Tahsin Reasat

Albany, CA, USA
greasat@gmail.com
+16157141563

[Google Scholar] [LinkedIn]
[GitHub] [Kaggle]

Summary

ML Engineer / Applied ML Engineer with 8+ years of experience building end-to-end ML systems across document AI, CV, NLP, and speech. Strong track record delivering production impact (80–90% reduction in manual processing time at 400 faxes/day; $3M/year claims recovery potential; 50% workload reduction). Experienced with workflow orchestration (**Temporal**), AWS (**Bedrock, Textract, S3, EC2, Aurora PostgreSQL**), backend APIs (**FastAPI/Flask**), evaluation/monitoring, and human-in-the-loop tooling.

Technical Skills

- **Languages:** Python, SQL, Bash, MATLAB, R.
- **ML/DS:** PyTorch, TensorFlow, Keras, Transformers, PEFT/LoRA, Scikit-learn, XGBoost, Pandas, NumPy, OpenCV.
- **Systems & Cloud:** AWS (Bedrock, Textract, S3, EC2, Aurora PostgreSQL), Temporal, Docker, SLURM, Distributed GPU Training.
- **Backend & Tools:** FastAPI, Flask, Streamlit, Jira.

Experience

Ambry Genetics — Albany, CA
AI Engineer (Remote) — *2025 - Present*

- Built a human-in-the-loop fax document processing system on AWS to automate document splitting, classification, and patient-to-order matching; **reduced manual processing time by 80–90%** and supported **400 faxes/day**.
- Designed **Temporal** workflow state transitions with conditional routing, retries/fallbacks, and human review paths for edge cases.
- Implemented OCR + LLM pipelines using **AWS Textract** and **AWS Bedrock** to extract structured data; persisted document metadata and workflow state in **Amazon Aurora PostgreSQL**.
- Built reviewer-facing UIs for document queuing, AI output review, and patient order matching; created a **Streamlit** dashboard tracking model performance, processing time, and throughput.
- Designed and implemented an **LLM**-based extraction system for insurance denial documents, enabling the potential recovery of approximately **$3M** annually in lost claims.
- Built an **LLM**-based pedigree image classifier on **AWS Bedrock**, achieving **99% accuracy** on an internal test dataset.
- Developed **ontology-aware features** for a graph data structure and trained an XGBoost model to prioritize genes; deployed on EC2 and **reduced manual variant analysis workload by 50%**.

Vanderbilt University — Nashville, USA
Graduate Research Assistant, ECE — *2018 - 2024*

- Built an active learning-based training pipeline for SimCLR on histopathology images; reduced data requirements by 93% and training time by 62%. [paper] [github]
- Designed multimodal UNet and Segment Anything Model (SAM) for MRI soft tissue tumor segmentation; achieved 80% Dice (state of the art). [paper][github]
- Created an MRI annotation platform on LabelStudio for physicians and curated a musculoskeletal soft tissue tumor segmentation (MSTT) dataset with 199 patients.
- Developed multimodal image + clinical data fusion models for MSTT malignancy prediction, improving AUC by 4% over single-modality baselines.

Bengali.AI — Dhaka, Bangladesh
Co-founder — *2017 - Present*

- Led open-source dataset/model releases for Bengali ASR/CV/NLP; winner of multiple **Kaggle research grants** totaling $200k. [Competition examples]
- Engineered normalizer and parser libraries for Abugida Unicode texts supporting 7 Indic languages; improved LLM robustness by 5–10 points across multiple metrics. [paper]
- Built the OOD-Speech benchmark dataset (1100+ hours; 22,000+ contributors; 17 domains) and fine-tuned GPT-based Whisper for state-of-the-art Bengali regional ASR. [paper] [Kaggle Comp] [Demo]
- Developed BaDLAD (33,695 annotated documents; 6 domains) and trained document layout analysis models (Mask R-CNN/YOLO). [paper]
- Organized a Kaggle Featured research competition for handwritten Bengali grapheme classification (411k curated samples). [Kaggle Comp]

Bangladesh University of Engineering and Technology Dhaka, Bangladesh
Graduate Research Assistant, EE *2016 - 2018*

- Engineered a CNN for inferior myocardial infarction detection from ECG signals with 84.54% accuracy. [paper] [github]

Selected Projects

- **Document AI Workflow Platform (Temporal + AWS):** Built end-to-end orchestration for OCR → split → classify → extract, with conditional routing, retries/fallbacks, and human review; tracked per-step latency for bottleneck analysis.
- **Insurance Denial Information Extraction:** Implemented LLM-based structured extraction for denial documents, enabling approximately $3M/year in recovered claims (potential).
- **Genomics Decision Support:** Built ontology-aware feature pipeline and deployed an XGBoost model on EC2 to prioritize genes, reducing manual variant analysis workload by 50%.
- **Pedigree Detection:** Built an LLM-based pedigree image classifier on AWS Bedrock with 99% accuracy on an internal test dataset.
- **Evaluation & Monitoring:** Set up internal test datasets and a Streamlit dashboard to track model quality, throughput, and processing time across pipeline stages.
- **Open Datasets for Low-Resource Languages:** Released large-scale ASR and document AI benchmarks via Kaggle competitions to drive adoption and reproducible baselines.

Grants and Awards

- Winner of multiple **Kaggle research grants** (2019, 2023) for hosting research code competitions, totaling $200k. [Competition examples]
- Winner of the **Kaggle Community Competition Creator** Prize (July 2022; $5k).

Selected Publications

- Data Efficient Contrastive Learning in Histopathology using Active Sampling. **Machine Learning with Applications**, 2024. [link]
- Abugida Normalizer and Parser for Unicode Texts. **LREC-COLING**, 2024. [link]
- OOD-Speech: A Large Bengali Speech Recognition Dataset for Out-of-Distribution Benchmarking. **Interspeech**, 2023.
- BaDLAD: A large multi-domain Bengali document layout analysis dataset. **ICDAR**, 2023. [link]
- Full list: Google Scholar.

Education

- **Vanderbilt University** Nashville, USA
  PhD in Electrical and Computer Engineering *2018 - 2024*