Mugunthan Kesavan

Houston, TX | (346) 204-1541 | mkesavan@cougarnet.uh.edu | LinkedIn | GitHub | Portfolio

SUMMARY

Engineering Data Science graduate student with a focus on **multilingual NLP, speech** and **AI for healthcare**. Experienced in building **end-to-end machine learning pipelines** in Python and Spark, from data collection and preprocessing to model training, evaluation, and analysis. Research spans **low-resource South Indian languages**, **speech-to-speech translation**, and **protein modeling** using transformer embeddings. Passionate about deploying AI systems that **reduce language and safety barriers** in real-world clinical and scientific settings.

EDUCATION

University of Houston, Houston, TX — *Expected Graduation: May 2027*
Master of Science, Engineering Data Science

Amrita Vishwa Vidyapeetham, Coimbatore, India — *Graduated: June 2025*
Bachelor of Technology, Artificial Intelligence Engineering
GPA: 3.7/4.0
Relevant Coursework: Machine Learning, Big Data Analytics (Hadoop, Spark, Scala), Distributed Systems, Probability & Statistics, Data Structures & Algorithms, Deep Learning

EXPERIENCE

Research Intern — FLAME University — Jan 2025 – Jul 2025

- Developed a **machine learning framework for protein property prediction** on 30K+ sequences using **transformer-based embeddings** and SVM, achieving **98.3% accuracy** and **98% F1-score**.
- Built **reproducible Python pipelines** (Pandas/NumPy) for ingestion, cleaning, and feature engineering, reducing manual dataset preparation time by 50%.
- Explored **distributed ML workflows** with **Hadoop and Spark** for scalable feature extraction and model training.
- Applied **unsupervised clustering** and feature selection to identify biologically meaningful patterns and improve downstream model quality.
- Conducted **10-fold cross-validation with Grid Search** to systematically tune models and ensure experiment reproducibility.

PROJECTS

Tamil-to-Telugu Speech-to-Speech Translation for Multilingual Communication GitHub

- Built an **end-to-end speech-to-speech translation prototype** from Tamil to Telugu by chaining **ASR, NMT, and TTS** components in Python.
- Implemented data preprocessing, text normalization, alignment, and audio generation pipelines; evaluated translation and audio quality using automatic metrics and listening tests.
- Designed the system with **real-world bilingual communication** in mind, focusing on robustness for **conversational, code-mixed speech**.

Natural Language Processing for Low-Resource Indian Languages GitHub

- Developed **extractive + abstractive summarization** for Tamil news/text using frequency-based and clustering-based extractive methods combined with transformer-based abstractive models.
- Applied **IndicBERT, LaBSE, and MuRIL embeddings** for **humor detection in Telugu social media text** with cost-sensitive learning; published in Springer LNNS.
- Created and cleaned **Tamil tweet sentiment corpora** and benchmarked multilingual transformers + classical ML models for **zero-shot vs supervised** sentiment classification (accepted ICDSA paper).

Machine Learning for Protein Analysis

- Processed 30K+ protein records using **ProtBERT/ESM2 embeddings** and clustering to discover structural and functional patterns.
- Benchmarked **Random Forest, XGBoost, and SVM**, achieving **10% higher accuracy** than baseline models; documented trade-offs between architectures.

Sales Data Analytics with Hadoop and Spark

- Designed a **distributed data processing pipeline** using **Hadoop MapReduce and Apache Spark** to analyze 15M+ transaction records.
- Implemented large-scale **data ingestion, transformation, and feature extraction** workflows for downstream ML tasks.
- Used **Spark MLlib** for classification and clustering, achieving a 40% reduction in computation time compared to serial processing.

Protein Methyl Transfer Identification (Ongoing)

- Building **hybrid pipelines** combining transformer-based embeddings and physicochemical features for protein classification.
- Evaluating **SVM, Random Forest, and XGBoost** with cross-validation to assess robustness on noisy biological data.

TECHNICAL SKILLS

Programming: Python (primary), Scala, R, C++, Java, SQL
ML & NLP: scikit-learn, PyTorch, TensorFlow, Keras, Hugging Face Transformers, clustering, model evaluation, multilingual NLP, sentiment analysis, sequence modeling
Data & Systems: NumPy, Pandas, Jupyter, Hadoop, Apache Spark, Spark MLlib, MapReduce, Git, Linux
Cloud & DevOps: AWS, GCP, Azure, Databricks, Docker (basics)
Domains: Multilingual NLP, Speech-to-speech translation, Bioinformatics / Protein modeling, Applied Machine Learning, Data Science

---

PUBLICATIONS

- Rupa, N. B., Hima, Y., **Mugunthan, K.**, & Premjith, B. (2025). *Humor Detection in Telugu Social Media Text Using Cost-Sensitive Learning and Indian Language Embeddings.* In: *Proceedings of ICRTC 2024*, Springer LNNS 885. Link.
- **Mugunthan, K.**, et al. (2025). *Machine Learning Framework for the Prediction of Pore-Forming Proteins.* Accepted (in press), Springer.
- **Mugunthan, K.**, Jain, R., & Jayaraman, V. K. (2025). *Pretrained Models for Zero-Shot Classification and Supervised Learning: A Simulation Study for Tamil Tweet Sentiment Analysis.* Accepted (forthcoming), *6th International Conference on Data Science and Applications.*

LEADERSHIP & ACTIVITIES

- Led workshops on **Python, Machine Learning, and Model Optimization** for 70+ students.
- Mentored 10+ juniors on end-to-end **AI and Data Science projects** (data collection, modeling, and evaluation).
- Ranked in the **Top 20%** on Kaggle machine learning competitions.