Sasi Jyothirmai Bonu

Boulder, CO • sasibonu@gmail.com • (720) 341-3315 • LinkedIn • GitHub
My Website

(Open to Relocate)

5 years of experience in ML /LLM / Data Science. Actively seeking new opportunities.

PUBLICATIONS

• Localizing epileptogenic network from SEEG using non-linear correlation, mutual information, and graph theory analysis, Journal of Engineering in Medicine (2022) Link
• Analysis of Classification Algorithms for Brain Tumor Detection, 9th International Symposium on Embedded Computing and System Design (ISED, IEEE) 2019
• Contributed to data analysis for "Yuan, L., Zhu, L., Johns, E., Mix, K., & Smith, L. Road to transfer and generalization: The role of knowledge hubs. (in preparation, DEL Lab, CU Boulder). Link
• Supported modeling for "Zhu, L. & Yuan, L. Cross-linguistic differences without cross-cultural confounds: Modeling the effect of linguistic systematicity on learning. (in preparation, DEL Lab, CU Boulder). Link
• Developed a Vision Transformer model and data analysis pipeline for"Hopkins, T. & Yuan, L. From visual features to hand-eye coordination: An eye-tracking study with musical notations. (in preparation, DEL Lab, CU Boulder). Link

PROFESSIONAL EXPERIENCE

NETWORK THEORY APPLIED RESEARCH INSTITUTE **Remote, US**
AI/ML Engineer **Aug 2025 - Present**
Enabled data-driven agricultural insights by fine-tuning domain-specific LLMs on FAO, USDA, and research data and operationalizing end-to-end pipelines.

- Worked with stakeholders to define real agricultural problems and translate them into clear use cases.
- Collected and cleaned data from FAO, USDA, and research sources to create reliable training datasets.
- Designed the data pipeline for ingestion, validation, and versioning so models are always trained on trusted data.
- Trained and fine-tuned domain LLMs, optimizing accuracy, cost, and latency for production-ready deployment.
- Mentored a graduate student in data preprocessing, pipeline design, and model evaluation, speeding up delivery and knowledge transfer.

SAMSTREAM.AI **Remote, US**
AI /ML Engineer **Jun 2024 -May 2025**
Enabled data-driven federal contract pricing by predicting bid success and automating ingestion pipelines for faster decision-making.

- Built an XGBoost model to predict federal contract bids, improving forecast accuracy by 20% and guiding pricing decisions.
- Engineered and validated interpretable features to help pricing teams understand and justify model predictions.
- Developed PySpark ETL pipelines for automated ingestion, reducing data latency by 60%.
- Tuned hyperparameters and compared XGBoost with baseline models to justify the final model choice.

DEL LAB **Boulder, US**
AI /ML / Data Scientist **May 2024 - May 2025**
Project 1: *Hours were spent by researchers on reviewing speech data and segmenting trials. The goal was to make a Streamlit dashboard powered by IBM Watson to segment trials and reduce the review time.*

- Developed a Streamlit dashboard with IBM Watson Speech-to-Text to transcribe and segment 100+ learning trials, enabling faster identification of spoken number responses and reducing manual review time by an estimated 60%.

- Automated the matching of behavioral trial data with speech transcripts by integrating timestamped trial metadata, eliminating hours of manual annotation for research assistants.
- Enabled seamless error correction and data export via an interactive dashboard, producing word-by-word segmented transcripts with millisecond-precision timestamps and trial annotations.

Project 2: *The goal was to train an attention-based LSTM to convert between numbers, words, and visual blocks, helping compare how children learn numbers with how well a model can mimic that learning.*

- Provided data-driven insights to advance the study of human cognition and the development, education, and learning of children, leading to publications.
- Trained an attention-based LSTM model to translate between Arabic numerals, number words, and visual blocks under three conditions: (1) numerals + words, (2) blocks + words, and (3) all three combined.
- Evaluated model performance using the BLEU score and tested trial correctness using a probability-based measure that compared raw attention-weighted probabilities across target and foil images.
- Found an average 75% accuracy on new testing data with Arabic numerals and 65% accuracy on new testing data with blocks.

Project 3: *Make a data analysis pipeline enabling analysis of how colors versus stickers influence how quickly new learners pick up typing on a keyboard.*

- Engineered data workflows for trial segmentation and gaze metric extraction (fixation area, duration, switching), resulting in high-quality derived datasets for downstream statistical modeling.
- Boosted frame-level classification accuracy of egocentric video data by 30% using a Vision Transformer, enabling precise AOI tracking.

WIPRO **Remote, India**
ETL Developer (Project Engineer): Data Analytics & AI **Sep 2020 - Jul 2023**
Client: Telkomsel, Indonesia

Project 1: *The goal was to develop backfill pipelines to process late-arriving/missing subscriber data and figure out the source of latency, ensuring accurate and timely campaign execution.*

- Extracted late-arriving records from upstream systems, processed, and analyzed payloads to categorize errors by type, enabling targeted fixes.
- Conducted multi-day analysis to identify patterns and error sources, informing preventive measures and improving pipeline reliability.
- Built ad hoc backfill pipelines to correct and merge delayed records into campaign datasets, ensuring no scheduled jobs are skipped/failed and reducing manual intervention.

Project 2: *Developed a data pipeline to generate clean, reliable datasets for weekly leadership reporting and business analysis.*

- Built an ETL pipeline to extract data from PostgreSQL tables, aggregate, and transform weekly data, ensuring accurate numbers for senior leadership reporting to stakeholders
- Optimized pipeline performance to efficiently process growing historical data in HDFS, improving processing speed and resource utilization while maintaining data accuracy.
- Produced clean, consistent datasets allowing business analysts to explore trends and perform deeper analysis without missing, duplicate, or inconsistent data.

Project 3: *Built a PySpark ETL job for automated monitoring to ensure reliable data processing every day and alerting otherwise.*

- Developed PySpark transformations to calculate record counts and other metrics, ensuring daily data consistency and correctness across inputs and outputs.
- Implemented validation checks to compare input and output records, triggering alerts automatically if discrepancies or failures occur.

- Reduced manual monitoring efforts and improved data reliability by 45%, enabling faster detection of issues and minimizing potential disruptions.

AMRITA ADVANCED CENTER FOR EPILEPSY, AIMS **Kochi, India**
Research Assistant **Jan 2020 - Aug 2020**
Improved epilepsy surgery planning by identifying seizure-onset zones from SEEG data and visualizing neural patterns to inform clinical decisions.

- Developed statistical and ML models to localize seizure onset zones from SEEG recordings.
- Built a clinical visualization tool mapping seizure origin nodes, supporting surgical decisions, and authored a peer-reviewed journal publication in the UK.

PROJECTS

Epi Graph, AIMS

- Designed and implemented a full SEEG data analysis pipeline, processing recordings from 10 epilepsy patients, applying bandpass and notch filtering to isolate ~50,000 interictal spikes across datasets.
- Analyzed SEEG recordings using nonlinear h^2 correlation to assess signal synchrony, engineered graph-theoretical features from correlation matrices, and localized seizure onset zone accurately.
  .

Lost & Found, CU Boulder

- Developed a web app for a lost-and-found platform, enabling users to report and search items via text and image inputs with an intuitive UI/UX.
- Used OpenAI's CLIP model and implemented secure image storage and user data management using Google Cloud Storage and MySQL backend APIs, ensuring scalable and reliable performance.

SKILLS

- **Databases:** PostgreSQL, MySQL
- **Programming Languages:** Python (NumPy, Pandas, Matplotlib, Scikit-learn, PyTorch), R (Shiny, dplyr, ggplot2), PySpark, SQL
- **Tools & Libraries:** Excel, PowerPoint
- **Data Visualization:** Tableau, Streamlit
- **Machine Learning:** Supervised ML (Linear and Logistic Regression, Decision Trees, RF, Xgboost), Unsupervised ML (K-means, PCA), Classification techniques (binary classification, multi-class classification)

EDUCATION

UNIVERSITY OF COLORADO, BOULDER **Colorado, US**
MS in Data Science **May 2025**

AMRITA VISHWA VIDYAPEETHAM **India**
B.Tech in Electronics & Communication Engineering **Jun 2020**