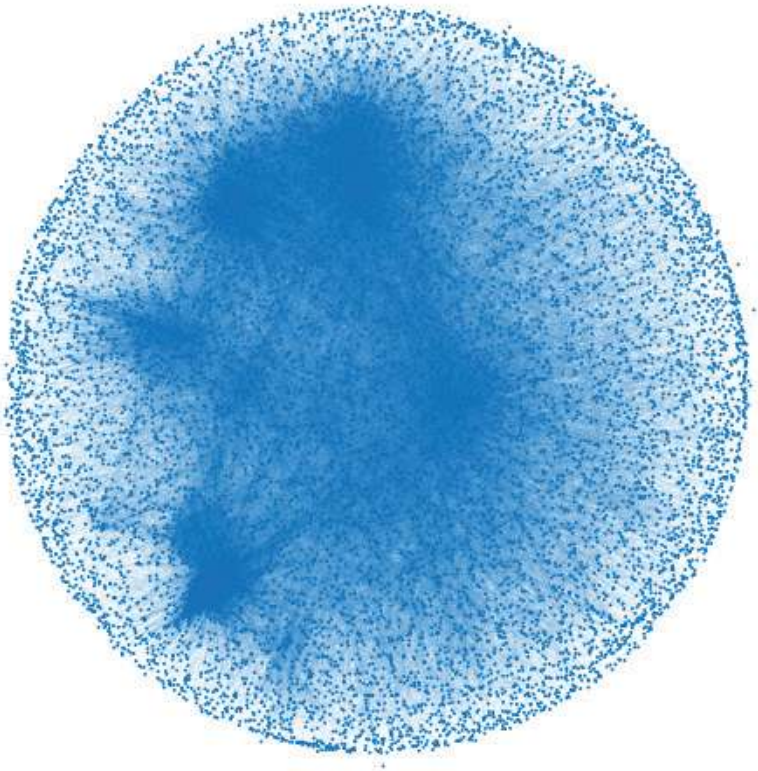
GANLA: GPU ACCELERATED NETWORK LAYOUT ALGORITHM



HASSAN ALSAYHAH

Network layout algorithms are used to visualize networks in a human friendly manner and are pivotal for data exploration. Currently, the Kamada-Kawai algorithm is commonly used for network visualization. However, Kamada-Kawai does not scale to large networks, such as those derived from human functional genomics data.

We have developed a fast, scalable version of Kamada-Kawai which takes advantage of graphics processing unit (GPU) hardware acceleration, called GPU Accelerated Network Layout Algorithm (GANLA). GANLA computes the layout positions for all nodes in the network in parallel, whereas Kamada-Kawai computes these positions sequentially. Moreover, GANLA utilizes a mini-batch technique which enables the scalable layout of networks with many nodes, with a fixed memory footprint. In conclusion, GANLA scales to larger networks and should be faster than Kamada-Kawai algorithm, enabling the exploitation of larger and more complex biological networks relevant to human functional genomics.



28

EVALUATION OF TRANSCRIPT-GUIDED CELL SEGMENTATION ON GBM-DERIVED SINGLE MOLECULE SPATIAL RANSCRIPTOMICS DATA

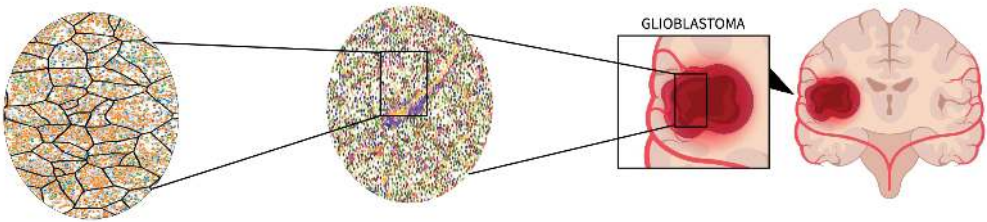


FATEMAH IBRAHIM ALSOLAIMAN

Glioblastoma (GBM) is a heterogeneous and aggressive brain tumor characterized by a variety of genomic mutations and aberrant gene expression patterns. This molecular heterogeneity significantly influences tumor development and progression. However, tumor progression is also a result of cells forming communities due to their spatial relationships, which gives rise to functional and behavioral attributes to promote survival and resistance. Thus, understanding spatial properties is critical to developing targeted therapies and improving patient outcomes.

Accurate single-cell segmentation is vital to measure the spatial distribution of GBM gene expression and tumor cell organization with their surrounding microenvironment and at multiple spatial scales. To address this, we are implementing, testing, and evaluating different segmentation methods to determine the best performing algorithms. This project will compare two different segmentation methods with original Xenium segmentation masks- Baysor and Proseg. This comparison involves integrating segmentation masks with images, annotating randomly selected regions, and evaluating their accuracy and suitability.

The expected outcome of this project is a comprehensive evaluation of segmentation tools, resulting in improved methods for analyzing spatial gene expression data in GBM. By advancing our understanding of tumor architecture, this study offers an opportunity to improve the prognosis and treatment outcomes of patients with GBM.



Cell Segmentation on GBM-Derived Single Molecule Spatial Transcriptomics Data

29

ENHANCING NAMED ENTITY NORMALIZATION IN BIOFACTOID USING NLP TECHNIQUES



FAISAL MELFY
M. ALKULAIB

Our project focuses on enhancing the accuracy of named entity normalization (NEN) in the Biofactoid web tool, which is used to curate biological interactions from scientific articles. By leveraging natural language processing (NLP) techniques, we aim to reduce common errors, such as orthologue mismatches (gene name is the same in multiple species), improving the reliability of bioentity data.

This enhancement will contribute to more accurate mapping of cellular processes, aiding in disease research and therapeutic development.

Enhancing Named Entity Normalization in Biofactoid Using NLP Techniques

Faisal Alkulaib[1,2] Jeffrey Wong[2] Gary Bader[2]
[1]KAUST Academy
[2]Terrence Donnelly Centre for Cellular & Biomolecular Research

Contact Info. Biofactoid

Introduction

Biofactoid is a web tool that helps authors curate the scientific findings in their article into a standardized, machine-readable format, making biological knowledge more accessible. A key step in this process is **gene normalization**, which means matching an author's name for a gene to a unique record in a curated database of genes However, gene normalization can introduce errors, such as orthologue *errors* (gene name is the same in multiple species). e.g., an author label "HDAC3" should be associated with mouse rather than the human gene.

Hypothesis:

"Mismatches can be avoided by using hints within articles."

A **hint** is a piece of information extracted from an article, it can be an organism in the article or a gene name under investigation.

Grounding Search is a flexible and high-accuracy search system designed to identify biological entities, such as genes and molecules, by mapping user-provided names to standardized database identifiers. This service improves the accuracy of data retrieval and integration in bioinformatics tools by disambiguating entity names, ensuring they are correctly linked to the appropriate database entries.

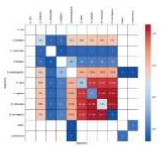


Figure 1: Gene names shared between species

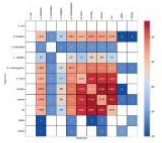


Figure 2: Gene names & synonyms shared between species

Figures 1 and 2 highlight the challenges in data integration due to frequent gene symbol sharing and increased overlaps when synonyms are included, complicating the gene normalization process. Accurate gene normalization is crucial because errors, such as orthologue mismatches, can lead to incorrect associations between genes and species, potentially skewing research outcomes. By improving gene normalization, particularly through the use of article-derived hints and tools like Grounding Search, we can enhance the reliability and accuracy of curated biological data. This, in turn, supports more robust and reproducible scientific findings, facilitating better integration and application of biological knowledge across research studies.

Objectives

- Use NLP to extract normalized entities detected in an author's article
- Reduce gene normalization errors
- Increase the accuracy and consistency of curated biological data

Method

To improve gene normalization accuracy, we used PubTator 3.0 to extract contextual information about organisms and bioentities from article titles and abstracts. These "hints" were integrated into Biofactoid's search engine to correct gene normalization, resulting in a more precise and reliable entity normalization process.

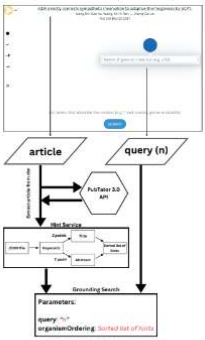


Figure 3: Hints pipeline for entity normalization

Results

The implementation of NLP-derived hints in the gene normalization process has significantly improved the accuracy of entity matching in Biofactoid. As demonstrated in Figures 4 and 5, the enhanced gene normalization process successfully corrects prior mismatches, accurately associating genes with the correct species. Before the enhancement, the grounding-search engine incorrectly matched "Aida" to a human gene, leading to wrong curation. However, after the enhancement, the correct match to the mouse gene "Aida" was achieved, demonstrating the effectiveness of the refined gene normalization process in improving data accuracy and reliability in Biofactoid. This enhancement reduces the likelihood of incorrect associations and enhances the quality of curated biomedical data.

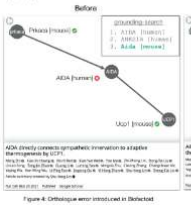


Figure 4: Orthologue error introduced in Biofactoid

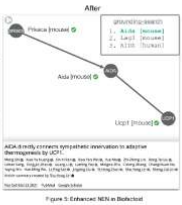


Figure 5: Enhanced NEN in Biofactoid

Conclusion

The use of NLP-derived hints in Biofactoid has significantly improved gene normalization accuracy, reducing errors like orthologue mismatches and enhancing data quality in biological research. Future work will focus on addressing synonym errors by using NLP to send "entity hints" with queries, helping to prioritize genes with competing synonyms and further improving accuracy.

References

1- Franz et al., (2021). A flexible search system for high-accuracy identification of biological entities and molecules.
2- Wei et al., (2024). PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge.

Acknowledgement

Faisal Alkulaib received a KAUST Academy sponsorship. The "Enhancing Named Entity Normalization in Biofactoid Using NLP Techniques" is supported by the Data Sciences Institute, University of Toronto.

30

UNDERSTANDING MISTAKES IN VARIANT CALLING



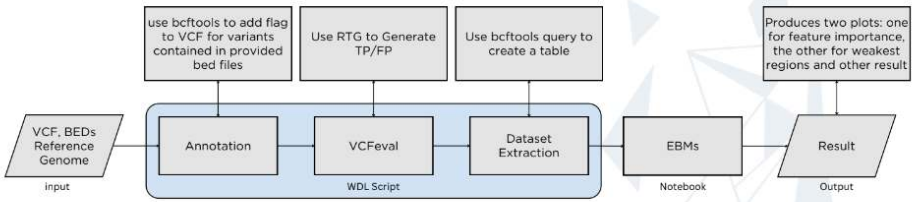
EYAD ABDULLAH ALQAYSI

Variant calling errors, including false positives and negatives, present challenges in genomic analyses. This study employs Explainable Boosting Machines (EBMs) to uncover the factors contributing to these errors.

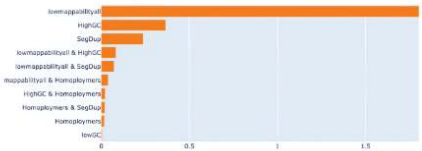
We developed a flexible pipeline capable of ranking the weakest genomic regions across multiple sequencing technologies, providing valuable insights for improving accuracy in future genomic studies.

The pipeline's adaptability allows it to work with various technologies and datasets, making it a powerful tool for investigating and mitigating variant calling errors in diverse genomic contexts.

Pipeline & Result

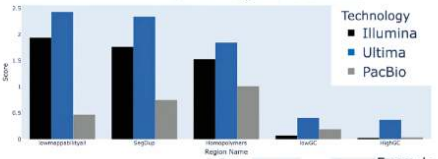


Feature importance



Example

Weakest regions



Example

BROAD INSTITUTE

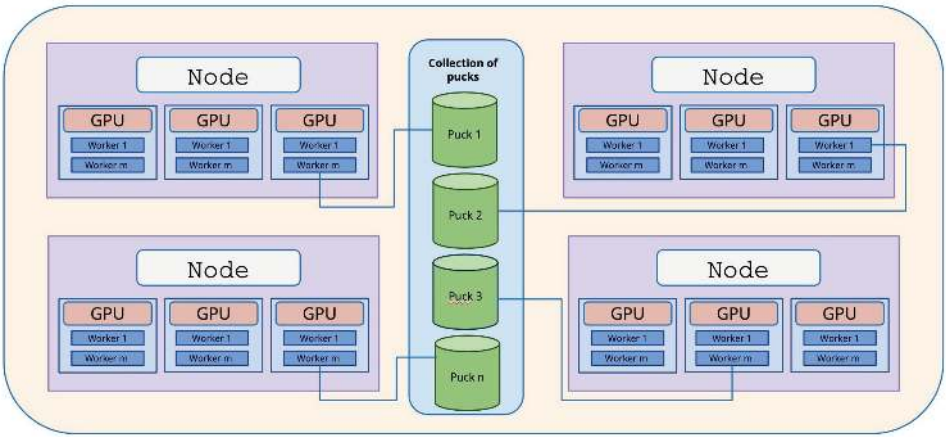
31

DISTRIBUTED DATALOADER FOR SPATIAL TRANSCRIPTOMICS



EBTISAM SALEH ALSHEHRI

Spatial biology explores the organization and interaction of biomolecules like RNA, DNA, and proteins within tissues to gain insights into cancer, neuroscience, and disease. Spatial Transcriptomics, a key aspect of this field, focuses on how gene expression varies across tissue regions, aiding in Disease Mapping. This project aims to develop a Spatial Transcriptomics Dataloader to enhance model training scalability within a distributed ecosystem by generating mini-batches from tissue pucks using various sampling strategies. Conducted in a cell biology AI lab contributing to the Cellarium ML core library, the project involved utilizing distributed systems and single-cell analysis libraries, and implementing the dataloader under expert guidance. Key features include pre-computed KNN computations, chunk-based partitioning with exclusive data assignment, and sampling strategies like random and KNN subsampling to capture different length scales. The Dataloader supports multi-node and multi-GPU distributed model training and includes batch tracking for workflow monitoring.



A workflow of the data distribution in the spatial data loader

AI FOR EXAMINING COMMERCIAL DETERMINANTS OF HEALTH IN ONLINE GAMING AND GAMBLING



EBTIHAL HANI ABDALLAH ALMASHAMA

Online gambling, driven by technical innovation and accessibility, poses significant challenges in regulation and public health.

In Ontario, gambling participation has surged, with $63 billion in online bets last year, raising concerns about gambling-related harms. Researchers have explored machine-learning methods to identify problematic gambling, but challenges like limited user data and insufficient focus on social aspects persist.

Our project addresses these gaps by using publicly available data to develop an interactive tool that highlights the social dynamics of online gambling.

This tool will identify key factors, create novel visualizations, and enhance data collection, providing insights to support public health strategies and responsible gambling practices

Methodology

Marketing Environment
(Evaluate marketing strategies within Canada and globally)

1



Marketing Images → **LlaVA Model** → **Marketing Strategies**

900+ Images

Obtained from prominent gambling websites. Images sourced from **5 countries** with established markets

56% US, UK, AU, GE

44% Canadian Context

Prompt Engineering

Conducted to enhance the accuracy of feature identification and classification.

65 Features

Broken into **6 main categories**: Visual Characteristics, Visual Elements, Marketing Messaging, Responsible Messaging, and Target Audience.

Socialization and Economic Stress
(Investigate online gambling social groups & identify trends in economic stress)

2 → 3

Social Media Comment Data → **Llama3 Model and Analysis** → **Visualizations**

33,000+ Youtube
40,000+ Reddit

from popular online gambling platforms.

6 Social Features

Online Gambling Platforms, Types of Bets, Gambling Outcomes, User Sentiments, Practices or Incidents.

Canada-Wide Survey

(*AGRI Project, 2018*) with demographic and financial stress data from 10,000 respondents

MAKING CLIMATE HAZARD RISK DATA USEFUL FOR NORTH CAROLINA COMMUNITIES

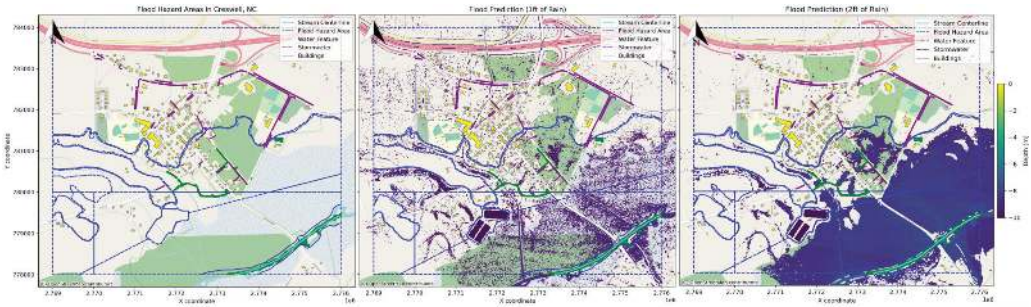


ALJOUHRA ALHARGAN

Creswell, a small, predominantly Black, and underserved community in North Carolina, has been identified by the National Risk Index as highly vulnerable to natural hazards and climate change. Located near the coast at a low elevation with a river running underneath, Creswell faces significant flood risks, leading to rising insurance costs and diminishing market availability.

This project aims to develop a pilot model using the HEC-RAS system to simulate various flood scenarios and estimate potential building damages. Our findings indicate that nearly half of the town is at high risk of severe flooding, with potential damages costs ranging from 12% to 33% of residents' annual income.

These results underscore the need for targeted flood mitigation strategies to protect Creswell's vulnerable infrastructure.


Flood Hazard Areas in Creswell, NC
Flood Prediction (1ft of Rain)
Flood Prediction (2ft of Rain)
Stream Centerline
Flood Hazard Area
Water Feature
Stormwater
Buildings
X coordinate
Y coordinate
Depth (ft)

FURTHER DEVELOPING DUKE'S WEARABLE INFECTION DETECTION PLATFORM



ALJAZI WALEED ALOMAIRI

This project is an exciting opportunity to bridge the gap between technology and healthcare, with the potential to make a significant impact on people's lives. By harnessing the power of wearable devices, we aim to create a system that not only captures and processes health data in real-time but also translates this data into actionable insights.The integration of machine learning will enable us to detect early signs of health changes, potentially catching infections or other conditions before they escalate into more serious issues. Our work goes beyond just data collection it's about making this information accessible and meaningful to users. By focusing on UI/UX development, we ensure that the platform is intuitive, user-friendly, and capable of delivering critical health information in a way that is easy to understand and act upon.

Ultimately, this project has the potential to empower users with timely health information, improving their well-being and possibly preventing severe health outcomes. It's a step toward a future where wearable technology plays a crucial role in proactive healthcare, offering users a better quality of life through early detection and intervention.

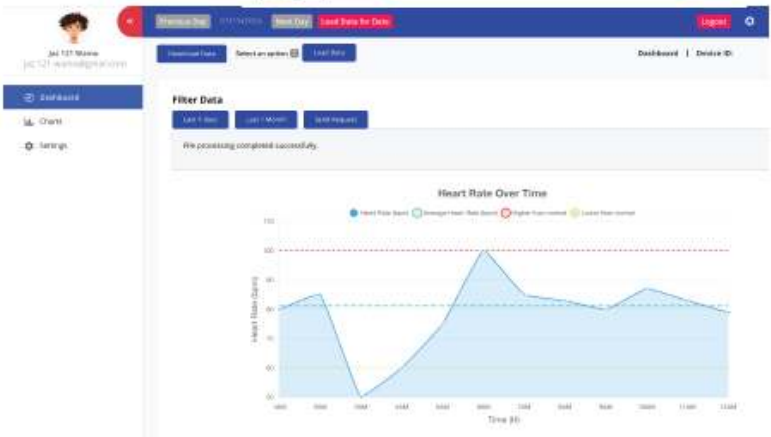
(Fig. 1): Device Sensors



(Fig. 2): System Desing



(Fig. 3): web Dashbord



35

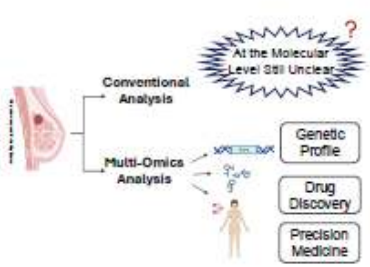
MULTI-OMICS DATA ANALYSIS & VISUALIZATION TO INVESTIGATE THE MOLECULAR PROFILE OF BREAST CANCER CELLS IN AN ORGAN-ON-A-CHIP MODEL

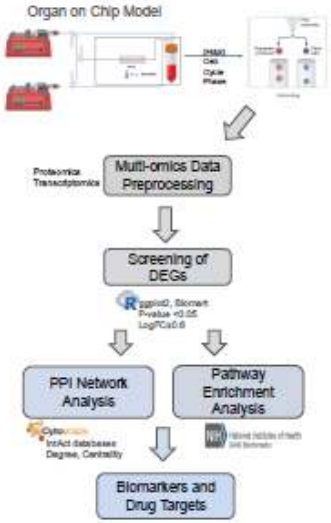


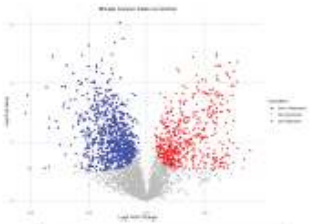
AISHA SOLIMAN ALSOMIRY

The emergence of high-throughput multi-omics technologies facilitate the comprehensive analysis of different data, such as genomes, transcriptomes, proteomes, and metabolomes. Multi-omics data, however, require powerful analytical pipelines to generate meaningful biological insights. Here, we propose a novel integrated omics data(proteomics and transcriptomics) analysis pipeline that applies Gene Ontology (GO) enrichment, pathway enrichment, and Protein-Protein Interaction (PPI) network analyses to determine the mechanisms and pathways most involved in a breast cancer.

Overall, this omics data analysis pipeline provides a robust methodology for interpreting large-scale biological data, facilitating the discovery of pathways, biomarkers, and drug targets in a disease environment. We apply this pipeline for breast cancer, but this methodology can be used for a variety of diseases.



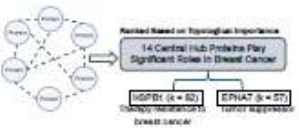




(Figure 1)
Volcano plot of breast cancer cells vs control using R(4.4.1). 826 Up-Regulated genes (upper-right dots colored red), 826 Down-Regulated genes (upper-left, dots colored blue).



(Figure 2)
The PPI Network was constructed from DEGs and nonDEGs. (14)crucial proteins were identified in the network. HSPB1 (k = 82) and EPHA7 (k = 57) were identified as hub proteins. The 14 proteins are colored according to their degree.



THUNDERPACK
FOR EHR INGESTION



ABDULRAHMAN MOHMMED ALHOMAYANY

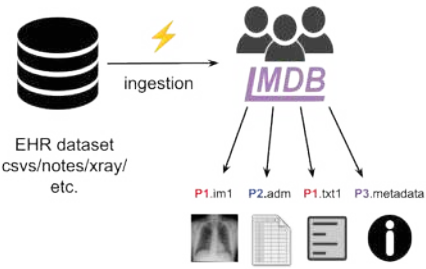
This project explores an alternative method for Electronic Health Record (EHR) ingestion using ThunderPack,

as compared to the current approach employed by the JEDI pipeline at the Broad Institute.

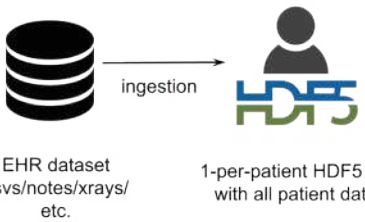
The results were promising: ThunderPack significantly reduced size on disk for the ingested dataset as well as ingestion time.

Furthermore, improved read speed metrics were measured for text and tabular data, the two primary modalities that JEDI works with.

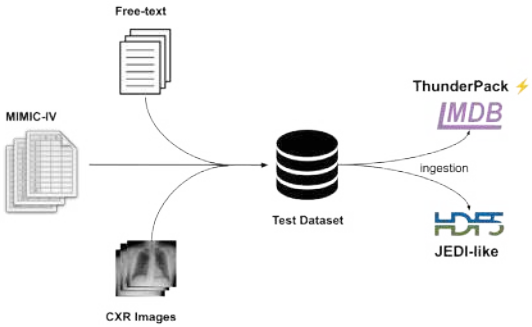
Alternative Approach



JEDI Approach



Workflow



37

A TEXTUAL ANALYSIS OF AGRICULTURAL RESEARCH



ABDULRAHMAN ABDULLAH ALALWAN

Understanding trends in research topics, funding, and citations over time offers valuable insights for researchers. This study focuses on identifying trends in research topics and funding in agricultural and applied economics research projects.

We analyze a corpus of research papers published in Applied Economics Perspectives and Policy (AEPP) and the American Journal of Agricultural Economics (AJAE), two journals with significant impact in the field of agricultural and applied economics.

We aim to provide insights into research priorities within the last two decades.

A Textual Analysis of Agricultural Research

Abdalrahman Alalwan, Darsh Mandera, Haozhou Zhang
Project Manager: Poojitha Balamurugan
Faculty Leads: Drs. Norbert Wilson, Leslie Collins, Boyla Mainsah

Introduction

Understanding trends in research topics, funding, and citations over time offers valuable insights for researchers. This study focuses on identifying trends in research topics and funding in agricultural and applied economics research projects. We analyze a corpus of research papers published in *Applied Economics Perspectives and Policy* (AEPP) and the *American Journal of Agricultural Economics* (AJAE), two journals with significant impact in the field of agricultural and applied economics. We aim to provide insights into research priorities within the last two decades.

Objectives

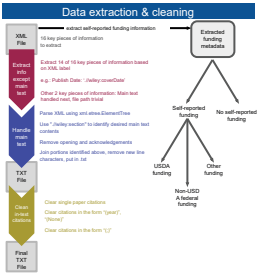
- Analyze a dataset of agricultural economics research papers to identify trends in research topics and funding patterns over time
- Use natural language processing techniques for topic extraction
- Perform time series analysis on the extract topics to identify longitudinal trends in research topics and funding

Data

Published articles

- 4,102 XML files representing 4,102 publications from the *American Journal of Agricultural Economics* (AJAE) and *Applied Economic Perspectives and Policy* (AEPP) spanning from February 2003 to June 2024
- 3,417 records of year-by-year citation counts for each document, identified by DOI, as extracted from the Web of Science
- After filtering non-research articles (book reviews, listings) and aligning filtered publication corpus with citation records based on DOI, the reduced dataset consists of 3,049 XML files

Data extraction & cleaning



Procedure

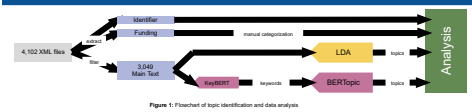
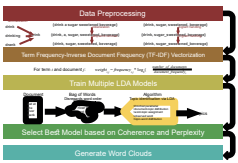


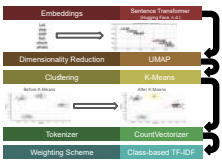
Figure 1: Flowchart of topic identification and data analysis

Methods

Latent Dirichlet Allocation (LDA)



BERTopic



Results

LDA

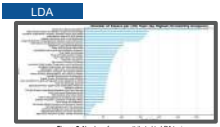


Figure 2: Number of papers attributed to LDA topics



Figure 3: LDA topic word clouds

BERTopic



Figure 4: BERTopic topic hierarchical grouping



Figure 5: BERTopic word clouds

- Extracted topics from each topic model were labeled by an internal subject matter expert
- LDA is more effective at capturing the thematic content of each paper, providing a general overview of research themes, while BERTopic is more effective at capturing the semantic nuances, offering more coherent topics
- A significant portion of papers in the corpus (total N = 3049) were attributed to the following topics: "Advances in Microeconometrics" (N = 332), "Public Sector Impacts on Markets, Growth and Trade" (N = 170), and "Economic Experiments: Auctions, Discrete Choice Models, and Incentives" (N = 167)

Longitudinal Trend Analysis



Figure 6: Longitudinal prevalence of topics grouped by increasing and decreasing trends

- There are similarities in topic themes and trends of topics extracted with both topic models, such as an increase in papers on "Household and Development Economics" and their related topics over the past two decades and a noticeable growth of COVID-19 related topics in 2020
- Example COVID-19 related topics include "Health Surveys, Demographics, and COVID-19 Impact" and "Pandemic Resilience and Supply Chains"
- We identify no coherent themes in topics with decreasing frequency over time

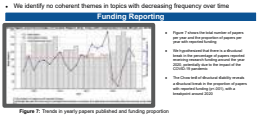


Figure 7: Trends in yearly papers published and funding proportion

- Figure 7 shows the total number of papers per year, and the proportion of papers per year with reported funding.
- We hypothesized that there is a structural break in the percentage of papers reported funding around the beginning of the year 2020, potentially due to the impact of the COVID-19 pandemic
- The Chow test of structural stability reveals a structural break in the proportion of papers with reported funding (p= 2001), with a breakpoint around 2020

Conclusion and Limitations

- There is consistency in themes and longitudinal trends of topics extracted by both topic models across the journals of high significance in agricultural and applied economics research
- The impact of the 2020 COVID-19 pandemic was reflected in changes in patterns in published research topics and funding reporting in agricultural and applied economics
- A limitation is the need for independent topic label assignment by external subject matter experts
- Future work will analyze relationships between topics, funding, and citations

References

Bird, S., Loper, E., & Klein, E. (2009). Natural language processing with Python. O'Reilly Media, Inc.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022. https://jmlr.csail.mit.edu/papers/v3/blei03a.html

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. https://arxiv.org/abs/1810.04805

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv. https://doi.org/10.48550/arxiv.2203.05794

Hugging Face. (n.d.). all-MiniLM-L6-v2 [Software]. Hugging Face. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Jenkins, M. (2021). Exported Mendeley bib. Retrieved June 10, 2024, from https://www.mendeley.com/reference-management/export/mendeley-bib

Kaushalya, S. P., Balaji, J., Hagedorn, G. J., & Li, Y. (2021). A machine learning analysis of the recent environmental and resource economics literature. Environmental and Resource Economics, 79(1), 39–73. https://doi.org/10.1007/s10640-021-00552-0

Pathak, D., Kumar, S., Bohra, B., & Jun, M.-W. (2022). Economic forecasting using Euro: A cointegrative bibliometric survey. Economic Modelling, 107, 105712. https://doi.org/10.1016/j.econmod.2021.105712

Princeton University. (2012). About WordNet. WordNet. Princeton University. https://wordnet.princeton.edu/about-wordnet

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2).

Yin, J., & Wang, J. (2014). A Dirichlet multinomial mixture model-based approach for short text clustering. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14). ACM. https://doi.org/10.1145/2623330.2623715

Acknowledgements

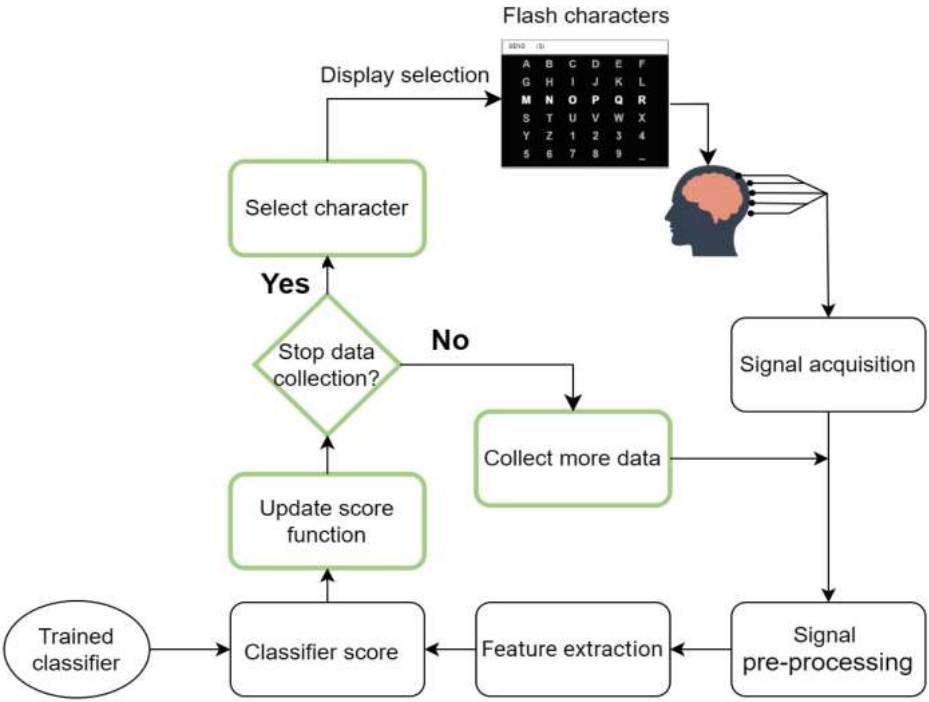
WILEY · AAEA · Duke University Libraries

EXPLORING NOVEL MACHINE LEARNING TECHNIQUES FOR BRAIN COMPUTER INTERFACE (BCI) APPLICATIONS



ABDULLAH ABDULRHMAN ALUMAR

This study explores the effectiveness of transfer learning and domain adaptation in brain-computer interfaces (BCI). Individuals suffering from severe neuromuscular disorders, such as amyotrophic lateral sclerosis (ALS), rely on BCI technologies to communicate. We compared the performance of machine learning models trained with ALS-specific data versus non-ALS data, focusing on character accuracy and estimated stopping time. By employing EEGNet and Stepwise Linear Discriminant Analysis (SWLDA), we determined the benefit of employing non-ALS data for pre-training and fine-tuning machine learning models. Our findings suggest potential improvements in accuracy for individuals with ALS through the use of non-ALS data.


Flash characters
Display selection
A B C D E F
G H I J K L
M N O P Q R
S T U V W X
Y Z 1 2 3 4
5 6 7 8 9 _
Select character
Yes
Stop data collection?
No
Signal acquisition
Collect more data
Update score function
Trained classifier
Classifier score
Feature extraction
Signal pre-processing

39

REFINING & EVALUATING THE USABILITY OF A MATERNAL EQUITY DASHBOARD



ABDULAZIZ ALTELMISSANI

In 2022 in the United States (US), 817 maternal deaths were recorded, with a disproportionate impact on racial/ethnic minority groups, as the maternal mortality rate among Black women was 2.6 times higher compared to non-Hispanic white women. Integrating equity dashboards within individual hospitals to identify and monitor racial variations in maternal health outcomes is one actionable step in addressing the issue.

This study aims to refine and evaluate the usability of a previously developed maternal health equity dashboard.

Method To develop the equity dashboard, we utilized incident reports and delivery dataset of a US-based academic health system from 2019 and 2020. Patient demographics were linked and disaggregated by race/ethnicity.

Measures were selected based on area with previously identified disparities in maternal health outcomes, areas of concern identified through hospital level data analysis.



Figure.3: Current dashboard iteration with additional features