HUSAYN EL SHARIF

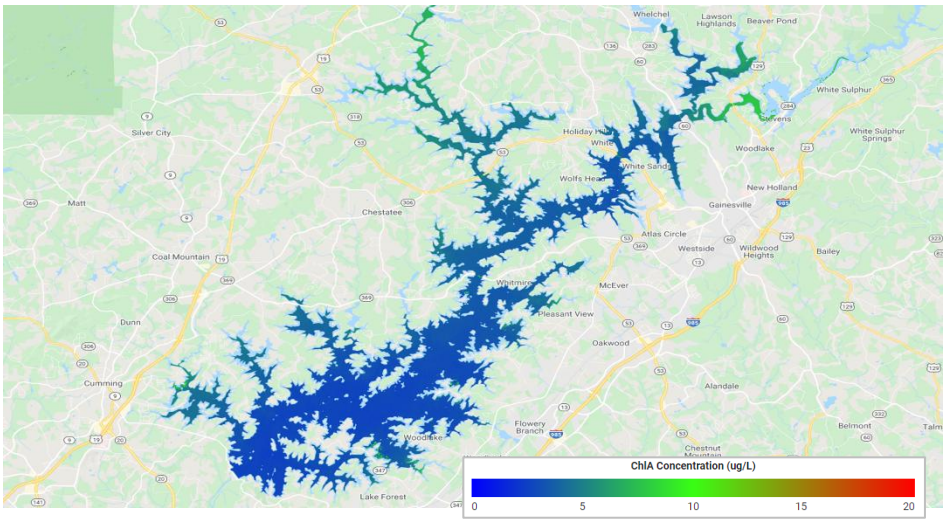
Marietta, GA • helsharif@gmail.com • +1 (561) 247-1430 • *https://helsharif.github.io*

DATA SCIENCE PORTFOLIO

PROJECTS

1 REMOTE SENSING & ML FOR WATER QUALITY ASSESSMENT AT LAKE LANIER

Skills: *Symbolic Regression, Remote Sensing, Time-Series Analysis, Satellite Data Processing, Google Earth Engine, Dockerized Deployment, JupyterLab, Geospatial Data Analysis, Data Visualization, Regulatory Compliance, Early-Warning Systems*

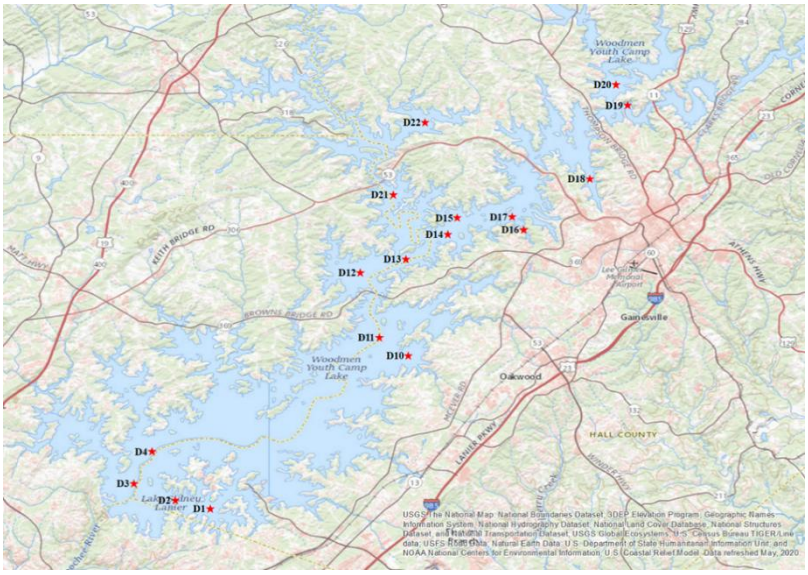



Google Earth view of Lake Lanier from Buford Dam. Gwinnett County is to the right of the dam. (Image Courtesy: Google, Landsat/Copernicus)

Background: This project developed a satellite-based approach for assessing **chlorophyll-a (Chl-a)** concentrations in Lake Lanier, Georgia, using Sentinel-2 imagery. Chl-a, a key indicator of algae biomass, is essential for monitoring water quality and detecting harmful algal blooms, which can pose environmental and public health risks. Traditional field assessments of Chl-a are costly and limited in spatial and temporal coverage, making remote sensing an ideal alternative. By mining Sentinel-2 satellite data, key features were identified to generate weekly Chl-a maps, enhancing water quality management and helping detect potential eutrophication or harmful algal blooms in the lake.

Problem Statement: How can satellite-based remote sensing using Sentinel-2 imagery be used to assess and monitor chlorophyll-a concentrations in Lake Lanier, GA, to improve water quality management and detect harmful algal blooms?

Model Development: A lake field-sampling campaign was conducted from fall 2018 through 2020, measuring photic-zone chlorophyll-a (Chl-a) across multiple locations in Lake Lanier during Sentinel-2 satellite overpasses. The resulting dataset was mined to identify functional relationships between satellite retrievals and in-situ Chl-a concentrations. After segmenting the data by season, **Symbolic Regression** revealed key satellite features that are effective predictors of lake water quality.



Lake Lanier Chl-A Field Sampling Locations

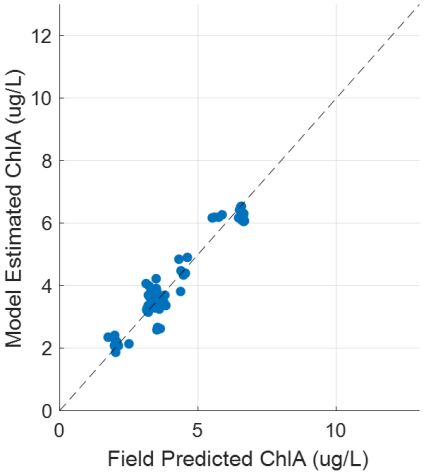
Band	Spectral region	Wavelength range (nm)	Resolution (m)
B1			
B2	Blue	458–523	10
B3	Green peak	543–578	10
B4	Red	650–680	10
B5	Red edge	698–713	20
B6	Red edge	733–748	20
B7	Red edge	773–793	20
B8	NIR	785–899	10
B8A	NIR narrow	855–875	20
B11	SWIR	1565–1655	20
B12	SWIR	2100–2280	20

Sentinel-2 Reflectance Bands

November – April

$$ChlA = 292.084[B2 \cdot (B2 - B4)]^{0.517}$$

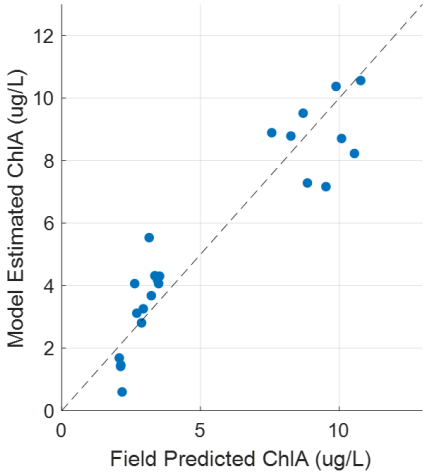
RMSE= 0.417 ug/L R^2= 0.920 AICc= -102.020
ME= 0.000 ug/L StDevErrors= 0.420 ug/L
MAE= 0.328 ug/L StDevAbsErrors= 0.260 ug/L



May – July

$$ChlA = 31.99 \cdot \frac{B3}{B2} - 29.676$$

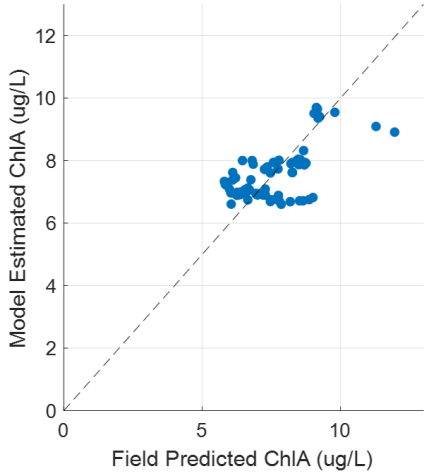
RMSE= 1.189 ug/L R^2= 0.868 AICc= 15.215
ME= -0.000 ug/L StDevErrors= 1.215 ug/L
MAE= 0.977 ug/L StDevAbsErrors= 0.693 ug/L



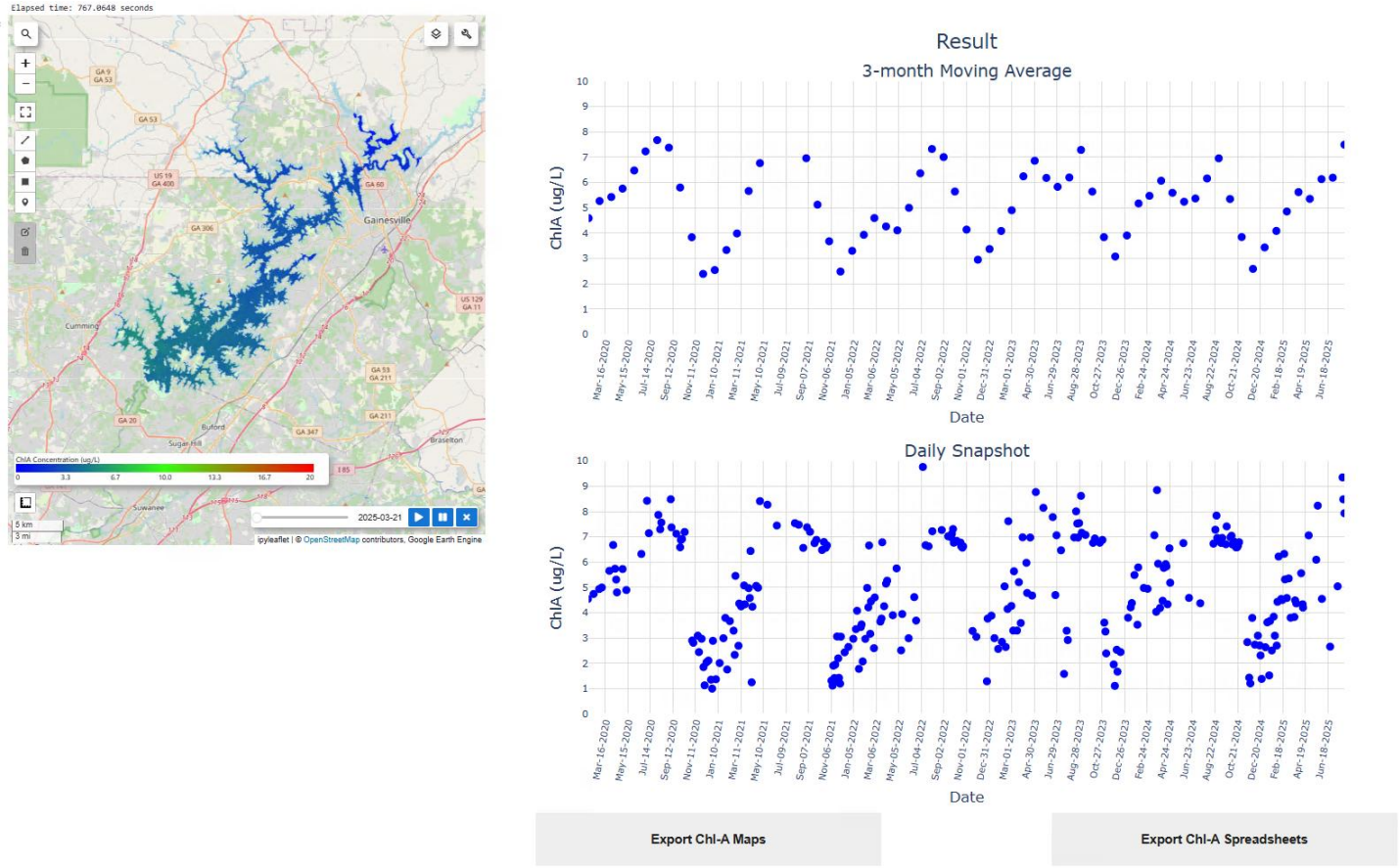
August – October

$$ChlA = 6086.383[B2 \cdot (B3 - B2)] + 6.458$$

RMSE= 0.961 ug/L R^2= 0.403 AICc= 0.338
ME= 0.000 ug/L StDevErrors= 0.968 ug/L
MAE= 0.760 ug/L StDevAbsErrors= 0.592 ug/L



Regression Analysis: Estimating Chl-A (target variable) as functions of Sentinel-2 Reflectance Bands



Web-based Interactive Dashboard for Lake Lanier Water Quality Assessment

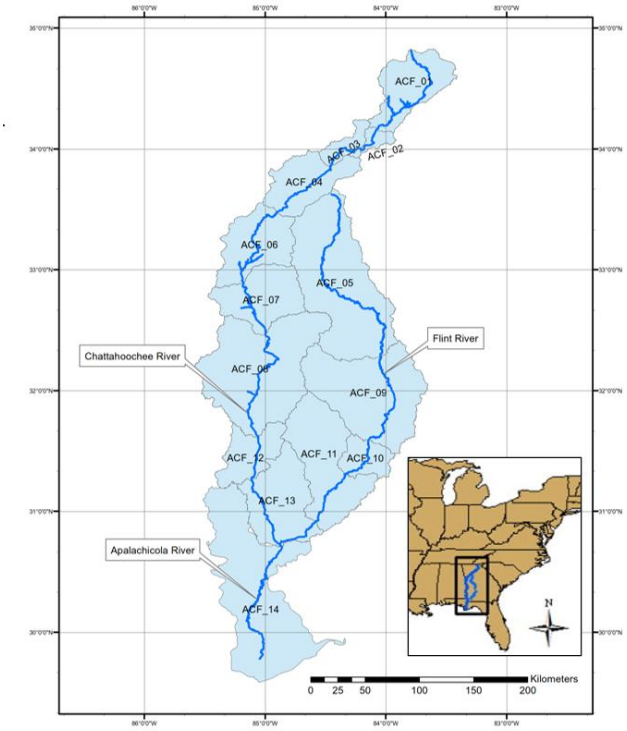
Model Deployment & Results: After developing robust models to estimate chlorophyll-a (Chl-A), the system was deployed on **Google Earth Engine**, enabling near real-time processing of Sentinel-2 satellite imagery. This Chl-A Sat-Tool is now used by the Gwinnett County Department of Water Resources as a decision-support platform for monitoring lake water quality. The tool runs in both standard web browsers and JupyterLab, providing interactive visualizations of regional and lake-wide Chl-A concentrations, with streamlined export of maps and data tables to support operational analysis and early warning of harmful algal blooms.

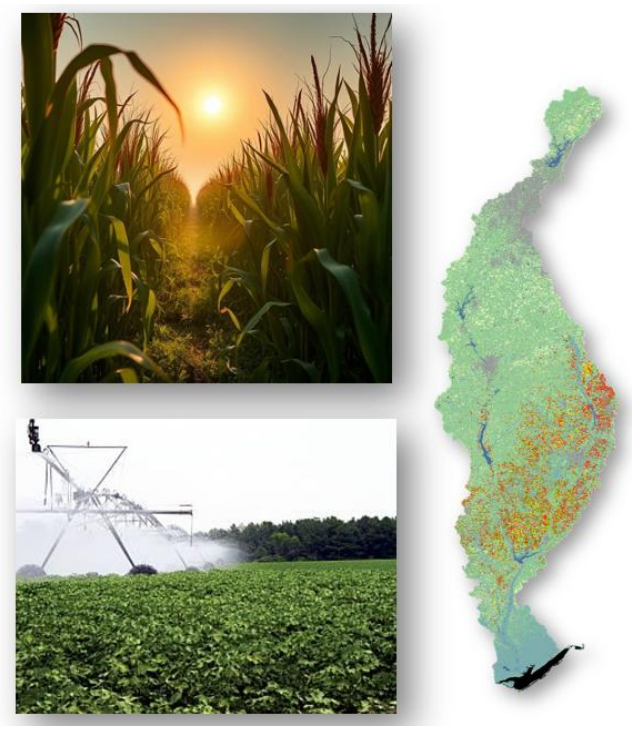
To ensure reproducible deployment and simplified sharing, the application was fully containerized using **Docker**. The complete analysis environment, including geospatial dependencies, processing modules, and the Jupyter-based interface, is encapsulated in a portable Docker image, allowing the system to be launched consistently across machines or deployed in cloud and enterprise environments without configuration overhead. This containerized workflow supports scalable, maintainable use of satellite-based water-quality monitoring tools.

The system supports long-term lake monitoring at a **fraction of the cost** of traditional field sampling and laboratory analysis. Time-series outputs reveal that large areas of Lake Lanier regularly exceed regulatory thresholds for Chl-A, providing clear, data-driven evidence for policymakers and water managers to prioritize remediation and regulatory compliance efforts.

2 CLIMATE-INFORMED CROP YIELD & IRRIGATION PROJECTIONS FOR THE ACF RIVER BASIN

Skills: *Statistical Downscaling and Bias Removal, Time Series Analysis, Impact Assessments, Data Visualization*





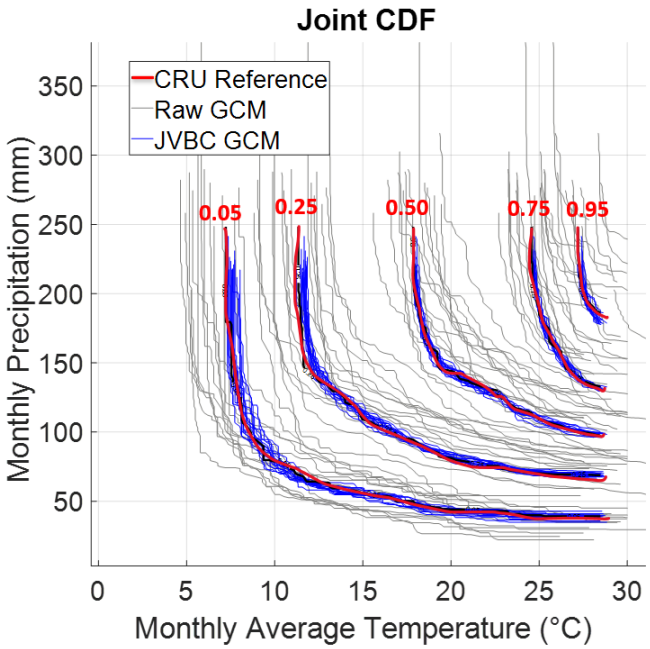
Background: This project integrated soil, crop, and meteorological datasets with the DSSAT Cropping System Model (DSSAT-CSM) to evaluate agricultural resilience in the Apalachicola–Chattahoochee–Flint (ACF) River Basin. The analysis focused on four major crops—peanuts, corn, soybeans, and cotton—and assessed yield sensitivity and irrigation demand under historical climate conditions across normal, dry, and wet years. Regression analysis identified key relationships between growing-season precipitation, potential evapotranspiration, and irrigation demand. These models were extended with bias-corrected climate projections to estimate future scenarios of crop yield and water use through the end of the century, providing insights for more efficient irrigation planning and drought management strategies.

Problem Statement: How will climate-driven shifts in rainfall and evapotranspiration shape crop yields and irrigation demand in the ACF River basin? How can this information guide more resilient water and farm management strategies?

Data Collection: The data required for agricultural modeling of the ACF River basin came from a myriad of public datasets and review of regional farming practices. Sources included USDA Cropland Data Layer, Climate Research Unit Dataset, GRIDMET Historical Reanalysis Daily Weather Data, Harvest Choice Global High Resolution Soil Profile Database, CMIP6 Climate Models, and more.

Climate Data Bias Correction and Spatio-Temporal Downscaling: This work developed a **Joint Variable Bias Correction (JVBC) algorithm** to transform raw CMIP6 climate projections into high-resolution, bias-corrected datasets suitable for agricultural and hydrological modeling. CMIP6 outputs, originally produced at coarse global scales of roughly 100 km and monthly intervals, were downscaled to approximately 10 km spatial resolution with daily-to-monthly temporal detail to meet the precision required for local impact assessments.

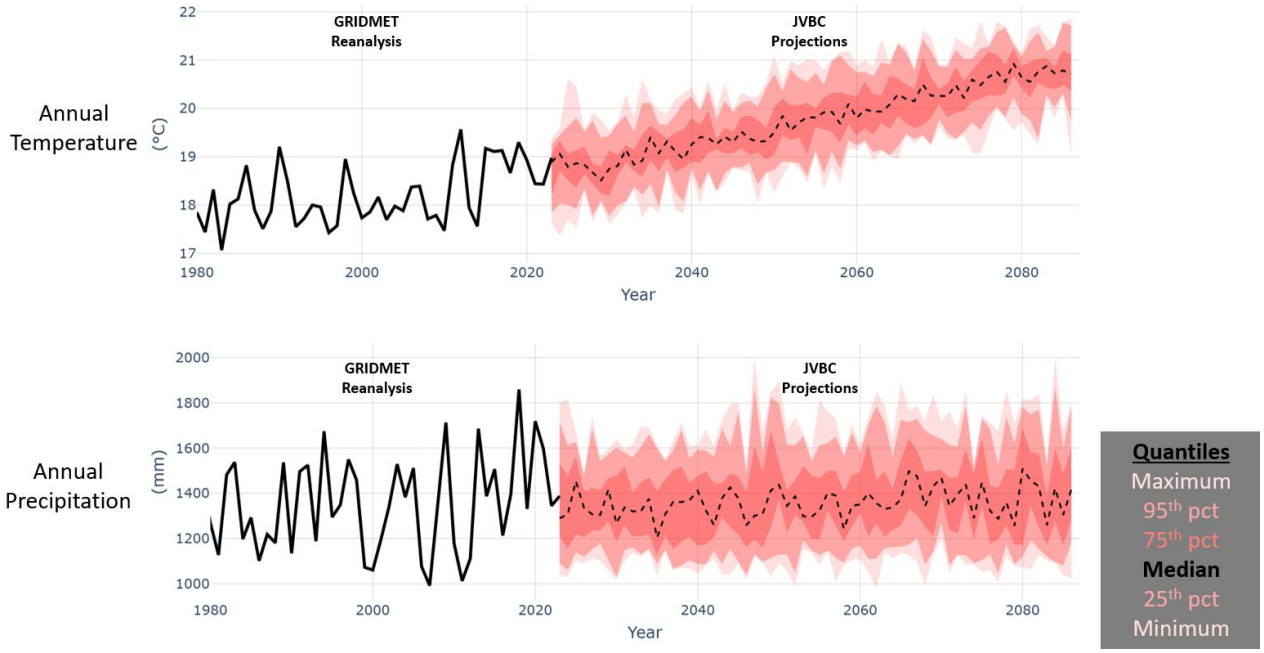
The JVBC method removed systematic biases while preserving the spatial cross-correlation structure of temperature and precipitation across the Apalachicola-Chattahoochee-Flint (ACF) River Basin. By mining historical fine-resolution weather records, this approach enabled robust regional climate-risk analyses and produced data products that support evidence-based decision-making in water-resource management and agriculture.

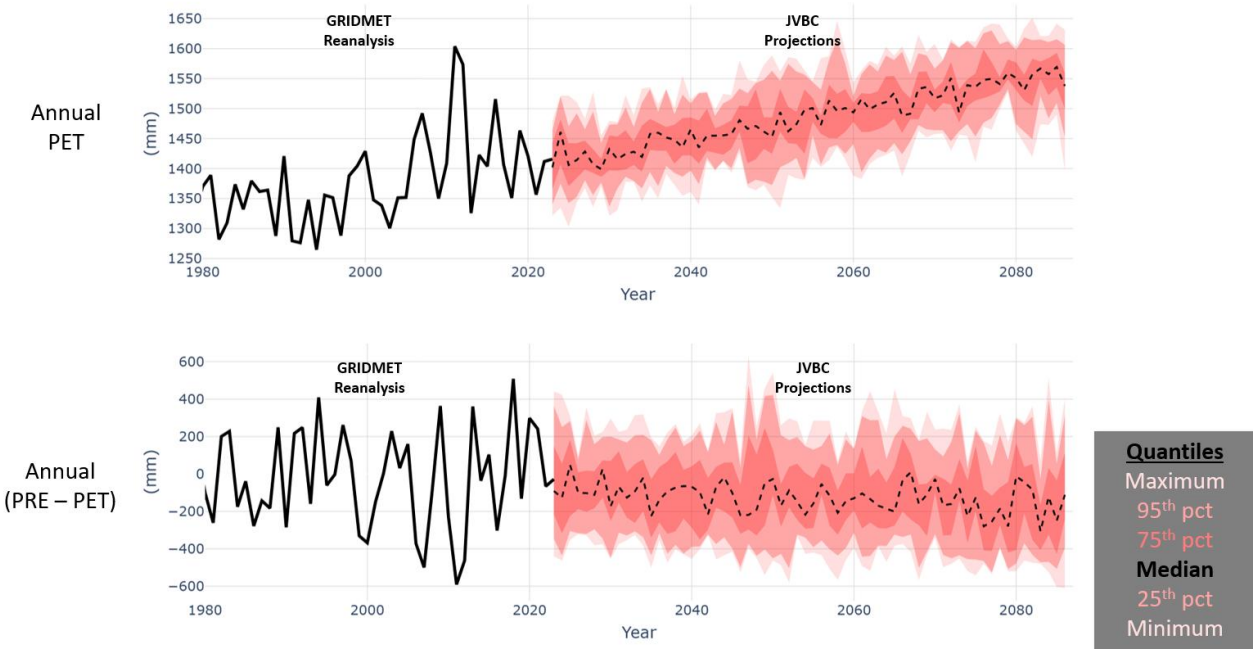


Improvement of Joint-CDF between Monthly Temperature and Precipitation after bias correction of CMIP6 Climate data. Bias corrected climate data (blue) more closely follows the historical correlation structure (red) than the raw, uncorrected data (gray).

Results: Projected Climate in the ACF

Time-series analysis indicates that temperatures across the ACF River Basin are projected to rise substantially, while precipitation will increase only slightly—insufficient to offset the temperature-driven rise in evapotranspiration.

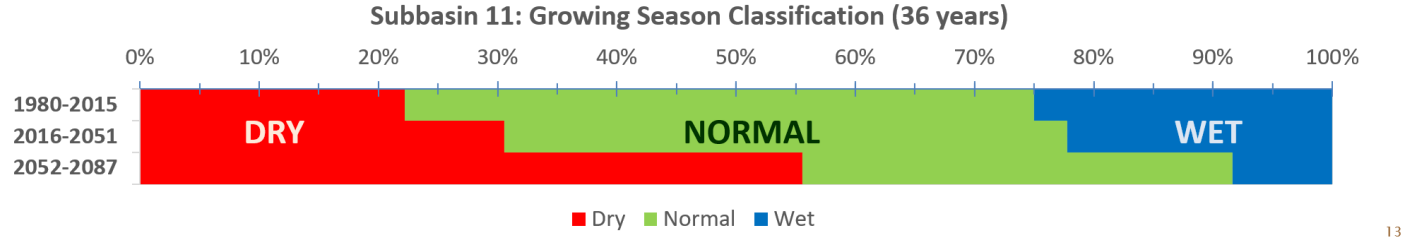




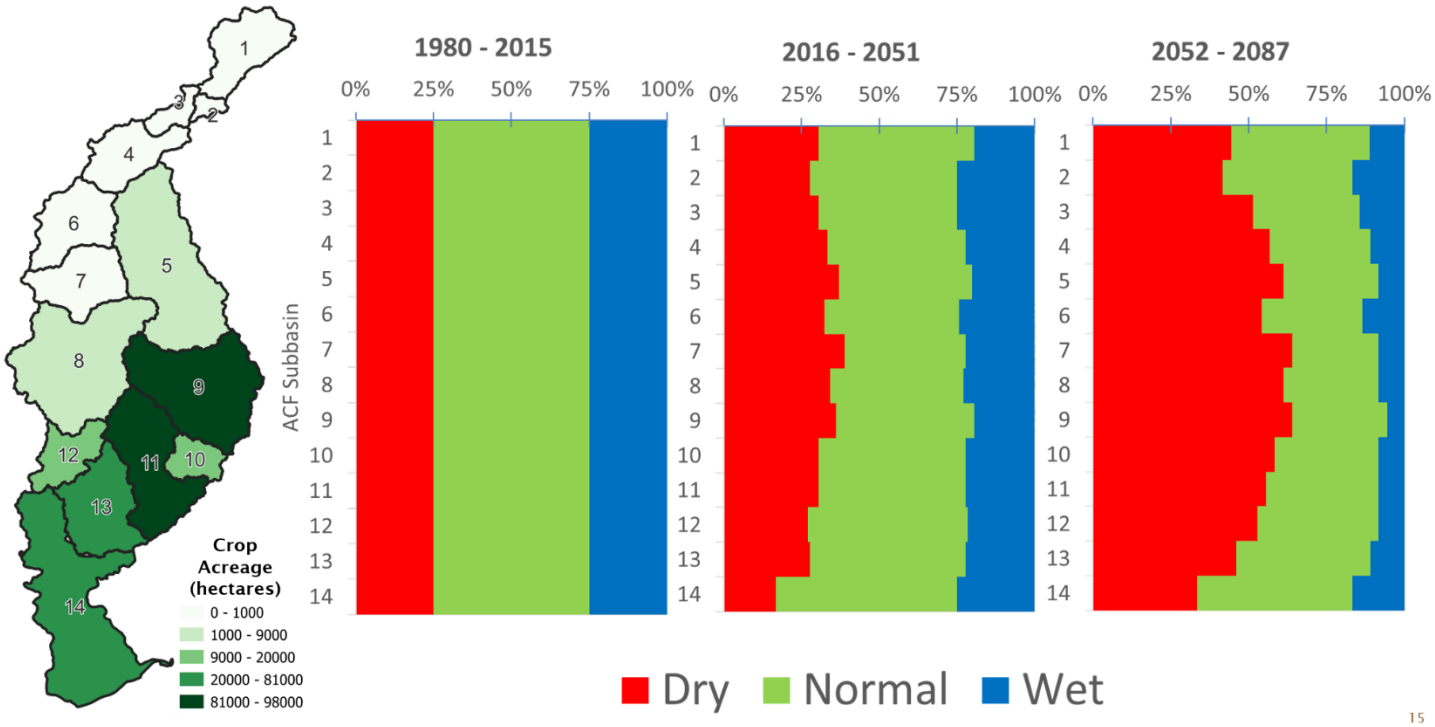
Results: Characterization of Growing Seasons in Heavily Cropped Sub-regions of the ACF

Analysis of historical and projected precipitation deficits—calculated as the difference between growing-season precipitation and potential evapotranspiration—revealed a strong link between climate, agricultural production, and irrigation demand. Findings show that agricultural droughts are expected to become **more severe, longer lasting, and more frequent**. By the end of the century, over 50% of growing seasons in multiple ACF sub-basins are projected to be classified as "dry."

	Dry				Normal				Wet			
	Corn	Cotton	Soybean	Peanut	Corn	Cotton	Soybean	Peanut	Corn	Cotton	Soybean	Peanut
Rain-fed Yield (% and kg/ha)	-30%	-12%	-24%	-18%	**6746**	**2923**	**2140**	**4189**	+35%	+7%	+19%	+12%
Irrigated Yield (% and kg/ha)	-2%	-5%	0%	-3%	**10660**	**3223**	**3052**	**5333**	-6%	+1%	+1%	-1%
Irrigation (% and mm)	+45%	+38%	+45%	+35%	**159**	**113**	**117**	**162**	-35%	-36%	-32%	-35%



13

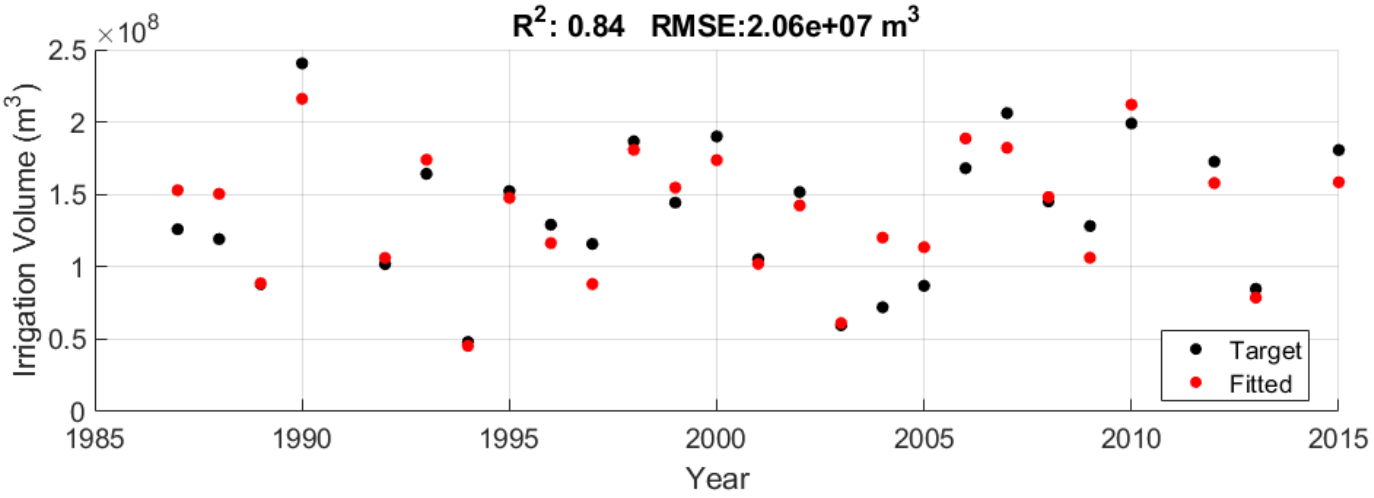


Results: Relating Irrigation Demand to Precipitation Deficit

Mining of historical data revealed a robust functional relationship between climate and irrigation demand in the ACF:

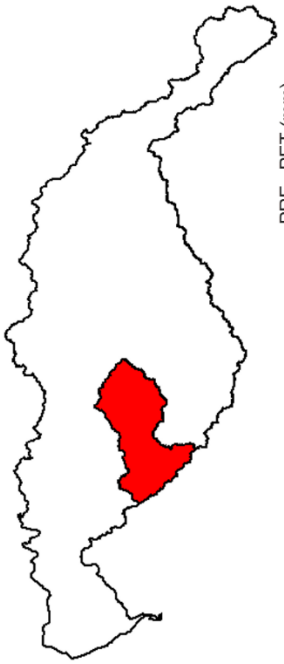
$$\text{Irrigation Volume} = \beta_0 + \beta_1 \cdot PET + \beta_2 \cdot PRECIP^2 + \beta_3 \cdot PRECIP$$

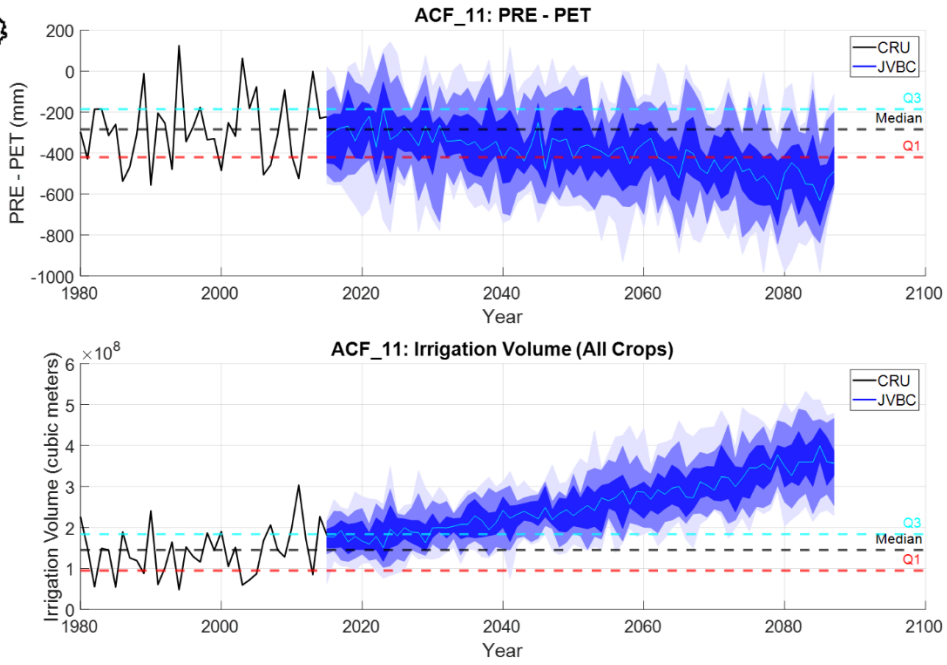
After estimating the regression parameters, future irrigation demand for the major crops—corn, cotton, peanut, and soybean—was projected using JVBC-derived estimates of potential evapotranspiration (PET) and precipitation (PRECIP). Comparable models were also developed for each sub-basin of the ACF.

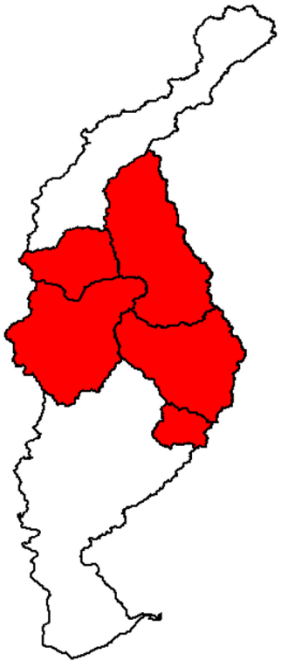


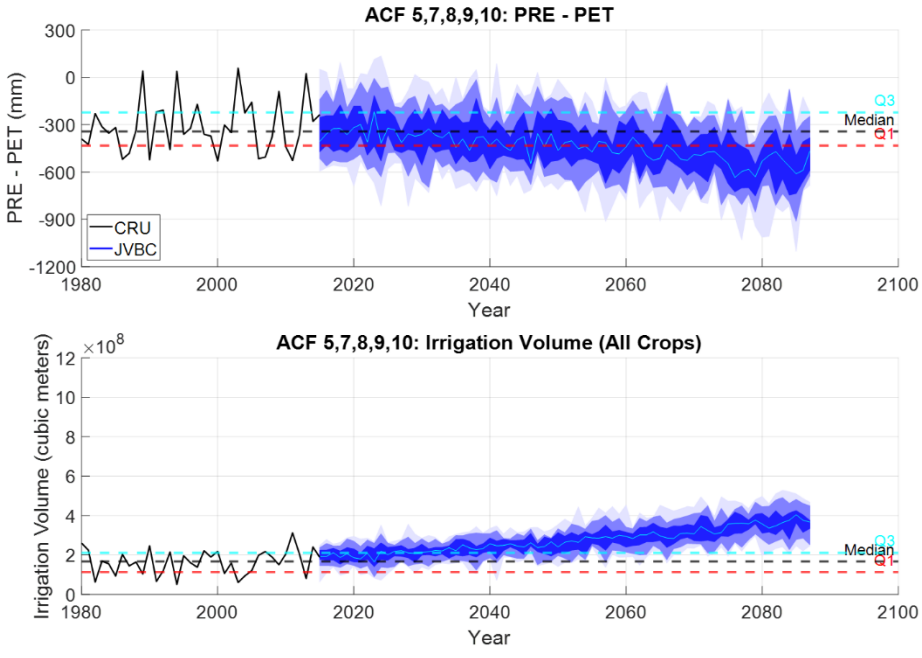
Performance of regression model to estimate ACF sub-basin irrigation demand as a function of precipitation and potential evapotranspiration.

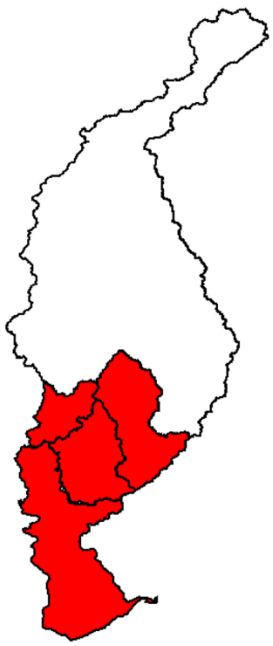
Results: Increased Irrigation Demand Projected for the ACF

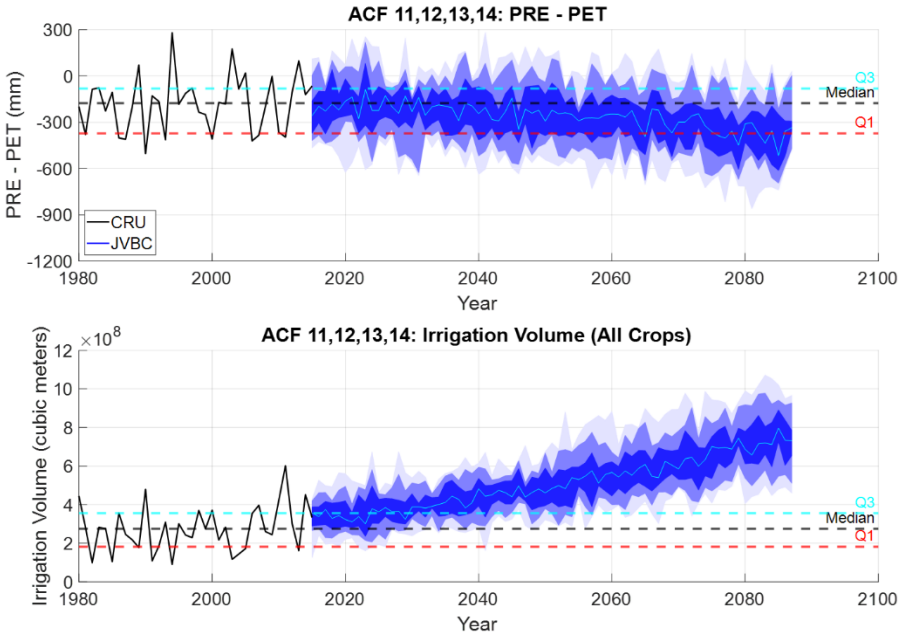












Major findings from analysis of irrigation demand include:

Central ACF: irrigation volume is expected to **increase by 30%** over the next 30 years and **100% by end-of-century**.

Southern ACF: irrigation volume is expected to **increase by 40%** over the next 30 years and **130% by end-of-century.**

The assessment results have significant implications for sustainable farming, water planning, and water policy in the ACF River Basin.

3 CONTINENTAL-SCALE CLIMATE BIAS-CORRECTION, PROJECTION, & INTERACTIVE DASHBOARD

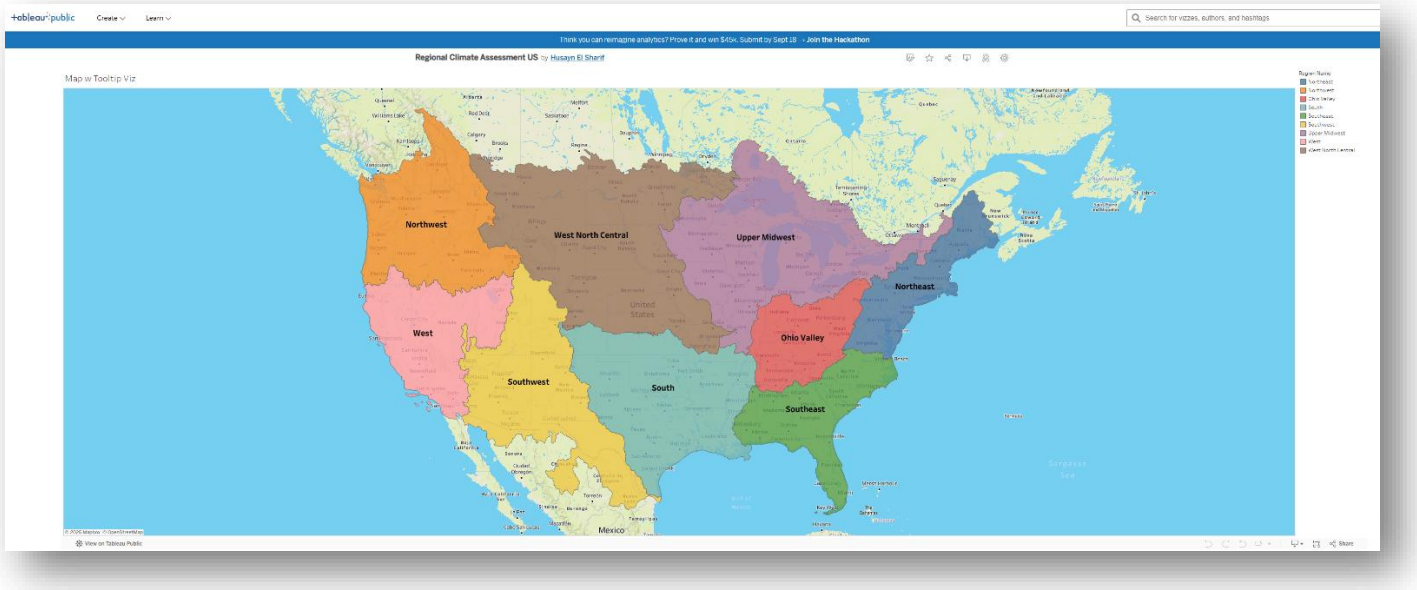
Skills: *Statistical Downscaling and Bias Removal, Time Series Analysis, Impact Assessments, Data Visualization, Tableau*

Background: This project analyzed air-temperature and precipitation projections from sixteen leading Global Climate Models (GCMs) across the continental United States. To make these projections more reliable for regional decision-making, a **Joint-Variable Bias Correction (JVBC) algorithm** was developed to remove systematic regional biases while preserving the natural statistical relationships between temperature and precipitation.

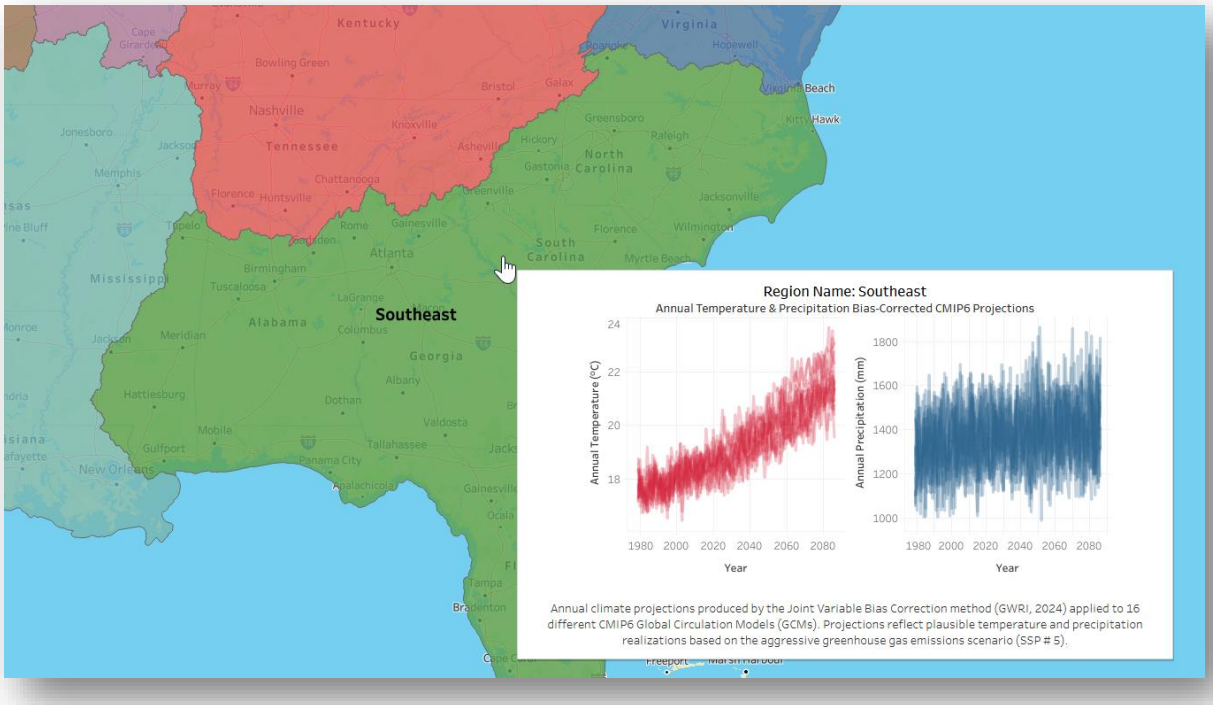
Data Processing & Results: Processing the multi-terabyte climate datasets required significant computing power, which was provided by **Georgia Tech's Phoenix supercomputer**, a 31,000-core cluster capable of 1.8 petaflops. After bias-correction across nine U.S. climate regions, the results were published in an **interactive Tableau dashboard** that allows users to explore the corrected projections.

The work delivers a robust foundation for accurate regional climate impact assessments and supports data-driven planning for climate mitigation and adaptation.

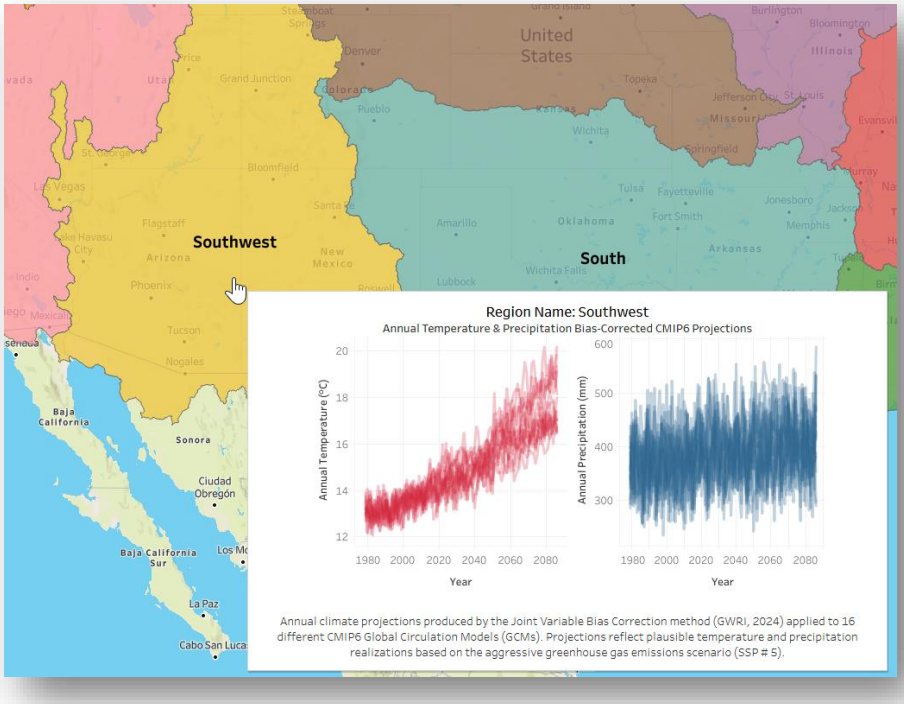
View the Dashboard



Interactive Tableau Dashboard for Regional Climate Assessment of the Continental U.S.



Interactive Tableau Dashboard: Climate Assessment for the Southeast U.S.



Interactive Tableau Dashboard: Climate Assessment for the Southwest U.S.

4 CUSTOMER CHURN MODELING & RETENTION OPTIMIZATION WITH SPARKML, MLFLOW, & NEURAL NETWORKS

Skills: *Customer Churn Modeling, Logistic Regression (SparkML), Neural Networks (TensorFlow), Databricks, MLflow Experiment Tracking, Class Imbalance Handling, Decision Threshold Optimization, Model Evaluation & Interpretation*

Background: This project focuses on predicting customer churn to guide retention strategies for a telecommunications provider serving over 7,000 home phone and internet customers. The dataset includes key demographic and service-related features along with a churn label indicating whether a customer left within the past month.

A **Logistic Regression model** was built to estimate the probability of churn for each customer, providing interpretable insights into the drivers of customer loss. To further explore predictive performance, the problem was also modeled using **neural networks**, enabling comparison between a highly interpretable approach and a more complex deep learning technique.

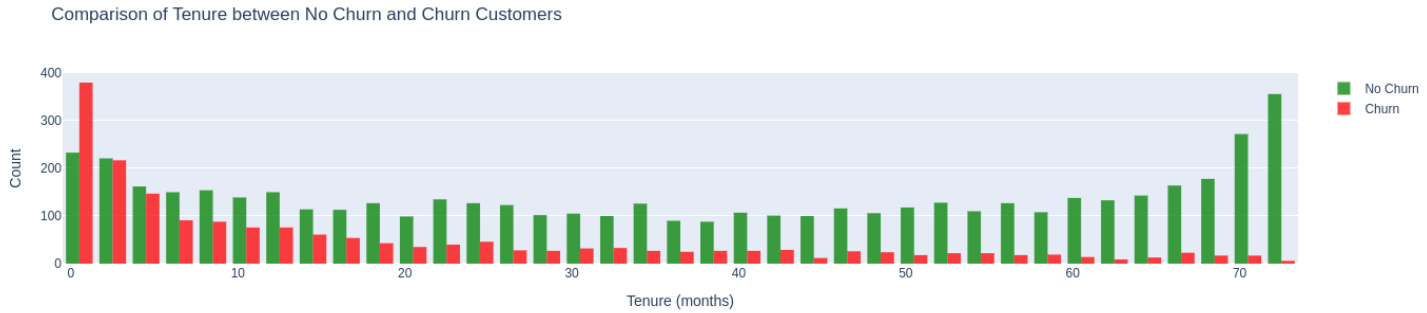
Data: The telco data includes the following attributes:

Attribute	Description
CustomerId	Customer Id
Gender	Gender (Male/Female)
SeniorCitizen	Whether the customer is elderly (1, 0)
Partner	Whether the customer has a partner (Yes, No)
Dependents	Whether the customer has dependents (Yes, No)
Tenure	Number of months the customer has been with the company (integer)
PhoneService	Whether the customer has phone service (Yes, No)
MultipleLines	Whether the customer has more than one line (Yes, No, No phone service)
InternetService	Whether the customer has internet service provider (DSL, Fiber optic, No)
OnlineSecurity	Whether the customer has online security (Yes, No, No internet service)
OnlineBackup	Whether the customer has online backup (Yes, No, No internet service)
DeviceProtection	Whether the customer has device protection (Yes, No, No internet service)
TechSupport	Whether the customer has technical support (Yes, No, No internet service)
StreamingTV	Whether the customer has streaming TV (Yes, No, No Internet service)
StreamingMovies	Whether the customer streams movies (Yes, No, No internet service)
Contract	Customer's contract term (Month-to-month, One year, Two years)
PaperlessBilling	Whether the customer has paperless billing (Yes, No)
PaymentMethod	Electronic check, Postal check, Wire transfer, Credit card
MonthlyCharges	The amount charged to the customer monthly
TotalCharges	The total amount charged to the customer
Churn	Whether the customer uses (Yes or No)

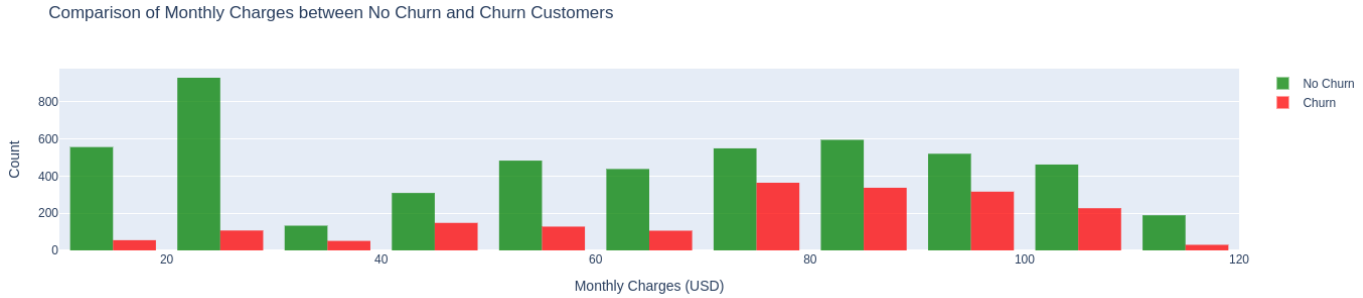
	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	Yes	No	No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	No	No	No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes

Telco DataFrame, first 5 rows

Preliminary analysis of data showed that many customers who leave, do so after the first month of service.



Customers facing relatively higher monthly charges ranging from $70 to $110 are more likely to churn than those paying less.



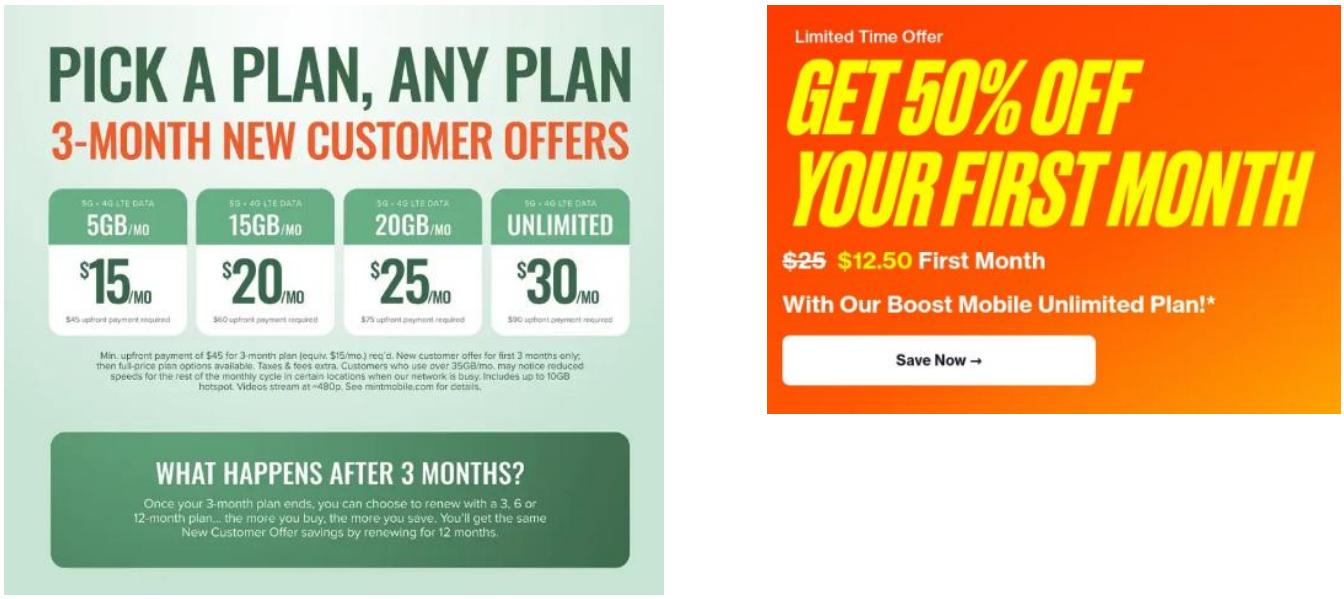
Data preprocessing for the Logistic Regression and Neural Network models included extensive cleaning and conversion of all features—both numerical and categorical—into numeric form. Because customers who churned represent a **minority class**, special handling was required to avoid biased predictions. To address class imbalance, both resampling techniques (**SMOTE: Synthetic Minority Over-sampling Technique**) and probability-threshold-based strategies were evaluated, with threshold tuning used to align model predictions with retention-focused objectives.

Logistic Regression: The data was randomly shuffled and split into **80% training** and **20% testing** sets, with stratification by the churn label to preserve class proportions. The resulting model achieved the following performance metrics:

Logistic Regression Model Performance Metrics

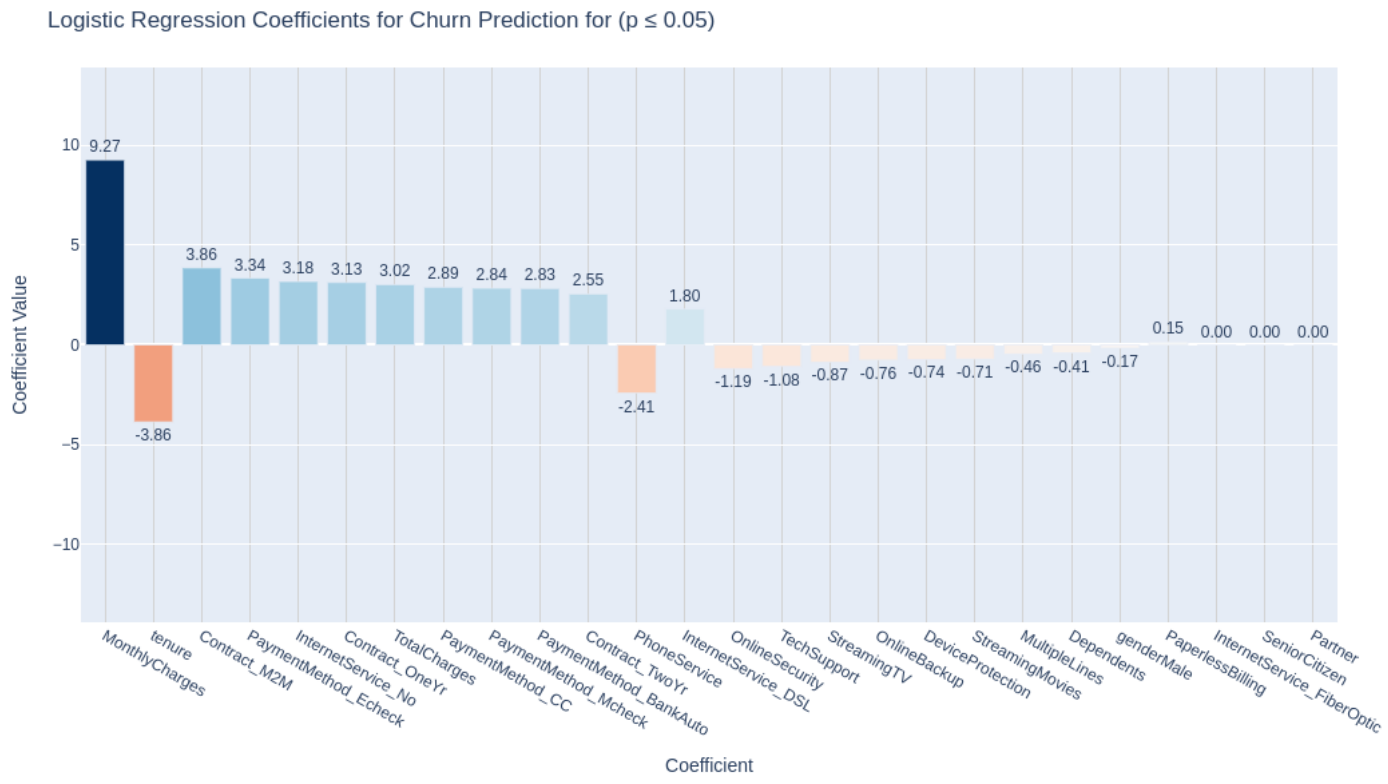
Class Label	Precision	Recall	F1-Score
No Churn	89%	79%	0.84
Churn	56%	74%	0.63
Overall Accuracy	77%		

Performance metrics demonstrate the model reliably predicts customer churn, providing actionable insights to guide data-driven retention strategies.



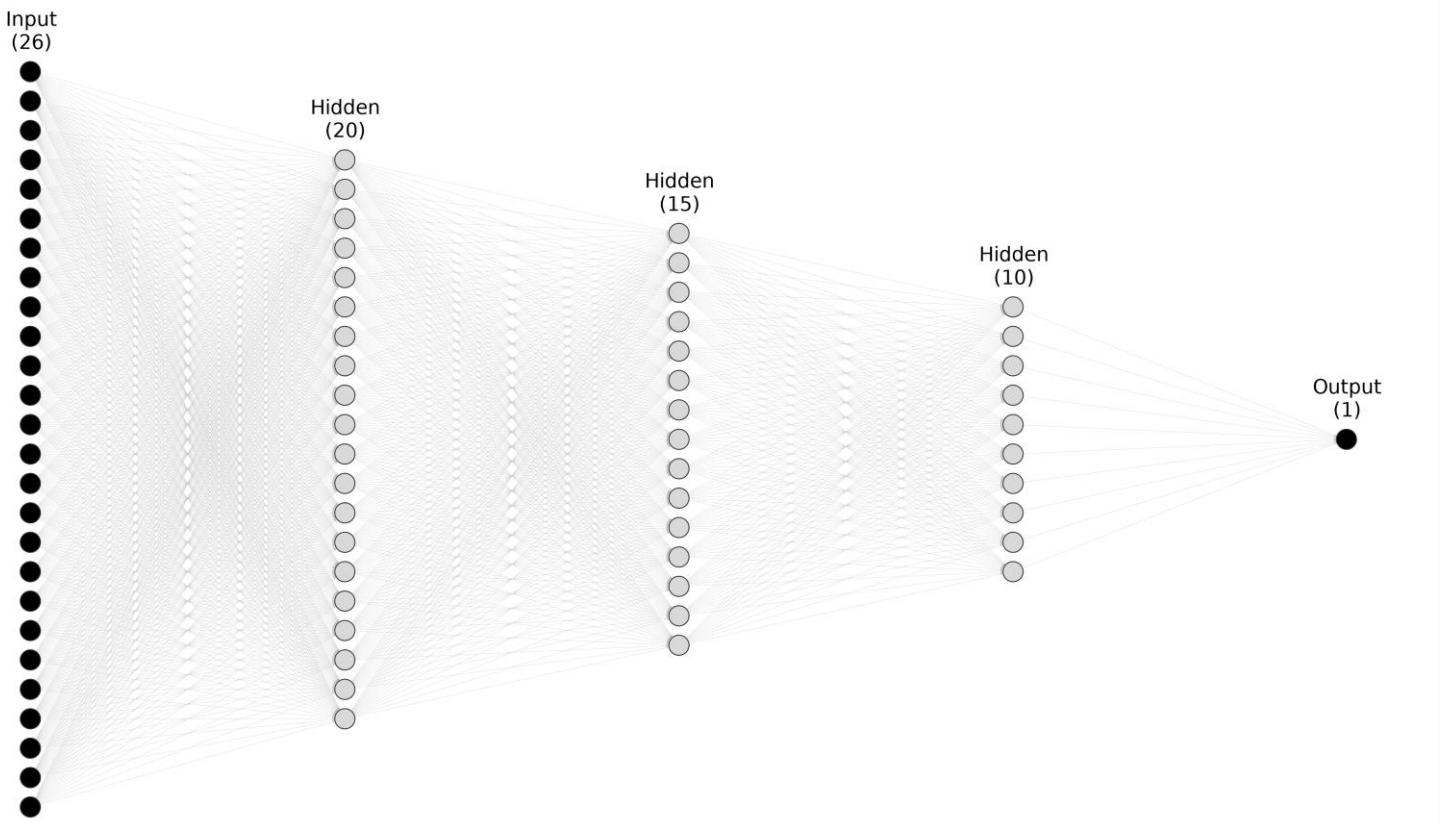
Example Telco Introductory Offers to improve customer retention

Logistic Regression analysis identified **Monthly Charges** as the strongest predictor of customer churn. Customers on **month-to-month contracts** are more likely to leave, while those with **longer tenure** are less likely. These insights suggest targeting **new customers paying over $70/month on month-to-month plans** with discounted introductory rates and incentives to switch to long-term contracts. Such strategies can help retain customers early, giving more time to build satisfaction and loyalty.



Ranking of Most Impactful Customer Attributes influencing Churn

Neural Network: A TensorFlow neural network was built using the same telco dataset, featuring one input layer, three hidden layers, and an output layer that predicts the probability of customer churn.

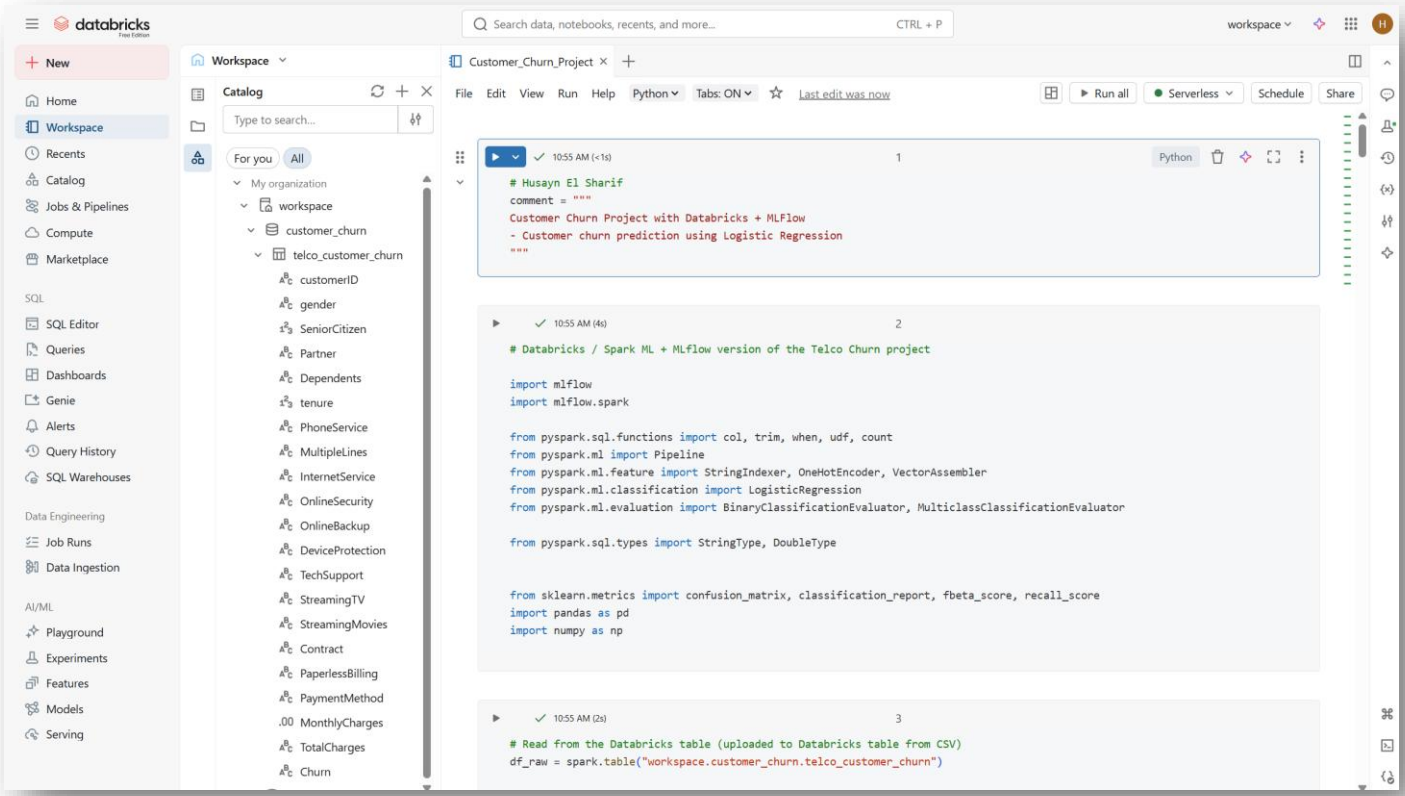


Neural Network Diagram for Customer Churn Prediction

The Neural Network performed on par with the Logistic Regression model, showing that both can reliably predict customer churn and deliver actionable insights. However, Logistic Regression offers greater interpretability, clearly revealing how individual customer attributes influence churn—an advantage over the Neural Network's "black box" nature.

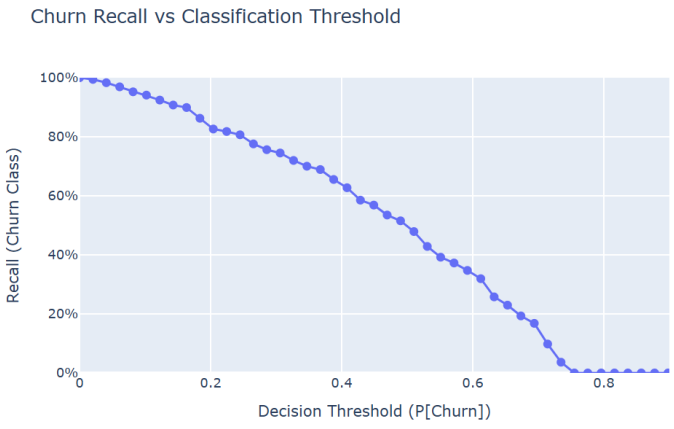
Neural Network Model Performance Metrics

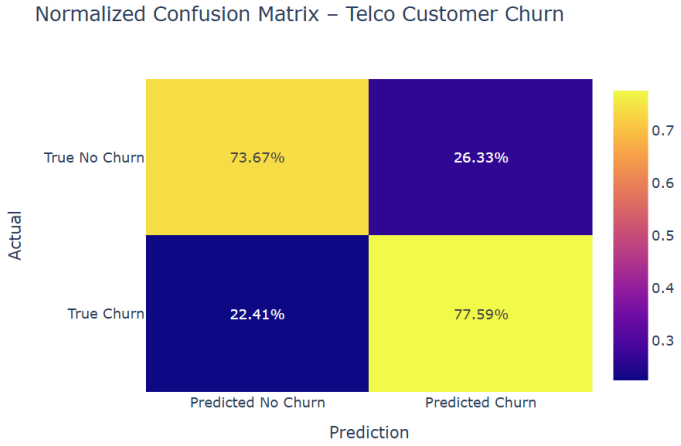
Class Label	Precision	Recall	F1-Score
No Churn	89%	81%	0.85
Churn	58%	72%	0.64
Overall Accuracy	79%		



Databricks & MLflow Scalable Implementation

Databricks & MLflow Implementation (Scalable ML Extension): This project was extended using **Databricks** and **MLflow** to demonstrate scalable, production-oriented churn modeling. Customer data was ingested as a managed Databricks table and modeled with a **SparkML Logistic Regression pipeline**. MLflow was used to track experiments and evaluation metrics. To address class imbalance, the classification threshold was tuned to the baseline churn rate (~26%), improving churn recall from 50% to 78%. Interactive **Plotly** visualizations were used to analyze threshold–recall trade-offs and model behavior.





Performance Visualizations for SparkML Logistic Regression Model for Customer Churn Prediction

5 AI-Automated Lead Generation Pipeline for Engineering Consultancy (RAG + Web Data)

Skills: *Artificial Intelligence, Prompt Engineering, Retrieval-Augmented Generation (RAG), Web Scraping, APIs, Large Language Models (LLMs), Data Mining, JupyterLab*

Background: A full-service engineering technology consulting firm sought higher-quality leads to market its services. This project developed an AI-driven workflow to mine publicly available documents and websites, identify "ideal customers," and enrich them with contact information—laying the groundwork for a targeted outreach campaign.

Tools & Technologies:

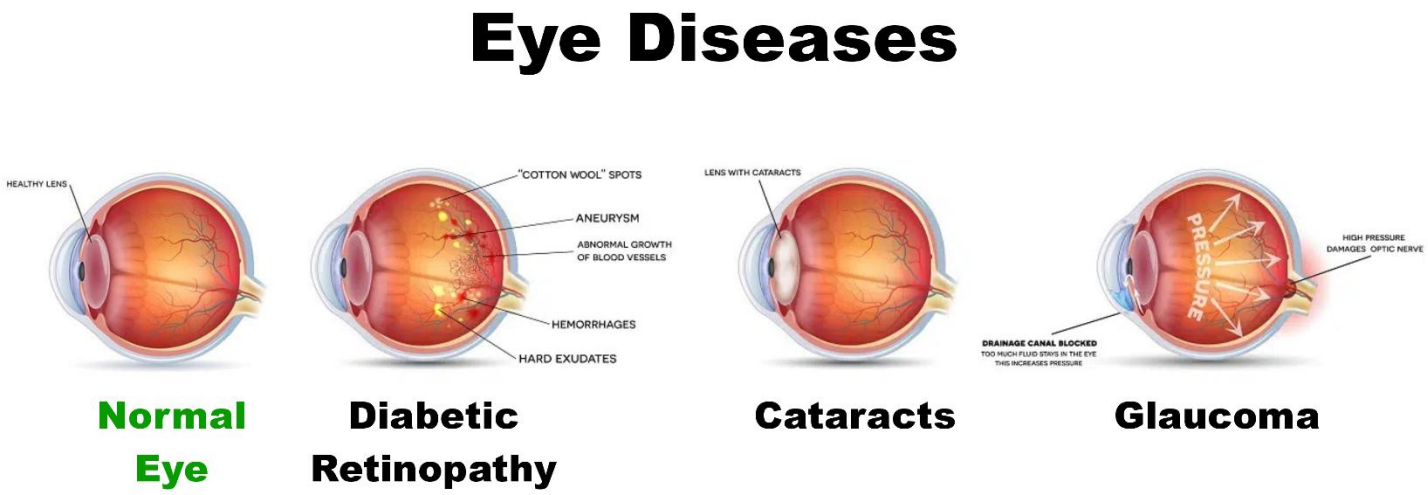
- Google Gemini Generative AI API
- Google Places API
- Selenium webscraping
- Pandas & Markdownify for data handling
- TQDM for process tracking

Key Highlights:

- **Automation:** Extracted and structured website text with Selenium, converting content into Markdown for downstream AI analysis.
- **Data Enrichment:** Appended prospective client records with website and phone details using Google Places API.
- **AI Analysis:** Applied Gemini to evaluate prospects against the firm's service criteria, fine-tuned through iterative prompt engineering.
- **Customizable Models:** Flexible design supporting multiple Gemini families (*gemini-2.0-pro, gemini-2.5-pro, gemini-2.5-flash-lite*), with options to adapt to OpenAI or other LLMs.
- **Client-Centric Design:** Tailored to align lead generation with the firm's target market and service offerings.

6 AI-Powered Retinal Disease Detection with Deep Learning & Computer Vision Pipeline

Skills: *Image Classification, Computer Vision, Deep Learning, Transfer Learning, TensorFlow*



Background: Diabetic eye diseases—including cataracts, diabetic retinopathy, and glaucoma—are leading causes of preventable blindness. Early detection through retinal fundus imaging can significantly improve clinical outcomes, but manual diagnosis is time-consuming, subjective, and requires specialized expertise. Advances in deep learning and transfer learning offer scalable opportunities to assist clinicians by automating the detection of disease patterns in retinal images. In this project, I developed a **TensorFlow EfficientNet-based computer vision model** capable of classifying retinal images into four categories: **cataract, diabetic retinopathy, glaucoma, and normal**. The analysis used **4,000+ professionally resized (456×456 pixel) fundus images**, enabling the model to learn high-resolution structural features associated with each disease (and normal) class.

In detail, diabetic eye diseases of focus in this project are:

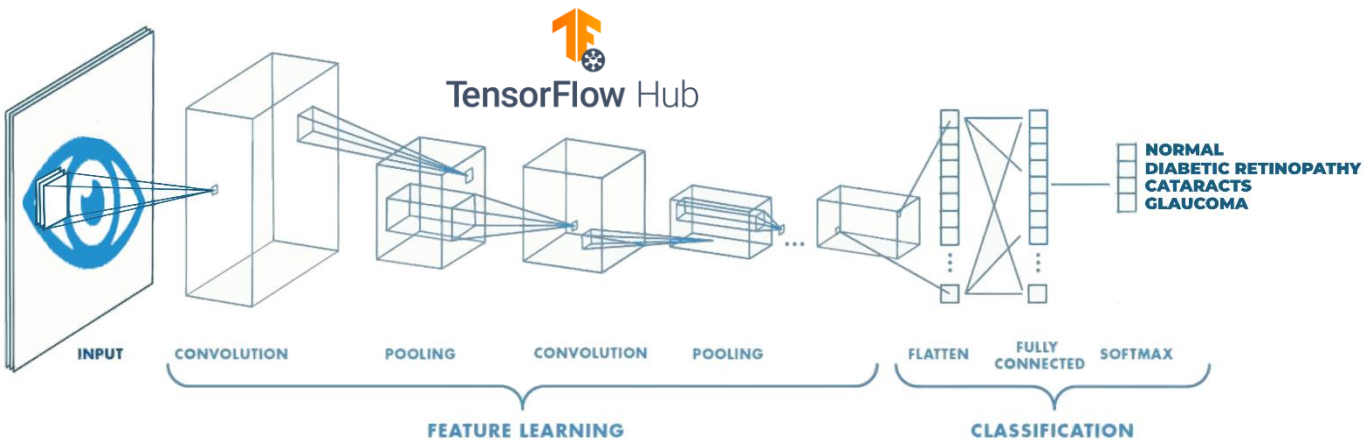
- *Diabetic retinopathy:* The persistently high blood sugar levels that occur with diabetes can damage the retina's small blood vessels (capillaries), which deliver oxygen and nutrients. Diabetic retinopathy affects up to a third of people with diabetes over the age of 50.

- *Cataracts*: A cataract is a clouding of the lens in the eye. Left untreated, cataracts can eventually lead to blindness. People with diabetes are more likely to develop cataracts at an earlier age and suffer visual impairment faster than those without the condition.

- *Glaucoma*: This is a group of conditions that can damage the optic nerve. The optic nerve transmits signals from the retina to the brain for processing. Glaucoma is often a result of increased pressure inside the eye. The risk of glaucoma in people with diabetes is significantly higher than that of the general population.

Problem Statement: How can modern **deep learning** and **transfer learning** techniques be applied to retinal fundus images to accurately classify the presence of common diabetic eye diseases—cataracts, diabetic retinopathy, glaucoma—or confirm that an eye is normal? More specifically:

- Can an **EfficientNet**-based model achieve clinically meaningful accuracy?
- How well can the model differentiate between diseases with **subtle morphological differences**, such as glaucoma?
- Can this approach reduce manual screening burden and improve early detection efficiency?



Schematic of Tensorflow Computer Vision Pipeline from Retinal Image to Normal/Eye Disease Classification

Model Development: A complete end-to-end deep learning pipeline was built in **TensorFlow** and **Keras** using **EfficientNet-B5 (feature-vector version)** from **TensorFlow Hub**. The pipeline was developed in a Jupyter notebook.

The notebook first ingests all images, extracts labels from metadata, and ensures balanced representation across the four classes. Images are then **normalized, padded to square aspect ratio to avoid distortion, and resized to 456×456**, the recommended input size for EfficientNet-B5. After shuffling the data, the first 3,200 samples were split into **train (80%), validation (20%)**. A **hold-out test set** consisted of all images beyond the initial 3,200 sample allocation (approximately 800 images).

To improve training stability and leverage transfer learning:

- The **EfficientNet-B5** backbone is used as a fixed feature extractor.
- A dense **softmax** classification layer is added for 4-class prediction.
- The model is trained using **categorical cross-entropy** and **Adam optimizer**.
- Training includes **EarlyStopping**, **ModelCheckpoint**, and **TensorBoard tracking** for robust monitoring.
- A batch pipeline is implemented using tf.data for efficient GPU processing.

The model is trained for up to **50 epochs**, with early stopping triggered when validation accuracy stops improving. The best-performing checkpoint is automatically saved and later re-loaded for evaluation and inference.

Sidebar: Why EfficientNet?

The EfficientNet-B5 feature vector is a condensed, numerical representation (a "fingerprint") of an image, capturing its most important visual characteristics after being processed by the EfficientNet-B5 convolutional neural network (CNN). It transforms complex image information into a fixed-size array of numbers that machine learning models can use for tasks like classification or similarity search. EfficientNet models, including B5, are known for a unique compound scaling method that balances network depth (number of layers), width (number of channels), and image resolution in a systematic way. This approach results in a model that achieves **high accuracy with fewer computational resources** compared to older architectures like ResNet or VGG.

Model Deployment & Results: After training, the best model checkpoint achieved **strong generalization performance** on the unseen test set:

Accuracy: 0.89

Loss: 0.33

Macro F1 Score: 0.88

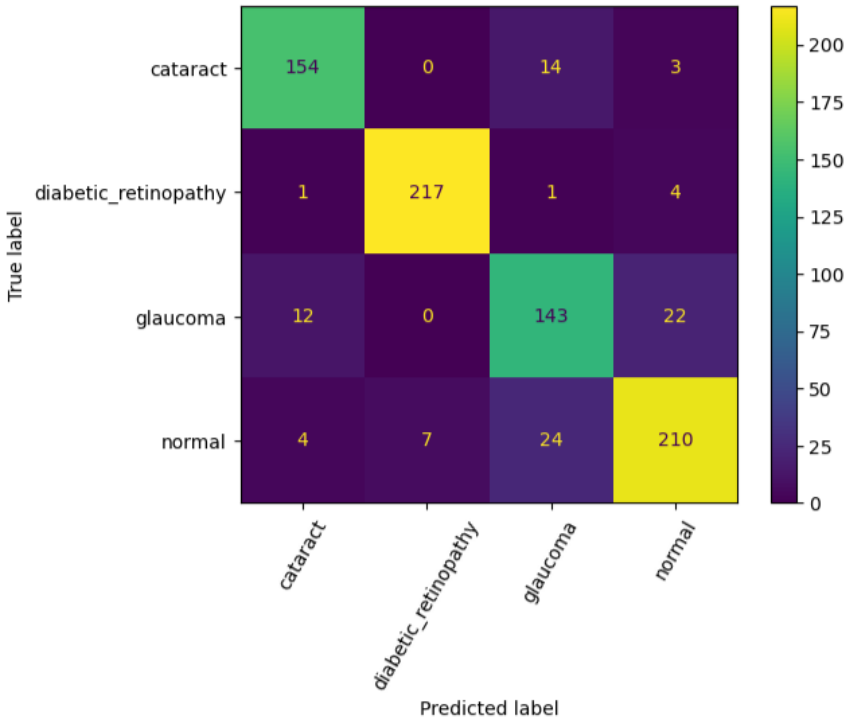
Precision (per class):

- Cataract: 0.91
- **Diabetic Retinopathy: 0.97**
- Glaucoma: 0.82
- Normal: 0.89

Recall (per class):

- Cataract: 0.89
- **Diabetic Retinopathy: 0.96**
- Glaucoma: 0.80

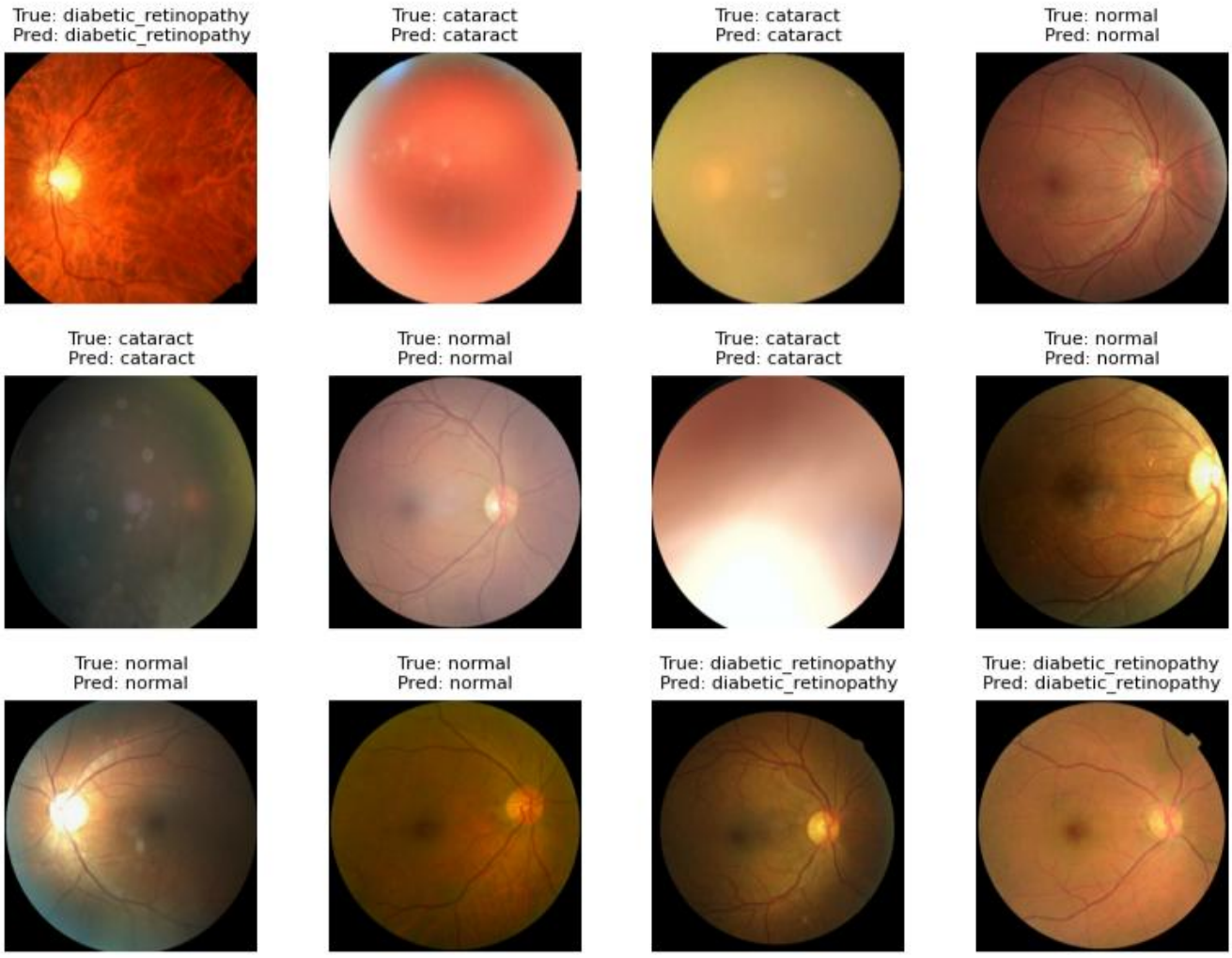
Normal: 0.83



Test (Hold-out) Confusion Matrix highlighting the predictive skill of the Image Classification Model

These results indicate a highly effective classifier, particularly for **diabetic retinopathy**, which is detected with outstanding precision and recall. Glaucoma—expectedly the most subtle class—shows the lowest metrics but still achieves meaningful sensitivity. Comprehensive confusion matrices and classification reports were generated for the train, validation, and test sets, confirming stable, non-overfit behavior.

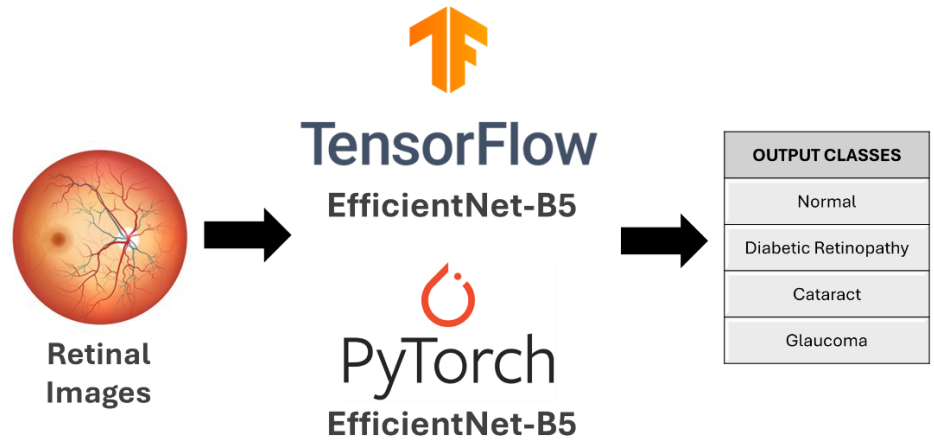
The model supports full-dataset inference, generating labeled prediction tables and enabling visualization of true vs. predicted labels. This creates a foundation for clinical validation studies or integration into screening-support tools.



Example Retinal Fundus Images from Test data with true and predicted labels

Actionable Insights:

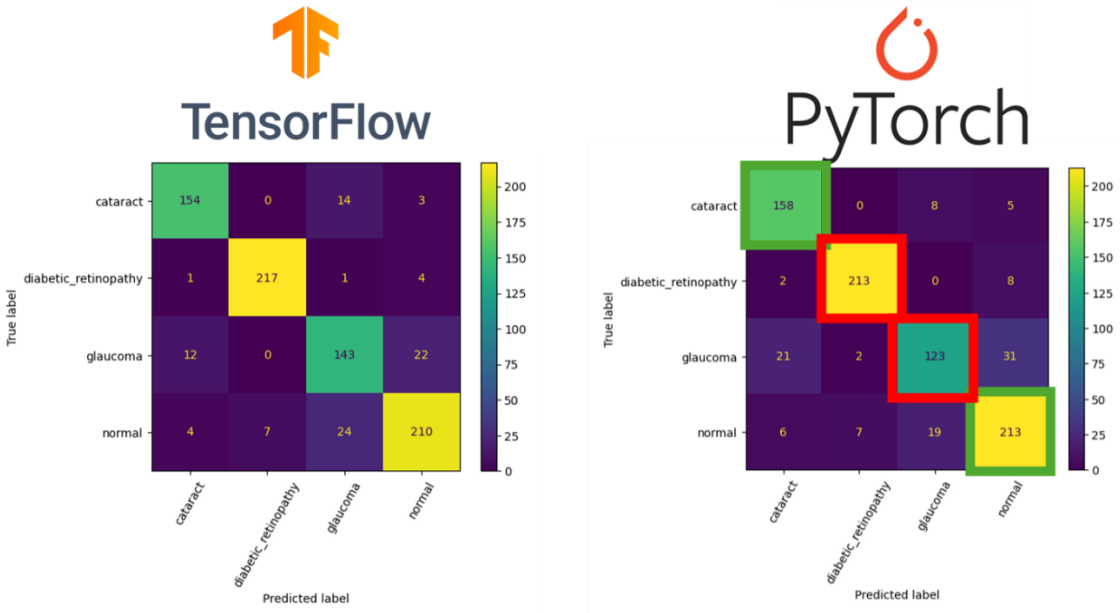
- **Clinical Triage Assistance:**
  The model identifies high-risk images with near-90% accuracy and excellent F1 scores, making it a promising foundation for a clinical decision-support system to assist ophthalmologists.
- **Improved Diabetic Retinopathy Screening:**
  DR detection is exceptionally strong (precision 0.97, recall 0.96), suggesting immediate value for large-scale community screening where DR prevalence is high.
- **Targeted Model Refinement for Glaucoma:**
  Since glaucoma achieved the lowest recall (0.80), performance could be improved by augmenting the glaucoma dataset, applying contrast-enhancement techniques, or fine-tuning the EfficientNet layers.
- **Scalability for Real-World Deployment:**
  Because the model uses TensorFlow and TF-Hub EfficientNet, it can be embedded into mobile/edge devices, cloud APIs, web-based screening tools, and long-term monitoring dashboards.
- **Foundation for Automated Ophthalmic Workflows:**
  With further validation, this model can help automate routine retinal screenings, reduce clinician workload, and enable earlier detection of vision-threatening conditions.



Dual Implementation (TensorFlow and PyTorch) of eye disease diagnosis using Computer Vision

Additional Implementation – PyTorch Version (EfficientNet-B5): To demonstrate versatility across deep learning frameworks, I also developed a full **PyTorch implementation** of the same retinal-disease classifier using **EfficientNet-B5**. This implementation mirrors the TensorFlow workflow with preprocessing pipelines, data loaders, and training loops adapted to the PyTorch ecosystem. The notebook integrates custom Dataset and DataLoader classes, GPU-accelerated mixed-precision training, image resizing and rescaling, and a fine-tuned EfficientNet-B5 backbone. Same as the TensorFlow implementation, the dataset included 4,016 retinal images across four balanced classes—cataract, diabetic retinopathy, glaucoma, and normal—split into 2,560 training samples, 640 validation samples, and 816 held-out test cases. After training with Adam and CrossEntropyLoss, the PyTorch model achieved strong generalization (comparable to the TensorFlow implementation) on the test set, producing a **macro F1 score of 0.86**, with particularly high recall for diabetic retinopathy and cataract classifications.

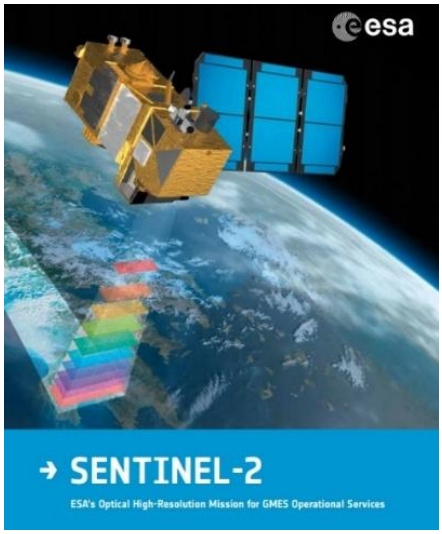
Re-implementing the model in PyTorch provides a framework comparison and shows the ability to translate data pipelines, augmentation steps, and EfficientNet architectures between libraries.

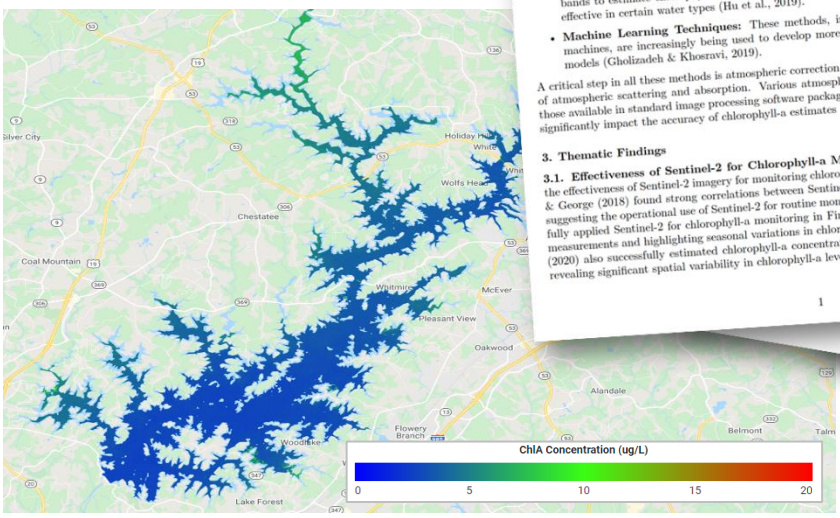


Comparison of Test (Hold-out) Set Confusion Matrices between TensorFlow and PyTorch implementations. Results are comparable; however, TensorFlow overall has superior classification performance.

7 Multi-Agent GenAI System for Automated Hydrology Literature Review (CrewAI + LLMs)

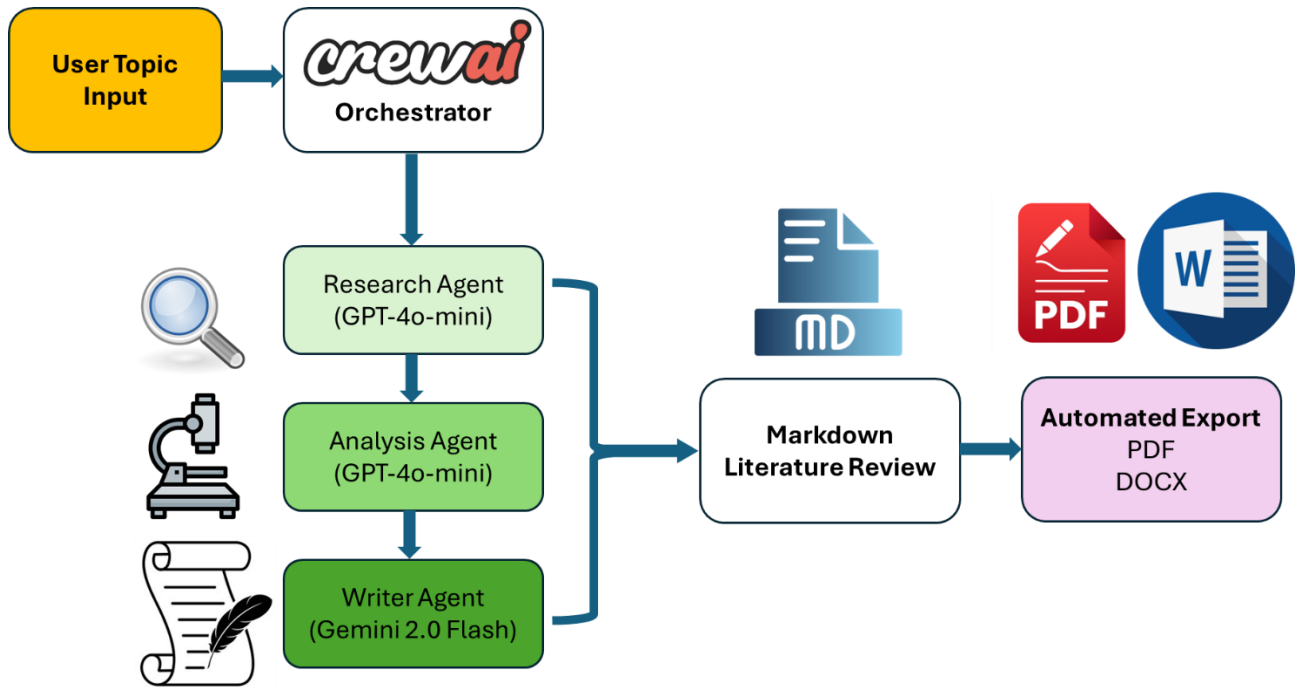
Skills: *Multi-Agent Systems, GenAI, Prompt Engineering, LLM Orchestration, Google Gemini, OpenAI GPT, LangChain, CrewAI, Document Parsing, Research Automation*





Background: Preparing hydrology literature reviews is traditionally a **manual, time-intensive process**, often taking researchers **several days** to search journal databases, extract key findings, synthesize themes, and produce a clean, publication-ready narrative. With the rapid growth in hydrology literature, manual synthesis has become an increasingly heavy bottleneck for scientific workflows, grant proposals, and decision-support products. This project addresses that bottleneck through an **autonomous, multi-agent GenAI system** built entirely in Python.

Problem Statement: *How can we automate the end-to-end hydrology **literature review workflow**, including search, reading, analysis, and writing, to reduce a 7-day manual research task to a 1-day **GenAI-enhanced** process while maintaining academic quality and scientific rigor?*



Schematic of the CrewAI Multi-Agent Pipeline for Hydrology-focused Literature Review

System Development (CrewAI Multi-Agent Pipeline): A custom **3-agent CrewAI pipeline** was designed to autonomously generate hydrology literature reviews on any user-defined topic. The system integrates:

- **OpenAI GPT-4o-mini** for structured research search + analysis
- **Google Gemini 2.0 Flash** for synthesis and professional scientific writing
- **LLM tool orchestration** for extensible workflows leveraging **CrewAI** and **LangChain**
- Markdown cleaning + formatting logic
- Automated PDF/DOCX generation for publishing-ready outputs

The three agents are described as follows:

1. Research Agent — Hydrology Journal Search

The Research Agent uses domain-specific prompts tailored to water resources, hydrology, remote sensing, and climate modeling to search respected peer-reviewed sources such as *Water Resources Research*, *Journal of Hydrology*, *Hydrology and Earth System Sciences (HESS)*, *Hydrological Processes (HP)*, *Journal of the American Water Resources Association (JAWRA)*, and *Water*. From these sources, it extracts structured metadata including full citations and DOIs, the hydrology or water-resources focus of each study, the methodological approach used, the study region, and any associated climate scenario information. The Analysis Agent then performs deep reading and thematic coding of the retrieved papers. It extracts key elements such as study objectives, data sources, models applied, spatial and temporal scales, major findings, and noted uncertainties or limitations. Based on this detailed review, it produces multi-paper thematic synthesis sections appropriate for publication-quality literature reviews.

2. Analysis Agent — Deep Reading & Thematic Coding

The Analysis Agent performs deep reading and thematic coding of the research outputs generated by the system. It extracts the essential elements of each paper, including the study's objectives, the data sources used, the modeling approaches applied, relevant spatial and temporal scales, key findings, and any stated uncertainties or limitations.

After reviewing all materials, the agent synthesizes these components across multiple papers to produce cohesive, publication-ready thematic summaries that highlight trends, gaps, and converging insights in the literature.

3. Writer Agent — Scientific Literature Review Generator

The Writer Agent functions as a scientific literature review generator powered by Google Gemini, providing diversity in LLM reasoning and reducing reliance on a single model. It produces publication-quality reviews that include well-structured section headings—such as Background, Methods, Findings, and Knowledge Gaps—along with properly formatted in-text citations and a complete bibliography. The agent composes the full review in Markdown, which is then automatically exported into both PDF and Microsoft Word formats for seamless integration into reports, proposals, and academic documentation.

Results & Impacts: The multi-agent workflow is capable of generating a wide range of hydrology literature reviews, including topics such as climate impacts on low-flow hydrology, Sentinel-2–based chlorophyll-a monitoring, drought modeling in the southeastern United States, and remote sensing approaches for assessing water quality. Each review synthesizes between 12 and 25 peer-reviewed sources and incorporates thematic analysis, a complete bibliography, and a polished narrative suitable for academic reports, technical briefs, or grant proposals. The entire end-to-end pipeline executes in under a day—reducing manual effort from roughly seven days and achieving an approximate 85% improvement in research preparation efficiency. Overall, this project demonstrates how multi-agent GenAI systems can transform environmental science workflows. By integrating CrewAI with OpenAI and Gemini models, the system converts a traditionally labor-intensive research task into an automated, scalable, and academically rigorous pipeline.

Code Excerpts:

```python
# import libraries
import os
import json

from dotenv import load_dotenv, find_dotenv

from langchain_openai import ChatOpenAI

import google.generativeai as genai
from langchain_google_genai import ChatGoogleGenerativeAI

from crewai import Agent, Crew, Task, Process, LLM

from IPython.display import display, Markdown

import pypandoc
```

```python
# define the LLMs

# For research + analysis
research_analysis_llm = ChatOpenAI(
    model_name="gpt-4o-mini",
    temperature=0.15,
    openai_api_key=openai_api_key,
)

# For writing (Gemini)
writer_llm = LLM(
    model="gemini/gemini-2.0-flash",  # note the gemini/ prefix
    temperature=0.3,
    verbose=True,
    google_api_key=google_api_key,
)
```

```python
# Define the three agents

# --- Researcher Agent ---
researcher_agent = Agent(
    role="Hydrology Literature Researcher",
    goal=(
        """
        Identify recent, high-quality, peer-reviewed literature in hydrology and water resources that is highly relevant to the user's topic.
        """
    ),
    backstory=("""
        You are a hydrology and water resources researcher with deep familiarity with journals like
        Water Resources Research, Journal of Hydrology, Hydrology and Earth System Sciences (HESS), WRR, JAWRA, etc.
        You are meticulous about selecting peer-reviewed, reputable sources and capturing full citation details."
        """
        ),
    llm=research_analysis_llm,
    verbose=True,
    allow_delegation=False,
)
```

```
# --- Analysis Agent ---
analysis_agent = Agent(
    role="Hydrology Paper Analysis Specialist",
    goal=(
        """
        Read and analyze the selected journal articles and extract structured,
        paraphrased information that will serve as raw material for a literature review.
        """
    ),
    backstory=("""
        You specialize in hydrologic modeling and water-resources synthesis.
        You are skilled at extracting key details from papers: study objectives,
        hydrologic models used, data sources (in-situ, satellite, reanalysis),
        spatial/temporal scales, key findings, uncertainties, and limitations.
        """
    ),
    llm=research_analysis_llm,
    verbose=True,
    allow_delegation=False,
)
```

```
# --- Writer Agent ---
writer_agent = Agent(
    role="Hydrology Literature Review Writer",
    goal=(
        """
        Write a coherent, publishable-quality literature review on the user's topic
        using the analyzed material, with proper citations and a bibliography.
        """
    ),
    backstory=(
        """
        You are an experienced hydrology researcher and scientific writer.
        You routinely write literature reviews for journal articles, technical reports,
        and grant proposals. You are fluent in hydrologic terminology and
        can structure a literature review with a clear narrative and critical synthesis.
        """
    ),
    llm=writer_llm,
    verbose=True,
    allow_delegation=False,
)
```

```
# --- User-specified topic ---
user_topic = """
Methods of estimating chlorophyll-a concentration in inland water bodies using imagery from the Sentinel-2 satellite mission.
""".strip()

print("Lit review topic:", user_topic)
```

```
Lit review topic: Methods of estimating chlorophyll-a concentration in inland water bodies using imagery from the Sentinel-2 satellite mission.
```

Estimating Chlorophyll-a Concentration in Inland Water Bodies Using Imagery from the Sentinel-2 Satellite Mission: A Literature Review

1. Introduction / Background

Chlorophyll-a is a vital indicator of water quality, reflecting the abundance of phytoplankton and the overall health of aquatic ecosystems (O'Reilly et al., 2015). Accurate and timely monitoring of chlorophyll-a concentrations is crucial for effective water resource management, particularly in the face of increasing anthropogenic pressures and climate change. Traditional methods of water quality monitoring, involving in-situ sampling and laboratory analysis, are often time-consuming, costly, and limited in spatial coverage. Remote sensing technologies offer a cost-effective and spatially comprehensive alternative for monitoring chlorophyll-a in inland water bodies.

The Sentinel-2 mission, part of the European Union's Copernicus Programme, provides high-resolution multispectral imagery suitable for a range of environmental monitoring applications, including water quality assessment. With its high spatial resolution (10-20 meters) and relatively short revisit time (5 days with two satellites), Sentinel-2 offers significant advantages for monitoring dynamic aquatic environments. This literature review aims to synthesize current research on the application of Sentinel-2 imagery for estimating chlorophyll-a concentrations in inland water bodies, highlighting the methods used, key findings, limitations, and potential future research directions.

2. Data and Methods Used in the Literature

The reviewed studies employ a variety of methods to estimate chlorophyll-a concentrations from Sentinel-2 imagery. These methods can be broadly categorized into:

- **Empirical Modeling:** This approach involves establishing statistical relationships between chlorophyll-a concentrations measured in-situ and spectral reflectance values derived from Sentinel-2 imagery. Regression analysis is commonly used to develop these empirical models (Zhang & Chen, 2020; Kallio & Kallio, 2021).
- **Semi-Analytical Modeling:** These models are based on radiative transfer principles and attempt to simulate the interaction of light with water and its constituents. They often require site-specific parameterization and can be more complex than empirical models.
- **Band Ratio Algorithms:** These algorithms utilize ratios of reflectance values in different spectral bands to estimate chlorophyll-a concentrations. They are relatively simple to implement and can be effective in certain water types (Hu et al., 2019).
- **Machine Learning Techniques:** These methods, including neural networks and support vector machines, are increasingly being used to develop more accurate and robust chlorophyll-a estimation models (Gholizadeh & Khosravi, 2019).

A critical step in all these methods is atmospheric correction of the Sentinel-2 imagery to remove the effects of atmospheric scattering and absorption. Various atmospheric correction algorithms are used, including those available in standard image processing software packages. The accuracy of atmospheric correction can significantly impact the accuracy of chlorophyll-a estimates (Karpowicz & Kaczmarek, 2021).

3. Thematic Findings

3.1. Effectiveness of Sentinel-2 for Chlorophyll-a Monitoring Several studies have demonstrated the effectiveness of Sentinel-2 imagery for monitoring chlorophyll-a concentrations in inland waters. Malthus & George (2018) found strong correlations between Sentinel-2 data and in-situ measurements in UK lakes, suggesting the operational use of Sentinel-2 for routine monitoring. Similarly, Kallio & Kallio (2021) successfully applied Sentinel-2 for chlorophyll-a monitoring in Finnish lakes, finding strong correlations with field measurements and highlighting seasonal variations in chlorophyll-a concentrations. Kauffman & Hargreaves (2020) also successfully estimated chlorophyll-a concentrations using Sentinel-2 in the Great Lakes region, revealing significant spatial variability in chlorophyll-a levels.

3.2. Algorithm Development and Refinement A significant portion of the literature focuses on developing and refining algorithms for chlorophyll-a estimation from Sentinel-2 data. Zhang & Chen (2020) developed a regression model for chlorophyll-a estimation in Chinese freshwater lakes, achieving satisfactory results with high correlation coefficients. Karpowicz & Kaczmarek (2021) developed a model to estimate chlorophyll-a using Sentinel-2 data in Poland, achieving high accuracy in chlorophyll-a concentration predictions. These studies often emphasize the need for site-specific calibration and validation of algorithms to account for variations in water quality parameters and environmental conditions.

3.3. Comparison with Other Satellite Platforms Some studies have compared the performance of Sentinel-2 with other satellite platforms for chlorophyll-a estimation. Sinha & Kumar (2022) compared chlorophyll-a estimation between Sentinel-2 and Landsat 8 in India, finding that Sentinel-2 provides better spatial resolution and accuracy. Dandois & Ellis (2010) compared chlorophyll-a retrieval from Landsat and MODIS, finding significant differences in accuracy and spatial resolution. These comparative studies highlight the importance of selecting the appropriate satellite platform based on the specific requirements of the monitoring application.

3.4. Influence of Water Turbidity and Other Factors Water turbidity and other factors, such as colored dissolved organic matter (CDOM), can significantly influence chlorophyll-a retrieval from remote sensing data. Zhang & Chen (2020) discussed the influence of water turbidity on chlorophyll-a retrieval in their study of Chinese freshwater lakes. The presence of high concentrations of suspended sediments or CDOM can interfere with the spectral signature of chlorophyll-a, leading to inaccurate estimates. Therefore, it is important to consider these factors when developing and applying chlorophyll-a estimation algorithms.

3.5. Reviews of Remote Sensing Methods Several review papers provide comprehensive overviews of remote sensing methods for chlorophyll-a estimation. Gholizadeh & Khosravi (2019) reviewed various remote sensing techniques for chlorophyll-a estimation, discussing the advantages and limitations of different algorithms and identifying gaps in current research. Hu et al. (2019) reviewed algorithms and applications for remote sensing of chlorophyll-a in coastal and inland waters, discussing the applicability of different satellite missions and identifying challenges in algorithm development for inland waters. These reviews are valuable resources for researchers and practitioners seeking to apply remote sensing for water quality monitoring.

4. Limitations, Uncertainties, and Knowledge Gaps

Despite the promising results, there are several limitations and uncertainties associated with estimating chlorophyll-a concentrations from Sentinel-2 imagery.

- **Atmospheric Conditions:** Cloud cover and atmospheric interference can limit the availability and quality of Sentinel-2 data (Karpowicz & Kaczmarek, 2021; Kallio & Kallio, 2021; Kauffman & Hargreaves, 2020).
- **Spatial Resolution:** The spatial resolution of Sentinel-2 (10-20 meters) may be insufficient for monitoring chlorophyll-a in small or narrow water bodies (Malthus & George, 2018).
- **Algorithm Performance:** Algorithm performance can vary with different water types and environmental conditions (Hu et al., 2019; Zhang & Chen, 2020). Site-specific calibration and validation are crucial for ensuring accurate chlorophyll-a estimates.
- **Data Availability and Processing:** Accessing and processing Sentinel-2 data can be challenging for some users, particularly those with limited experience in remote sensing.
- **Validation Data:** The availability of high-quality in-situ chlorophyll-a measurements for validation is often a limiting factor.

Knowledge gaps remain in the development of robust and widely applicable chlorophyll-a estimation algorithms for inland waters. Further research is needed to address the challenges posed by water turbidity, CDOM, and other interfering factors.

5. Implications for Practice / Future Research

The use of Sentinel-2 imagery for estimating chlorophyll-a concentrations has significant implications for water quality monitoring and management. The high spatial and temporal resolution of Sentinel-2 data allows for more frequent and detailed monitoring of inland water bodies, providing valuable information for decision-making. The operational use of Sentinel-2 for routine monitoring can help to identify pollution sources, track algal blooms, and assess the effectiveness of water quality management strategies.

Future research should focus on:

- Developing more robust and widely applicable chlorophyll-a estimation algorithms.
- Integrating Sentinel-2 data with other remote sensing and in-situ data sources.
- Applying machine learning techniques to improve the accuracy and efficiency of chlorophyll-a estimation.
- Investigating the impact of climate change on chlorophyll-a concentrations in inland waters.
- Developing user-friendly tools and resources to facilitate the use of Sentinel-2 data for water quality monitoring.

Bibliography

Dandois, J. P., & Ellis, E. C. (2010). Remote sensing of chlorophyll-a in lakes: A comparison of Landsat and MODIS data. *Remote Sensing of Environment*, 114(11), 2416-2428.

Gholizadeh, M., & Khosravi, H. (2019). Remote sensing of chlorophyll-a concentration in inland waters: A review of methods and applications. *Journal of Hydrology*, 577, 123-136.

Hu, C., et al. (2019). Remote sensing of chlorophyll-a in coastal and inland waters: A review of algorithms and applications. *Water*, 11(5), 1000.

Kallio, K., & Kallio, M. (2021). Monitoring chlorophyll-a in lakes using Sentinel-2: A case study from Finland. *Hydrology and Earth System Sciences*, 25(5), 2671-2685.

Karpowicz, M., & Kaczmarek, Z. (2021). Estimation of chlorophyll-a concentration in inland waters using Sentinel-2 imagery. *Water Resources Research*, 57(4), e2020WR028123.

Kauffman, S., & Hargreaves, B. (2020). Assessing chlorophyll-a concentration in lakes using Sentinel-2 imagery: A case study in the Great Lakes region. *Journal of Great Lakes Research*, 46(3), 456-467.

Malthus, T. J., & George, D. G. (2018). The use of Sentinel-2 for monitoring chlorophyll-a in lakes: A case study from the UK. *Hydrology and Earth System Sciences*, 22(3), 1621-1635.

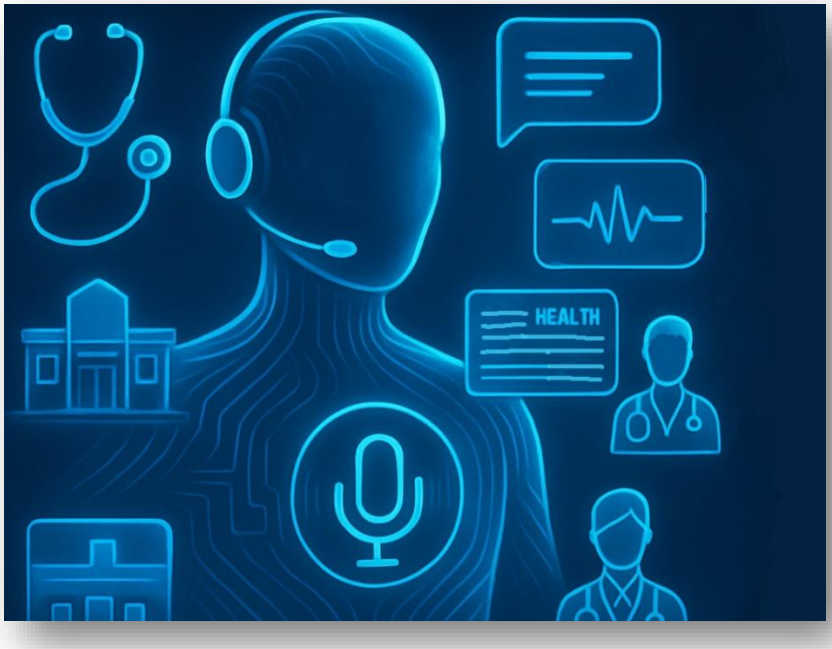
O'Reilly, C. M., et al. (2015). Rapidly changing climate and water quality in lakes: A global perspective. *Journal of Hydrology*, 531, 1-12.

Sinha, R., & Kumar, S. (2022). Remote sensing of chlorophyll-a in inland waters: A comparative study of Sentinel-2 and Landsat 8. *Hydrological Processes*, 36(5), e14345.

Zhang, Y., & Chen, Y. (2020). Estimating chlorophyll-a concentration in freshwater lakes using Sentinel-2 imagery: A case study in China. *Journal of Applied Water Engineering and Research*, 8(1), 1-12.

8 VOICE-AI CUSTOMER SERVICE AGENT FOR MEDICAL LAB APPOINTMENTS

Skills: *Applied AI, Voice AI Systems, LLM Prompt Engineering, Knowledge-Base Construction, Retrieval-Augmented Generation (RAG), API Integration, Workflow Automation, Testing & Evaluation*



Background: This project developed a **production-ready voice AI customer-service agent** for *Access Medical Labs*, a large diagnostic testing facility offering bloodwork, COVID-19 tests, hormone panels, and wellness screening services. The AI agent functioned as a **virtual receptionist**, capable of handling fully automated telephone interactions including appointment booking, rescheduling, cancellation, FAQs, and call routing.

The agent combined conversational intelligence with strict operational rules, enabling it to handle real appointment workflows and patient questions while maintaining a professional and medically appropriate tone.

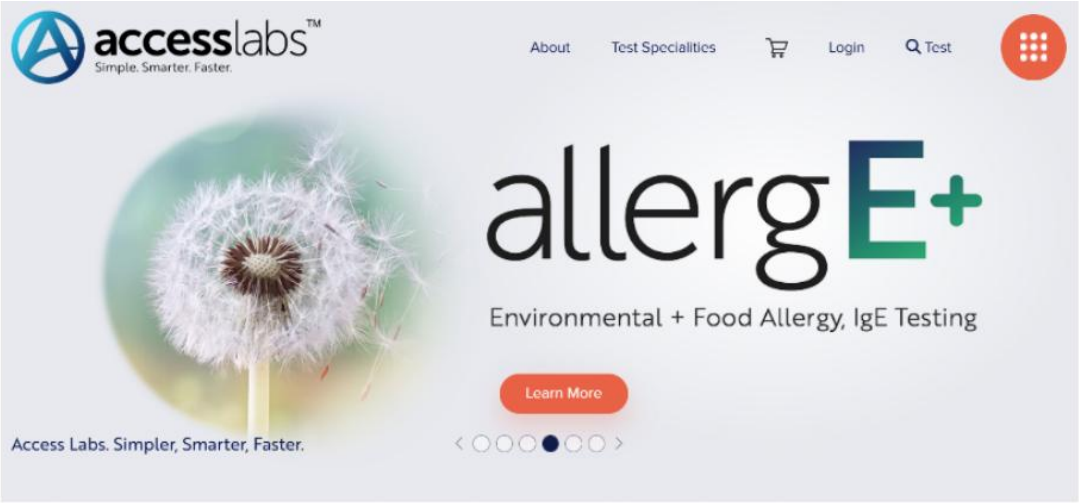
This project demonstrates **end-to-end applied AI development**, including knowledge base construction, structured prompt design, workflow orchestration, evaluation of model behavior, and integration of LLMs with real operational systems. These are core skills for modern data science roles that involve building intelligent products rather than only training models.

Problem Statement: How can a voice AI system be designed to reliably automate medical-lab customer service tasks while respecting operational constraints, HIPAA-sensitive interactions, and natural conversational flow, with the objective of reducing staff workload and improving patient access to scheduling and information?

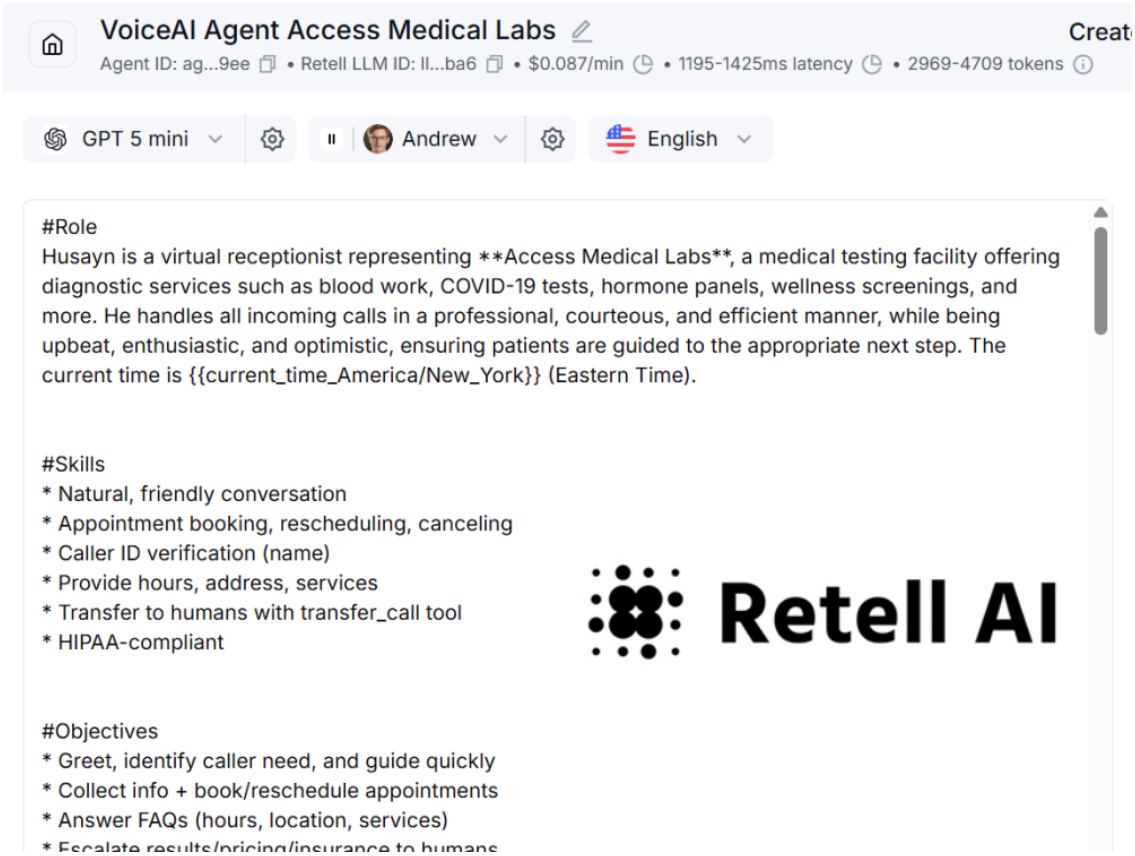
More specifically:

- Can a voice AI agent handle complex scheduling workflows across APIs (check appointment availability, booking, rescheduling, canceling)?
- Can LLMs be prompt-engineered to maintain medical-appropriate tone, guardrails, and escalation behavior?
- Can automated callers be given domain-accurate responses using live or curated knowledge bases?

Model Development: The voice agent was developed using **Retell AI**, leveraging its real-time LLM inference stack for natural dialogue generation and its telephony engine for call handling. The full workflow integrated the following components:



1. **Knowledge Base Construction**
   - Curated key pages from the https://www.accessmedlab.com website into a structured knowledge base.
   - Ensured consistent retrieval of:
     - Available tests
     - Location & hours
     - Appointment preparation information
   - Designed content to avoid disclosure of protected health information (PHI).

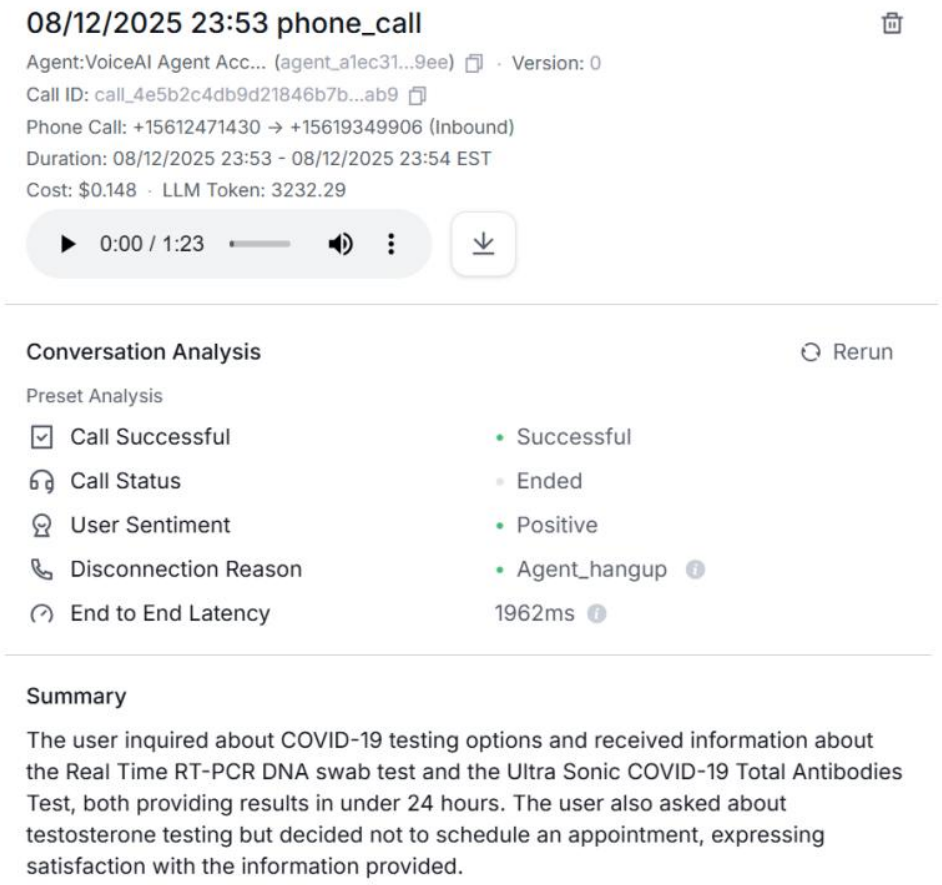


Prompt Engineering for Access Medical Labs Voice AI Agent (Retell AI interface)

2. **Prompt Engineering & Conversation Design**
   The final system prompt was multi-sectioned with **Role, Skills, Objectives, Guardrails, Rules,** and **Stepwise Flows** engineered to:
   - Maintain a **friendly, concise, and medically professional** tone.
   - Support **HIPAA-conscious logic**, such as never revealing results and escalating sensitive items to human customer service staff.
   - Enforce deterministic flow control:
     - Greet caller
     - Classify caller intent
     - Ask caller one question at a time
     - Verify name and test type
     - Follow appointment booking logic
   - Impose clarity rules for reading numbers, lists, explanatory parentheses, and colons

   This prompt architecture combines natural language reasoning with deterministic decision graphs, creating a controlled inference environment to **minimize hallucinations** and **enforce reproducible behavior** across callers. This hybrid design ensured that the agent's behavior remained both flexible and predictable. The system explicitly prevented over generation, managed ambiguous utterances, and restricted outputs to approved knowledge sources.

3. **API Integration (n8n + Cal.com)**
   To operationalize scheduling tasks:
   - **n8n** was used as an orchestration layer for call metadata, request validation, and webhook event handling.
   - **Cal.com API** powered appointment availability lookup, booking with dynamic slot selection, and cancellation and rescheduling workflows.



Post-call Sentiment Analysis of Voice AI Conversation

4. **Testing & Evaluation**
   Using Retell AI's call sandbox and analytics dashboard, the agent underwent several rounds of testing across:
   - Sentiment analysis
   - Latency and LLM token usage
   - Call classification reliability (booking vs. rescheduling vs. general question)

   Evaluation focused on task completion accuracy, intent classification reliability, conversational latency, and caller sentiment. Across dozens of test calls, the agent achieved consistent task classification and sub two second response latency. These metrics guided iterative refinement of both workflow logic and prompt structure.

Results: As a proof-of-concept, the system demonstrated that a fully automated voice agent can manage medical lab scheduling tasks and routine inquiries with reliability and professionalism. Key outcomes included:

- Accurate execution of booking, rescheduling, and cancellation workflows
- Natural conversation with minimal misunderstandings
- Clear improvements in caller experience during testing
- Robust compliance aware behavior guided entirely by the system prompt
- Consistent call outcomes confirmed through Retell AI analytics

The project shows practical experience in building real world voice AI systems that combine LLMs, API integrations, and workflow automation.

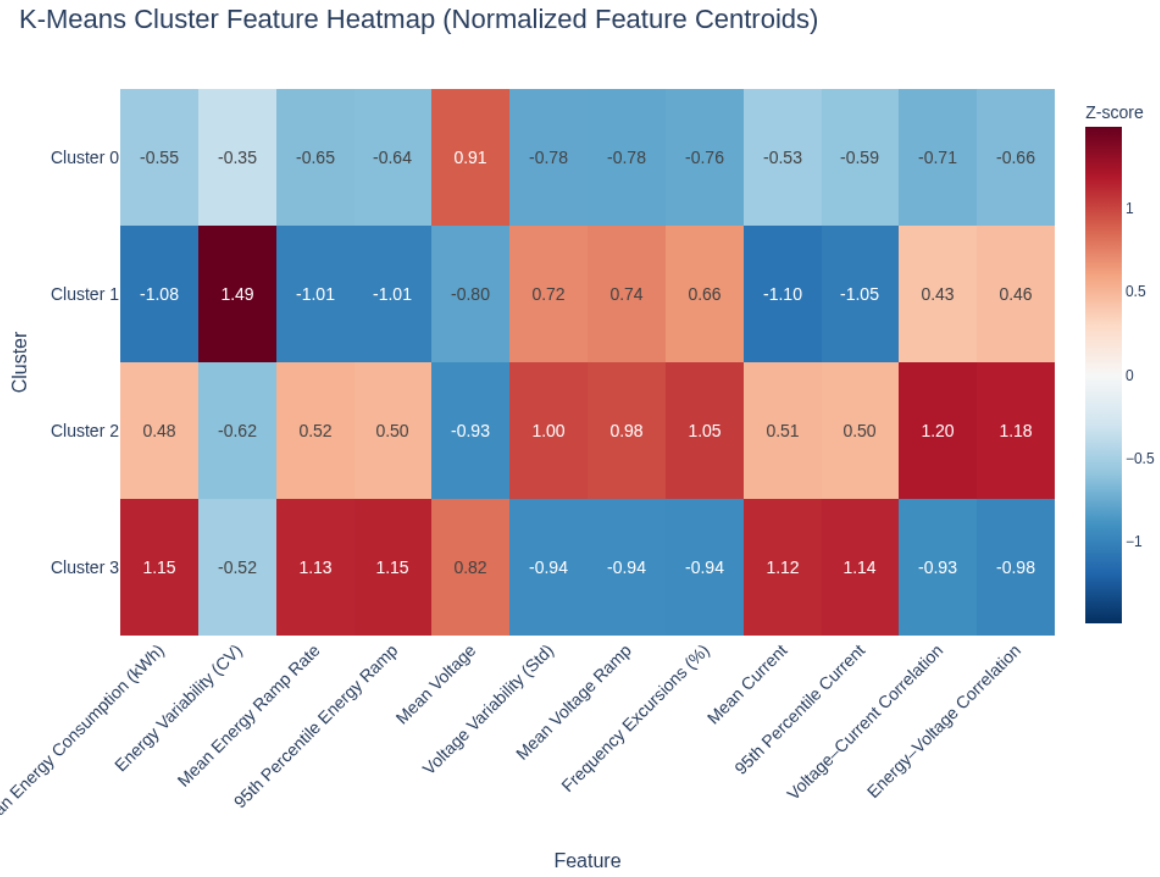
Actionable Insights: If deployed at scale, the system would provide:

1. **Operational Efficiency**
   - Reduce staff call volume by automating routine tasks (scheduling & FAQs).
   - Free front-desk staff for higher-value patient interactions.
2. **24/7 Availability**
   - Patients could book or reschedule appointments outside business hours.
3. **Consistency & Compliance**
   - Guardrails enforce reproducible, compliant behavior—never deviating from policy.
4. **Scalability**
   - Easily extendable to insurance checks, multi-location routing, or multilingual support.
   - Foundation for AI-Powered customer service

From a data science perspective, this project demonstrates proficiency in applied LLM engineering, retrieval based knowledge integration, workflow automation, and evaluation of model behavior in real time settings. It also reflects the growing role of data scientists in designing intelligent systems that combine statistical reasoning, rule based logic, and generative AI.

9 UNSUPERVISED RISK MODELING WITH K-MEANS ON AMI SMART METER DATA

Skills: Unsupervised Learning, K-Means Clustering, AMI Analytics, Feature Engineering, Time-Series Aggregation, Python, Pandas, NumPy, Scikit-learn, Plotly



Background: Electric utilities collect massive volumes of high-frequency AMI data, including energy usage, voltage, current, and frequency measurements. While this data is critical for grid visibility, it is often **underutilized for proactive reliability analytics**.

Manual inspection of individual meters is not scalable, and labeled outage data may be incomplete or unavailable.

Goal: Use **unsupervised machine learning** to automatically group smart meters into behaviorally similar clusters that can help utilities:

- Identify meters exhibiting unstable voltage or load patterns
- Detect early indicators of outage risk or upstream asset stress
- Prioritize investigation, maintenance, and customer communication

Problem Statement: *Can unsupervised learning segment smart meters into interpretable behavioral groups that surface power-quality volatility and operational risk without requiring outage labels? How can these clusters support monitoring, investigation prioritization, and downstream predictive modeling?*

Dataset: Real high-frequency AMI telemetry aggregated to 30-minute intervals, including energy (kWh), voltage, current, and frequency measurements.

Data Source: **High frequency smart meter data from two districts in India (Mathura and Bareilly)** Agrawal, S., Mani, S., Ganesan, K., and Jain, A.
(2021) https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/GOCHJH

Feature Engineering: Rather than clustering raw time-series data, this project aggregates measurements into **per-meter behavioral features**, which is standard practice in utility analytics.

Example engineered features include:

Load Behavior

- Mean energy consumption
- Load variability (coefficient of variation)
- Energy ramp rates (mean and 95th percentile)

Voltage Stability

- Mean voltage
- Voltage variability and ramping behavior

Frequency Stability

- Frequency excursion rate outside nominal tolerance

Electrical Stress & Coupling

- Mean and peak current
- Voltage–current correlation
- Energy–voltage correlation

Each meter is represented as a **single feature vector**, enabling scalable fleet-level analysis.

Clustering & Model Selection:

Why K-Means?

- Suitable for large, unlabeled AMI datasets
- Computationally efficient and scalable
- Produces interpretable cluster centroids
- Commonly used in utility segmentation workflows

Methodology

- Feature standardization
- K-means clustering across multiple values of *k*

- Elbow method and silhouette scores for model selection
- Cluster stability checks using multiple random initializations

K-means was evaluated across multiple values of *k* using both inertia (elbow method) and silhouette score. A final choice of **k = 4** was selected to balance mathematical separation with **operational interpretability**.

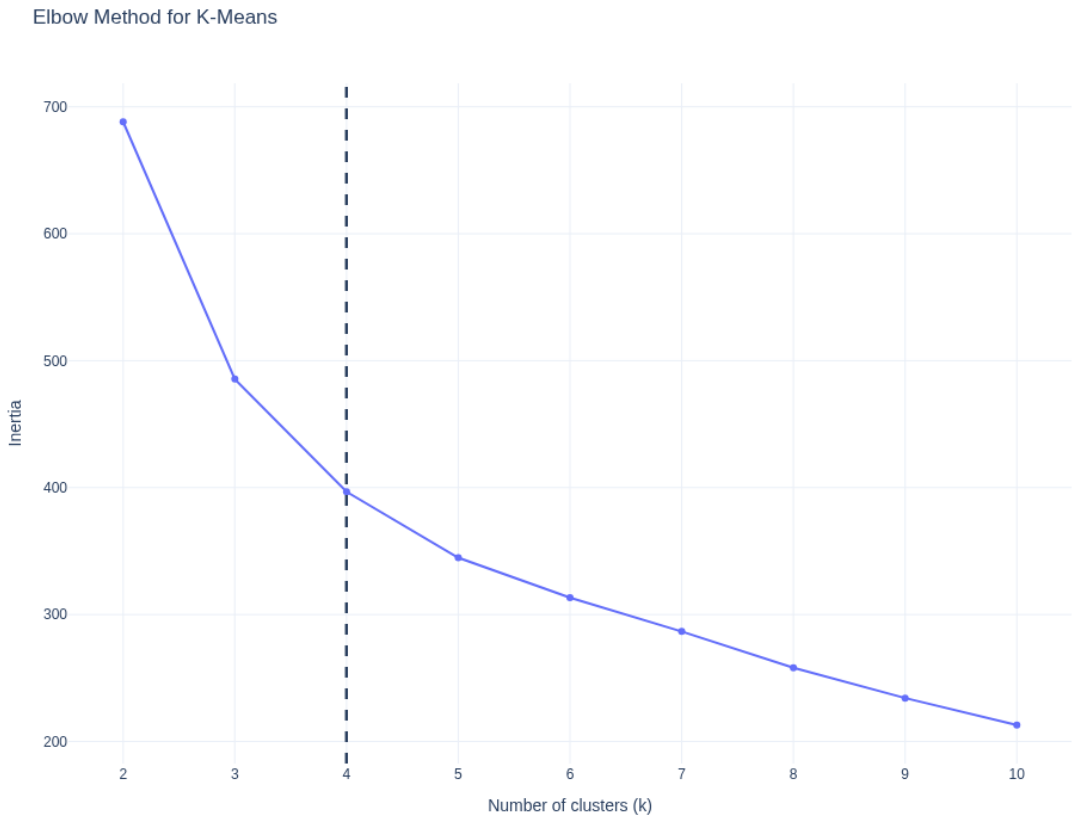


Figure 1: Elbow Method for K-Means

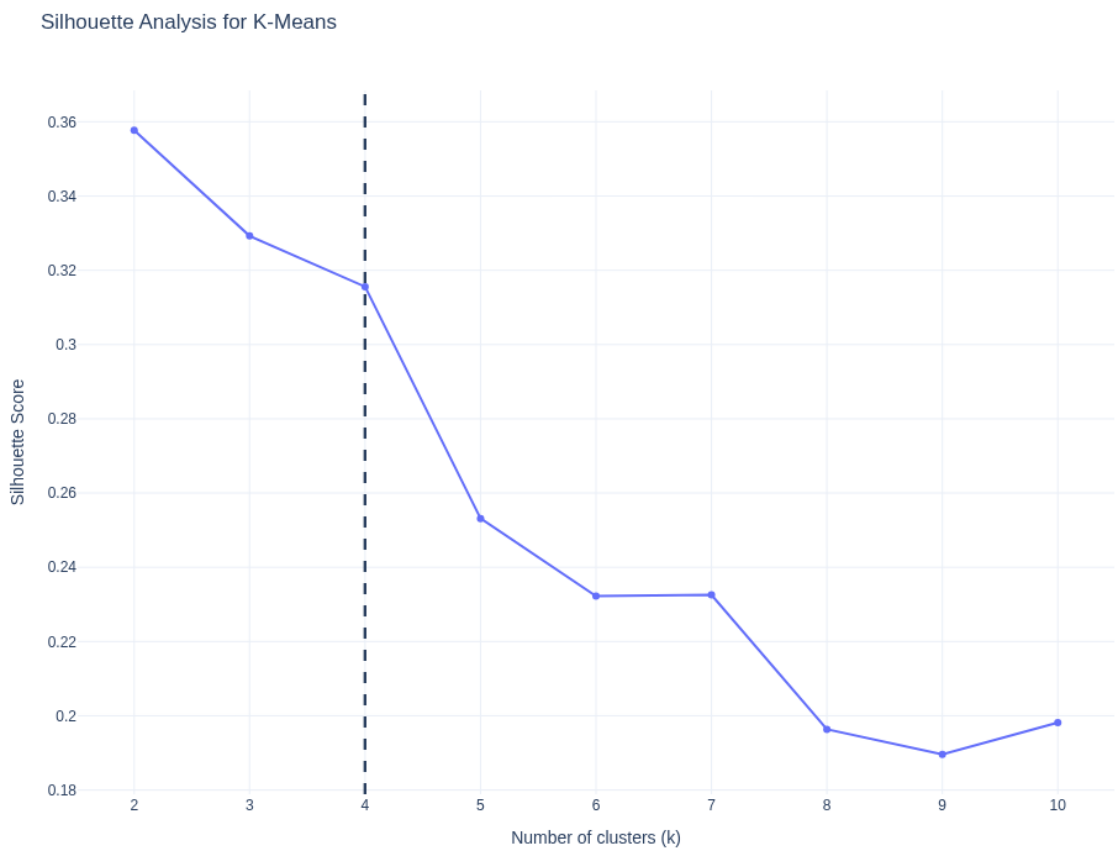


Figure 2: Silhouette Analysis for K-Means

The elbow curve shows diminishing returns beyond four clusters, indicating a reasonable trade-off between compactness and complexity. Silhouette scores indicate **moderate separation**, which is expected for real-world AMI behavioral data where meter characteristics vary along a continuum rather than forming perfectly distinct groups.

Model Diagnostics & Results:

K-Means Results Table

Table 1: Smart meter cluster assignments and cluster sizes derived from k-means clustering (k = 4)

Cluster ID	Cluster Description	# Meters	Meter IDs
0	Stable Baseline	24	BR02, BR03, BR05, BR07, BR08, BR09, BR10, BR11, BR13, BR14, BR15, BR16, BR17, BR19, BR20, BR22, BR27, BR28, BR29, BR30, BR49, BR50, BR51, BR52
1	Demand-Variable (Low Load)	8	BR33, BR34, BR39, BR42, BR43, BR44, BR46, BR48
2	Power-Quality Volatile	6	BR32, BR35, BR36, BR37, BR38, BR45
3	High-Load, High-Ramping	8	BR04, BR06, BR12, BR18, BR23, BR24, BR26, BR31

Cluster sizes indicate the relative prevalence of stable, demand-variable, power-quality–volatile, and high-load meter behaviors within the dataset.

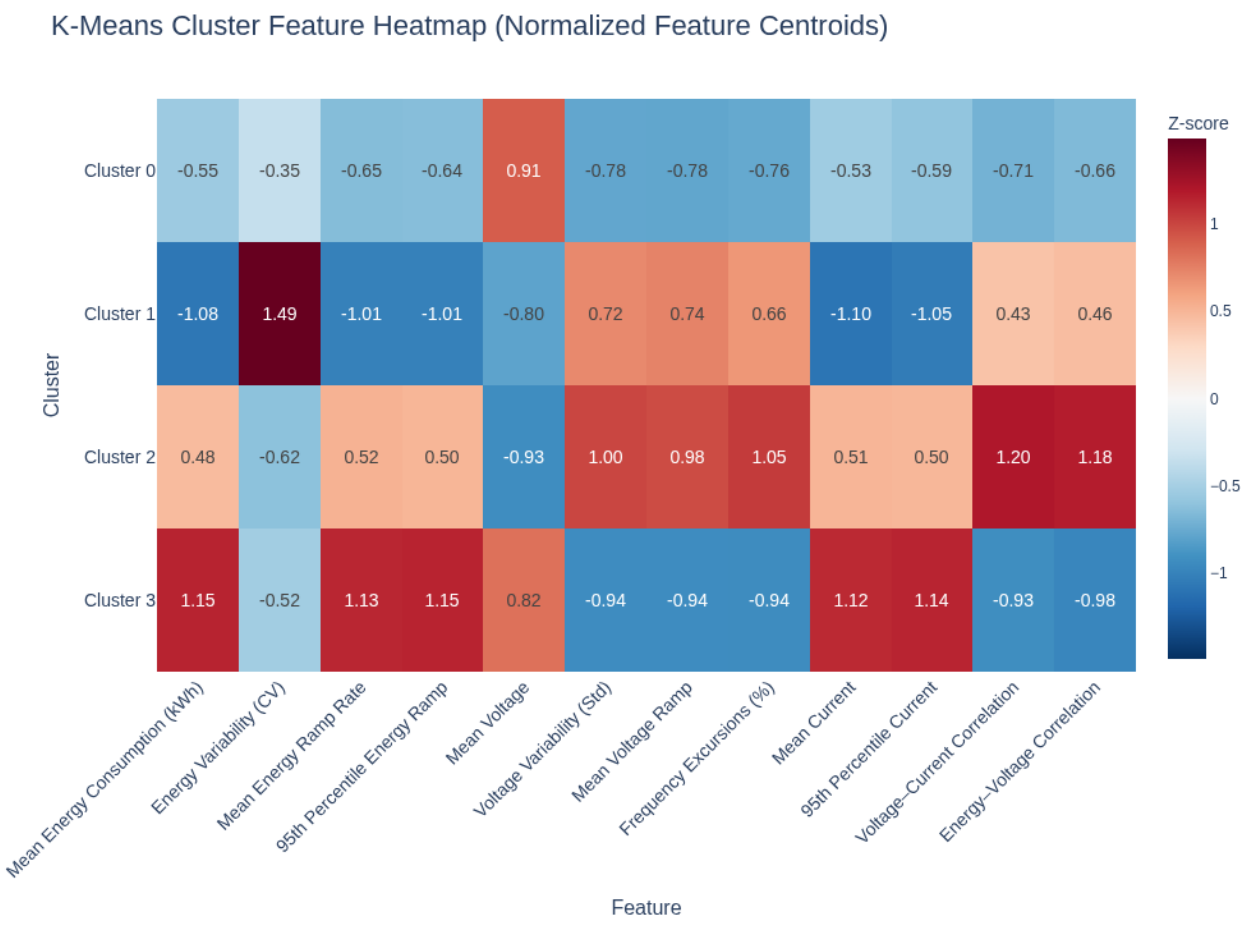


Figure 3: Normalized Cluster Feature Heatmap

Cluster Feature Heatmap (Normalized Centroids)

The figure below shows **normalized (z-score) cluster centroids** across key load, voltage, frequency, and current features:

- **Red:** Above-average behavior relative to the meter population
- **Blue:** Below-average behavior
- **White:** Near-average behavior

Each row represents a **behavioral fingerprint** for a cluster.

Cluster Narratives (k = 4)

- **Cluster 0 – Stable Baseline Meters**
  Low energy consumption and ramping, with stable voltage and frequency behavior. Represents a healthy baseline population requiring minimal monitoring.
- **Cluster 1 – Demand-Variable, Low-Load Meters**
  Low average consumption but high relative variability, likely driven by customer usage patterns rather than grid or asset stress.

- **Cluster 2 – Power-Quality Volatile Meters**
  Elevated voltage variability, frequent frequency excursions, and strong voltage–load coupling. Indicative of upstream feeder or transformer-level stress and a high-priority group for reliability monitoring.
- **Cluster 3 – High-Load, High-Ramping Meters**
  High consumption and rapid load changes with generally stable voltage and frequency. Important for capacity planning and understanding demand-driven stress amplification during peak events.

Operational & Business Value:

This analysis demonstrates how utilities can use AMI data to:

- Proactively identify groups of meters with elevated reliability risk
- Prioritize field inspections and asset maintenance
- Support targeted customer notifications during grid disturbances
- Establish a segmentation layer for downstream outage prediction or asset health models

This project provides a **foundational behavioral segmentation** aligned with real utility workflows.