Amitabha Dey

Greensboro, NC 27405
amitabhadey.github.io

amitabhadey.drive@gmail.com
linkedin/amitabhadey
github/amitabhadey

Education

- **University of North Carolina at Greensboro** — Greensboro, North Carolina
  Master of Science in Computer Science — *August 2021 - May 2023*
  - Relevant Coursework: CSC 454: Algorithm Analysis/Design, CSC605: Data Science, CSC610: Big Data and Machine Learning, CSC640: Software Engineering, CSC656: Foundations of Computer Science, CSC671: Advanced Database Systems, IAF605: Data Visualization, STA631: Introduction to Probability
  - Thesis: *LDEB: Label Digitization with Emotion Binarization and Machine Learning for Emotion Recognition in Conversational Dialogues* supervised by Dr. Shan Suthaharan.

- **BRAC University** — Dhaka, Bangladesh
  Bachelor of Science in Computer Science — *August 2013 - December 2017*
  - Relevant Coursework: CSE220: Data Structures, CSE221: Algorithms, CSE230: Discrete Mathematics, CSE310: Object-Oriented Programming, CSE320: Data Communications, CSE321: Operating Systems, CSE330: Numerical Methods, CSE340: Computer Architecture, CSE331: Automata and Computability, CSE370: Database Systems, CSE421: Computer Networks, CSE422: Artificial Intelligence, CSE423: Computer Graphics, CS470: Software Engineering, STA201: Elements of Statistics and Probability
  - Thesis: *Fake News Pattern Recognition using Linguistic Analysis* supervised by Dr. Amitabha Chakrabarty.

Work Experience

- **Optum (UnitedHealth Group)** — Greensboro, North Carolina
  Senior Data Scientist — *August 2025 - Present*
  - Lead a specialized pod advancing healthcare-focused Large Language Models (LLMs) — architecting, fine-tuning, and deploying domain-adapted models optimized for long-sequence processing, automated medical coding, and clinical text understanding. Work includes domain adaptation, retrieval-augmented generation, and evaluation across provider–patient communication, coding accuracy, and compliance-sensitive use cases.
  - Engineer multi-cloud, large-scale training and inference pipelines leveraging AWS SageMaker, Azure AI Foundry, and distributed GPU clusters (e.g., NC80adis_H100_v5) with advanced Python automation, CI/CD orchestration, and containerized workflows. Pipelines integrate large-scale data ingestion, preprocessing, model training, versioning, and inference optimization across heterogeneous compute environments.

- **University of North Carolina at Greensboro** — Greensboro, North Carolina
  Lecturer, Department of Computer Science — *August 2023 - July 2025*
  - CSC 105: Data, Computing, and Quantitative Reasoning: FA23; CSC 250: Foundations of Computer Science I: FA23; CSC 261: Computer Organization & Assembly Language: SP25; CSC 350: Foundations of Computer Science II: FA22, SP23; CSC 330: Advanced Data Structures: SP24; CSC 362: System Programming: SP24, FA24, SP25
  - Mentored students in academic development, providing guidance on research, programming practices, and career pathways in computing. Coordinated multiple course offerings each semester, balancing teaching, curriculum planning, and departmental service responsibilities.

- **DevResonance Ltd.** — Dhaka, Bangladesh
  Data Scientist — *January 2018 - May 2020*
  - Utilized Python to implement a CNN model on 1TB of unstructured data, performed PCA and other dimensionality reduction techniques to reduce process time by 20% and improved classification accuracy by 15% by optimizing loss function. Increased customer retention rate by 13% as a result of these improvements.
  - Generated dynamic and interactive 3D visualizations with both linear and non-linear trendlines by integrating Plotly and Streamlit to allow clients to evaluate the impacts of interventions and monitor progress. Won grants of over $50,000 from Bill & Melinda Gates Foundation, UNICEF, WHO, Save the Children, etc.

- **Redgreen Corporation** — Dhaka, Bangladesh
  Data Science Intern — *July 2017 - September 2017*
  - Developed and implemented predictive regression models to project future sales by constructing feature space, performing data-preprocessing steps, and doing PCA resulting in an 87% accuracy, 12% better than previous years.
  - Built models to predict the possibility of faulty products and identify the manufacturers responsible. Cutting these manufacturers reduced the number of faulty components in the next quarter by 35% and increased MRR by $5K/mo. Developed a marketing analytics metrics dashboard to monitor sales conversion rate from Facebook Ads.

Technical Skills

- **Languages:** Python (advanced), R, SQL (PostgreSQL, MySQL), Java, C/C++, JavaScript (Node.js), PHP, MATLAB, Bash
- **Frameworks:** Django, Flask, Streamlit, FastAPI
- **Libraries:** TensorFlow (TF2.x/Keras), PyTorch (Lightning), Scikit-learn, XGBoost, LightGBM, CatBoost
- **Advanced Topics:** Contrastive Learning, Self-Supervised Learning, Adversarial ML, Active Learning
- **Transformer Architectures:** BERT, RoBERTa, BioBERT, ClinicalBERT, SciBERT, FinBERT, T5, mT5, mBART, XLNet, Longformer, DeBERTa, GPT-2/3/4, LLAMA, Mistral, Claude, Gemini, DeepSeek
- **Toolkits:** HuggingFace Transformers & Datasets, spaCy, NLTK, TextBlob, AllenNLP, Gensim
- **Healthcare NLP:** cTAKES, medSpaCy, MetaMap, BlueBERT, FLAIR, Stanza, ScispaCy
- **Information Extraction:** NER, RE, Co-reference resolution, Entity Linking (with UMLS/MeSH vocab), clinical assertion status classification
- **Computer Vision:** OpenCV, PIL, Tesseract OCR, PyTorchCV, Detectron2, Image segmentation (U-Net), medical imaging preprocessing (DICOM/NIfTI with nibabel & pydicom)
- **CI/CD Tools:** GitHub Actions, Jenkins, MLflow, DVC, Airflow, Prefect
- **Serving:** ONNX, TorchServe, TensorFlow Serving, Triton Inference Server, FastAPI + Docker
- **Monitoring:** Prometheus + Grafana, EvidentlyAI, Seldon Core, BentoML
- **Packaging:** Docker, Kubernetes, Helm Charts, Conda/Pipenv/Poetry
- **Model Explainability:** SHAP, LIME, Captum (PyTorch), ELI5
- **Clouds:** AWS (SageMaker, Athena, Redshift, Comprehend Medical), GCP (Vertex AI, BigQuery, Dataflow), Azure (Azure ML Studio, Data Lake)
- **Big Data:** Apache Spark, Hadoop, Kafka, Hive
- **Data Processing:** Pandas, Dask, NumPy, Vaex, PyArrow, Polars, Modin
- **Database & Query Systems:** PostgreSQL, MongoDB, Redis, Neo4j (Graph DB), Elasticsearch, Pinecone (Vector DB), Weaviate, Milvus
- **Healthcare/Clinical Standards:** FHIR (Fast Healthcare Interoperability Resources), HL7, OMOP CDM, HIPAA Compliance, PHI Masking, De-identification techniques (Scrubber, Presidio, Philter)
- **Visualization & Dashboards:** Plotly Dash, Tableau, PowerBI, Seaborn, Matplotlib, D3.js, Streamlit, Panel, Altair

Selected Publications

- **Dey, Amitabha**, et al. "Fake news pattern recognition using linguistic analysis." 2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision& Pattern Recognition (icIVPR). IEEE, 2018.
- **Dey, Amitabha**, and Shan Suthaharan. "LDEB–Label Digitization with Emotion Binarization and Machine Learning for Emotion Recognition in Conversational Dialogues." arXiv preprint arXiv:2306.02193 (2023).

Selected Research

- **Multimodal Transformers for Cross-domain Document Retrieval:** Architected a dual-encoder transformer-based retrieval system designed for large-scale, cross-domain document collections. Leveraging OpenAI's CLIP for visual representation and fine-tuned BERT/RoBERTa models for textual embeddings, I constructed a shared vector space enabling efficient cross-modal retrieval. The fusion of textual and image modalities was achieved using cross-attention transformers, allowing dense alignment of semantically linked content. Training involved contrastive loss with dynamic hard-negative mining on academic and clinical corpora, significantly improving retrieval precision on multimodal benchmark datasets.

- **Knowledge Distillation for Efficient Transformer Compression:** To reduce inference latency and deployment costs in real-world NLP applications, I implemented knowledge distillation from large-scale transformer models such as T5 and BART into a lightweight student model based on DistilBERT. The distillation pipeline incorporated multi-objective optimization combining cross-entropy on student logits with Kullback-Leibler divergence on teacher-student output distributions and intermediate layer matching. This approach was validated on CNN/DailyMail and XSum datasets for abstractive summarization tasks, where it preserved over 90% of the teacher's performance with nearly 60% reduction in model size and compute requirements.

- **Adversarially Robust Named Entity Recognition:** This work focused on improving NER robustness against adversarial perturbations by integrating adversarial training techniques into transformer-based models. A BERT-based NER model was fine-tuned using adversarial samples generated via TextAttack's gradient-based and black-box strategies including FGSM, HotFlip, and PWWS. These samples were injected during training to enhance generalization under distributional shifts. The model was further improved by incorporating dynamic embedding noise injection and evaluated on the OntoNotes 5.0 and BioNLP NER datasets, where it demonstrated marked resilience and improved F1 scores under adversarial conditions.

- **Unsupervised Machine Translation with mBART and Dual Learning:** Developed an end-to-end unsupervised neural machine translation framework leveraging multilingual pretraining via mBART combined with back-translation and denoising autoencoding strategies. This framework supported zero-resource translation pairs and employed a dual-learning loop with cycle-consistency loss and dynamic data augmentation. Fine-tuning involved iterative pseudo-labeling and translation round-trip consistency on noisy parallel corpora. This setup showed promising BLEU score improvements in low-resource translation pairs, particularly for medically relevant domain-specific datasets

- **Contextual Financial Sentiment Modeling with FinBERT:** For this research, I customized and fine-tuned FinBERT to capture context-rich sentiment signals within financial texts such as news headlines, earnings reports, and analyst commentary. The model was extended to perform multi-task learning by jointly training on sentiment classification and auxiliary financial tasks such as volatility prediction and risk tagging using FiQA and MarketNews datasets. I implemented domain-adaptive pretraining and contextualized attention mechanisms to identify key financial triggers and investor sentiment indicators, resulting in a robust pipeline capable of outperforming classical rule-based and keyword-centric sentiment models.

Selected Projects

- **Clinical Outcomes Prediction via Multimodal EHR:** Developed a multimodal deep learning system to predict patient outcomes, such as 30-day readmission risk and in-hospital mortality, by integrating structured EHR (tabular) data with unstructured clinical notes. TabNet was used for modeling structured features, while ClinicalBERT followed by a Bi-LSTM and attention mechanism was used to extract signals from clinical narratives. A late fusion architecture combined the two modalities, and optimization involved focal loss with oversampling techniques like SMOTE to manage class imbalance. The pipeline was tested on MIMIC-III datasets and demonstrated state-of-the-art performance across all metrics.

- **Loan Default Risk Modeling using Elastic Net and Survival Analysis:** Performed end-to-end modeling for predicting the likelihood and timing of loan defaults using a combination of linear models (Ridge, Lasso, Elastic Net), Gradient Boosted Decision Trees, and survival analysis techniques such as Cox Proportional Hazards. I engineered time-sensitive financial features, used recursive feature elimination for dimensionality reduction, and calibrated model probabilities using Platt scaling. Model performance was validated with ROC-AUC and Brier scores, where survival models provided additional insights into default timing distributions.

- **Voice-Controlled Intelligent Assistant (Jarvis):** Designed and developed a fully offline-compatible, voice-controlled virtual assistant named Jarvis using Python libraries such as gTTS, SpeechRecognition, and custom-built intent parsers. The system used a keyword extraction-based NLU module supported by fuzzy logic to handle ambiguous queries. Functionalities included music playback on YouTube, weather updates via API calls, real-time news aggregation, and context-aware responses. All components were modularized and integrated into a CLI/GUI hybrid interface, demonstrating real-time responsiveness even under low-bandwidth conditions.

- **Sentiment Analysis of Financial News using LSTM with Attention:** This project involved building a sentiment classification system for financial news articles to assist in short-term stock movement prediction. I employed a BiLSTM model augmented with a word-level attention mechanism to capture contextual dependencies in token sequences. The dataset was preprocessed using advanced NLP techniques including stopword removal, token normalization, and TF-IDF filtering. Pre-trained GloVe vectors were used to initialize embeddings. The model was trained with Adam optimizer and dropout regularization, achieving substantial accuracy improvements over baseline SVM and logistic regression models.

- **Legal Domain NER using BERT and CRF Decoding:** Developed a Named Entity Recognition system tailored for legal documents by fine-tuning LegalBERT on a labeled dataset comprising contracts, court judgments, and regulatory texts. To improve sequence labeling consistency, I added a Conditional Random Field (CRF) layer on top of the transformer encoder to capture label transitions and dependencies. The model was benchmarked against legal-specific tag sets and demonstrated enhanced entity precision in extracting contractual obligations, named parties, and references to statutes, achieving over 12% improvement over BiLSTM-CRF baselines.

Selected Honors and Awards

- **UNCG Outstanding Graduate Student Award (2023):** Awarded the most prestigious award by the Department of Computer Science for the academic year 2022-23 in recognition of scholarly accomplishment and contribution to the department.
- **UNCG Merit Scholarship (2021):** Awarded $16,000 for 14 months and In-state and Out-of-state full tuition waiver by the Department of Computer Science and the Graduate School.
- **The Daily Star Award (2010):** Awarded the National Daily Star Award for Edexcel IGCSE Students in 2010 for academic results - Further Mathematics (A*), Mathematics (A*), Chemistry (A*), Physics (A*), Economics (A), English (A), Bengali (A).
- **Positions Held:** Student Senator (UNCG Graduate Student Association), Department Central Advisor, Scholarship Coordinator, Study Abroad Coordinator, Industrial Advisory Board Liaison, Advisor to UNCG Bangladesh Student Association and Nepali Student Association