

ANKUR DEBNATH

+91-7002296838 | [Email](#) | [LinkedIn](#) | [Google Scholar](#) | [GitHub](#)

PROFESSIONAL SUMMARY

AI/ML leader with 5+ years of experience building and scaling end-to-end AI platforms in enterprise environments. Led teams of 5+ Data Scientists and Engineers while architecting production systems processing millions of documents, generating \$15M revenue opportunities and achieving 100+ FTE cost savings. Proven track record in technical leadership and team mentorship, combining deep expertise in LLMs, GANs, NLP, Speech AI, Time Series and production AI systems with research innovation (4 publications, 3+ patents) and operational excellence.

CORE COMPETENCIES

Data Science: Large Language Models, RAG, Deep Research (Agentic), ASR, TTS, Natural Language Processing, Machine Learning, Deep Learning, Tensorflow, Keras, PySpark, Pytorch, Big Data

Programming: Python (NumPy, SciPy, Matplotlib, Pandas, Scikit Learn), C, C++

Others: Celery, Redis, Kafka, Websockets, LangChain, CrewAI, LlamaIndex, PEFT, FAISS, Pinecone, Docker, pgvector, Web Automation, SQL, Impala, Hadoop, Apache Airflow, FastAPI, Flask, Git, Microsoft Office Suite, LaTeX, Maths and Statistics skills

Cloud: AWS (Lambda, EventBridge, Sagemaker, API Gateway, Kubernetes, S3, ECR, EC2, SQS, CloudWatch)

Frontend: Streamlit

KEY ACHIEVEMENTS & AWARDS

Filed 3+ US patents on AI innovations including Generative Time Series, Web Data Extraction and Authorization Optimization

4 publications in premier venues: KDD, CIKM, ECML-PKDD, IEEE INDICON

S&P Global Inventor Award and 2H'23 EEA Data Innovation Champion Award

99.17 percentile in GATE 2018 Electrical Engineering (among 1.5M+ candidates)

Secured 154 Rank in Assam CEE 2013

EXPERIENCE

Lead Data Scientist, Senior Data Scientist

S&P Global - Enterprise Data Organisation

Sep 2022 – Present

Gurugram, India

AI Live Transcripts (Ongoing)

- Designed, developed and productionalized a live transcription system to transcribe Corporate Calls for the product in S&P Capital IQ Pro
- OpenAI's Whisper Large V3 as ASR, DIART for real-time diarization, and NVIDIA's Titanet large for speaker embeddings and identification.
- Comprises of Audio Sourcing (Web hosted calls), Modular multi tenant Websocket-based Server-Client architecture dockerized and deployed in Kubernetes.
- An end-to-end automation pipeline designed on Human-out-of-the-loop principle to reduce high human dependency on existing product, targeted towards an opportunity of **\$12 Million every year**
- Currently exploring architectures like DiarizeLM and leverage LLMs for transcript post-processing and correction.

DocSense - Document Relevancy & Routing (Ongoing)

- Designed and deployed a platform for document routing, replacing legacy Content Collection Workflow
- **80% reduction in noise** and **65k+ man hours saved** as a part of DocSense platform
- Comprises of Document Deduplication, Document Relevancy Model Zoo, Document Routing, Continual Learning and Model Monitoring.
- A single platform for encapsulating various document routing use-cases
- Uses a rich tech-stack including, LLMs, transformer-based models, VectorDB, Grafana for Model Monitoring, Celery-Redis Asynchronous Task Management and Kubernetes for containerized deployment

Project Nova - Controversy Tracking & Bundling (Ongoing)

- Extension of DocSense Platform : An innovative solution to track ESG related controversies and bundle updates from news articles over time
- An end-to-end automation pipeline designed to replace existing 3rd party vendor and automate HITL workflow , targeted towards an opportunity of **\$2.5 Million every year**
- Comprises of Article Deduplication, Relevancy Model, NER (Tagging impacted companies), Entity Linking (with S&P Company IDs, Online Clustering over document embeddings.
- A hybrid system involving both small and large language models, RAG and pgvector for document retrieval and embeddings from LLM. Also includes, NER and entity linking achieved by prompt-based LLM extraction.

Web IQ - Bank Rates, ETF Benchmarking Solutions

- Successfully built and deployed end-to-end bank rates extraction from websites and alert on any change detected over time
- Covered **153K urls and 2.49M domains** with over **90% success rate** with increase of **27% TAT**
- Designed and deployed an ETF data extraction pipeline following the same lines
- Avoided cost of **40 FTEs** and increased dataset update frequency from half-yearly(legacy manual process) to daily. Also increased coverage of 300 ETFs to **7000 ETFs on per-day basis**
- Used **graph embeddings** for identifying areas of interest and track changes on webpage over time

Research and Leadership

- Lead a team of **5 Data Scientists and Data Engineers** and mentored and supported them towards upskilling and being productive
- Initiated and contributed towards development of critical boilerplate pipelines for asynchronous task management, web data extraction and pipeline orchestration. The same is now being used for creating large and scalable pipelines
- Filed a system patent based on Web IQ **"System and Method for Web Data Extraction Using Meta-path Graph Data Model"**

Data Scientist 2 - AI Specialist, Product Development and Innovation

MasterCard, AI Garage

Aug 2020 – Sep 2022

Gurugram, India

Product design and development of "ESG Q&A System For Investors" use case

- Successfully built and prototyped end-to-end Neural Q&A system on Mastercard reports
- Used deep learning techniques including **BERT**, advanced document preprocessing and softly deployed as a Streamlit application

Product Innovation on "Intelligent Retry Scores" use case

- Modelling customer behaviour on failed retrieval recurring payments across merchants like Apple, Google, Amazon, Netflix & Spotify
- Using tree models as well as deep models for predicting days to approval for customers across merchants
- Managed stakeholders to understand their requirements and present efficient solution

Research and Leadership

- Led research thread on **temporal data synthesis using GANs**
- Published two intermediary works as workshop papers at premier venues **KDD and ECML-PKDD**
- Lead a team on this thread, mentoring freshers, interns and peers towards quality research
- Worked on another research thread on **Neural Temporal Point Processes**
- Published a **CORE-A conference paper at CIKM-2022** on the same thread
- Filed **3 US Patents** and cleared reviews on Next Transaction prediction, and Synthetic Time Series Generation, currently under process

EDUCATION

Master of Technology in Electrical Engineering

Indian Institute of Science, CGPA : 8.0/10.0

Aug 2018 – June 2020

Bengaluru, India

Bachelor of Engineering in Electrical Engineering

Assam Engineering College, CGPA : 7.82/10.0

Aug 2013 – May 2017

Guwahati, Assam, India

PUBLICATIONS & PATENTS

Modeling Inter-Dependence Between Time and Mark in Multivariate Temporal Point Processes (2022)

Waghmare G., **Debnath A.**, Asthana S., and Malhotra A.

31st ACM International Conference on Information & Knowledge Management, CIKM 2022

Adversarial Generation of Temporal Data: A Critique on Fidelity of Synthetic Data (2021)

Debnath A., Gupta N., Waghmare G., Wadhwa H., Asthana S., Arora A.

Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML PKDD 2021

Exploring Generative Data Augmentation in Multivariate Time Series Forecasting: Opportunities and Challenges (2021)

A Debnath, G Waghmare, H Wadhwa, S Asthana and A Arora

MileTS'21: 7th KDD Workshop on Mining and Learning from Time Series

Low-Resource End-to-end Sanskrit TTS using Tacotron2, WaveGlow and Transfer Learning (2020)

A. Debnath, S. S. Patil, G. Nadiger and R. A. Ganesan

2020 IEEE 17th India Council International Conference (INDICON)

Artificial intelligence based methods and systems for improving accuracy of authorization optimizer (2024)

A Debnath, Ammar A Khan, A Arora, A Gupta, G V Waghmare, H Wadhwa, L Dheekollu, S Asthana

US Patent Application 18/473,737

ACADEMIC PROJECTS

- Good Quality End-to-end Hindi TTS using Tacotron2, WaveGlow and Transfer Learning** Aug 2019 – June 2020
- Designed, develop and test an end-to-end TTS for Hindi in a well-known voice
 - With only 20 hrs of speech, able to achieve MOS of 4.45
- Text-Independent Speaker Identification using Scattering Transform** Jan 2019 – May 2019
- Text-independent speaker identification is attempted using features generated by scattering transform, achieved **accuracy of 99.78%**
- Relation Extraction using Neural Network** Jan 2019 – May 2019
- Analyzed different architectures like **PCNN, BiGRU and GCN**, that have been shown to be very good in Relation Extraction
- Explainable Deep Learning** Jan 2019 – May 2019
- Addressed problem of model explainability, focussed on local and global explanations in image classification task

ACHIEVEMENTS

Secured 99.17 percentile in GATE 2018 Electrical Engineering among 1.5 M + Students
Secured 10.0 CGPA in CBSE AISSE 2011

CERTIFICATIONS & ACCOLADES

S&P Global Inventor: Awarded to individuals in recognition of their commitment to protecting S&P Global's intellectual property and filing their S&P Global patent with the U.S. Patent & Trademark Office.

LLM/GenAI Engineer/Architect Certification: This certificate is awarded for acquiring knowledge around large language model topics for technical engineer and architect roles (S&P Global)

LLM/GenAI Business Practitioner: This certificate is awarded for acquiring knowledge around Generative AI and large language model topics for non-technical roles. (S&P Global)

2H'23 EEA Data - Innovation Champion Award: The Essential Excellence Awards (EEA) set high-performance standards, values accountability and reward teamwork, ownership, and results across Market Intelligence Data. Recipients of this special achievement badge make outstanding contributions aimed at transforming operational efficiencies.