Ashtik Mahapatra

Buffalo, Ny • ashtikkm@gmail.com • +1 (716) 573-8945 • linkedin.com/in/ashtik-mahapatra
github.com/fieryash

Experienced Data Scientist skilled in NLP, computer vision, and generative AI, with award-winning ML pipelines deployed at Wolters Kluwer generating millions in revenue. IEEE-published researcher proficient in transformers, diffusion models, end-to-end system development, cloud automation (AWS/Azure), and building production-grade AI for accessibility, legal analytics, and diverse applications.

EDUCATION

Masters of Science in Computer Science & Engineering (AI/ML) — **Aug 2024 - Jan 2026**
University at Buffalo (SUNY Buffalo) — **GPA:** 3.917

Completed graduate coursework at the University at Buffalo, specializing in Deep Learning, Applied ML at Scale, Algorithms, Operating Systems, and Data-Intensive Computing. Developed predictive models leveraging LightGBM for real-time forecasting, built CNN-based classifiers achieving over 95% accuracy, and designed LSTM-based forecasting and NLP systems utilizing transformers and attention mechanisms.

WORK EXPERIENCE

Data Scientist II — **Aug 2021 - Jun 2024**
Wolters Kluwer — **Pune, India**

- Developed patent-pending Borrower Analytics - Collateral Intel platform ($6M+ revenue), winner of Global Innovation Awards 2023, processing 22M+ UCC lien filings nationwide over 23 years.
- Built document classification pipelines using Document Understanding Transformer (DONUT), achieving 98%+ F1 across 20+ form categories; utilized LiLT models with regex extraction achieving 90%+ F1.
- Created POCs using open-source LLMs for extracting structured data from unstructured images; implemented workflows for model monitoring, quality assessment, and error reporting.
- Enhanced Wolters Kluwer's LegalView BillAnalyzer ($20M+ revenue), fine-tuning NLP models for billing guidelines (80% precision, 70% recall), and optimized Prior Approval module with SpaCy and FastText, boosting precision by 14% and recall by 10%.
- Designed high-performance billing rate management system processing over 1M records/minute and migrated key components to Dataiku for improved MLOps and governance.

Intern — **Jan 2021 - Jul 2021**
IBM (multiple internships) — **India**

Developed a Document Denoiser using OpenCV and CNNs, as well as classifying documents for denoising with 95.4% accuracy. Built automated pipelines using Step Functions and CloudFormation to tie S3, EMR and Redshift for the backend of an upcoming Asset Management business. Researched AI strategies and integration opportunities in Asset Management.

Data Science Intern — **Sep 2020 - Jan 2021**
Lawnics — **India**

Created Named Entity Relationship Recognizer using Spacy and BERT for Information Retrieval of Legal Documents and achieved an accuracy of 85%+ in 5 target values (Cases, Sections, Acts, Articles and Citations) and implemented a model for ranking of documents based on similarity index of the query using BM25 and indexed them into ElasticSearch.

PROJECTS

Text-to-ASL Video Generation using Diffusion Models

Developed an end-to-end deep learning pipeline for generating realistic American Sign Language (ASL) videos from text inputs. Leveraged Stable Diffusion fine-tuned with Low-Rank Adaptation (LoRA) to efficiently produce accurate, fluent, and visually coherent ASL gestures. Implemented Mediapipe for precise 2D skeleton extraction, ensuring consistent gesture representations. Enhanced video fluency through sophisticated frame blending and transition strategies, achieving high-quality sentence-level ASL videos suitable for accessibility applications, educational tools, and assistive technology.

Memory-Augmented Multi-Hop QA System on HotpotQA

Built a multi-hop QA system on HotpotQA using a Memory-Augmented Neural Network with BERT embeddings, dual LSTM controllers, and a Neural Turing Machine-style memory for reasoning across paragraphs. Integrated BM25/BERTScore for paragraph retrieval and optimized training with mixed precision and learning rate scheduling. Achieved 62.5% Exact Match and 90.5% F1 on the validation set.

PUBLICATIONS

A Novel Approach for Identifying Social Media Posts Indicative of Depression on **IEEE Explore** — **Jan 2021**

Developed an LSTM-based NLP model leveraging word embeddings and word2vec to accurately detect depression-indicative posts on social media forums.

SKILLS

Programming Languages: Python

Frameworks and Tools: PyTorch, TensorFlow, NumPy, Scikit-Learn, SpaCy, FastText, Dataiku, Apache Solr, ElasticSearch

Machine Learning: NLP, GenAI, Computer Vision

Cloud Technologies: AWS, Azure, Dataiku, Docker, GitHub