Vinamra Baghel

IBM Research | Indian Institute of Technology Bombay | LinkedIn | Email: baghelvinamra@gmail.com | +91 7977905377

Education

Indian Institute of Technology Bombay

Dual Degree (B.Tech + M.Tech), Electrical Engineering (CGPA: **8.56**) [2019 – 2024]

M.Tech Specialization: Communication and Signal Processing (CSP), **Minor**: Artificial Intelligence and Data Science (C-MInDS)

Experience

Research Engineer, IBM Research

[Aug 2024 – Present]

Time Series Foundation Models

Time Series Semantic Intelligence Agent

- Built a real-time **Kafka** & **Flink** streaming pipeline for **Industry 4.0** TSFM multivariate **anomaly detection** and **semantic search**
- Developed **ReAct**-based **LangGraph** agent with an **MCP** server for automated root cause investigation and failure mode analysis
- Designed a **Neo4j knowledge graph**-based context layer with **Skills** to ground agent reasoning over fleets of industrial assets
- Catalyzed industry collaboration with **Confluent** for building TS Semantic Intelligence Agents for plants and data centers

Data Evaluation Framework for Forecasting Time Series Foundation Models (TSFMs) **[*ICASSP 2026*]**

- Developed a **KNN**-based algorithm for optimal pretraining data **subset selection**, identifying influential and **diverse** samples
- Designed a low-dimensional representation space using **temporal**, **statistical**, and **spectral** features and **TSFM embeddings**
- Characterized pretraining (PT) data **diversity** for data selection by relating neighborhood **distances** with **forecasting** metrics
- Implemented and **optimized robust** attribute estimators and embedding extractors handling ill-defined statistics and **outliers**
- Achieved **3% improvement** in forecasting performance with over **50% reduction** in PT corpus, tested across TSFM architectures

Targeted Synthetic Data Generation for Pre-training Corpora Enhancement for TSFMs **[*CODS 2026*]**

- Quantified gaps in PT corpus using **KNN** distance metrics by projecting time series to representation **TSFM embedding** spaces
- Mapped series to an **FFT** space and enriched PT corpus via targeted **synthetic data generation** through inverse-FFT transform
- Led extensive pretraining and methodology evaluation across **Mixer**, **Transformer**, and **State Space-based** TSFM architectures
- Demonstrated up to **9% improvement** in **MASE** on **in-distribution** (GIFT-Eval) and **out-of-distribution** (BOOMLET) benchmarks

Climate and Sustainability – Accelerated Discovery (AI)

Spatio-temporal CO_2 Flux Estimation from Sparse Remote-Sensing Data using Open-sourced Proxies **[*US Patent*]**

- Developed a method to estimate **CO_2 emission flux** on a high-resolution grid using **remote-sensed** and **ground-level** proxies
- Resolved **satellite NO_2** observation sparsity using **RandomForest**-based data **imputation** to obtain high res. **XCO_2** conc. maps
- Ideated a **novel** approach to spatio-temporally allocate CO_2 **flux** using a **UNet**-based architecture with temporal **self-attention**

Research Intern, IBM Research

Data Driven Decarbonization for Enterprise Sustainability [May 2023 – Aug 2023]

- Developed a **counterfactual explainable** recommendation system using **DiCE** to guide enterprise sustainability initiatives
- Clustered enterprises by emission behavior via **K-Means** on *Carbon Disclosure Project* reports for peer-based benchmarking
- Boosted CO_2e emission estimation **MAPE** via feature selection (**ANOVA F-test**) on open-sourced data for an **XGBoostRegressor**
- Awarded a **Letter of Recommendation** from the Mentor for exemplary performance and foundational contribution to **ESG-FM**

Summer Internship Hackathon - CCC Bot

- **1st runner-up** in a 48-hour hackathon by building a natural-language to computing-cluster-command translator bot
- Created a dataset and fine-tuned **Flan-T5** to map natural language to a structured **syntax schema** to generate commands
- Demonstrated the bot via a user-friendly **Textual** terminal interface, substantially enhancing the lab-wide cluster usability

Summer Technology Analyst, Morgan Stanley (Institutional Securities Technology)

Flaky Test Filtering for Healthy Software Development Life Cycle (SDLC) [May 2022 – Jul 2022]

- Architected a **Scala** and **Java**-based classification engine for a CI platform on the **Optimus** framework for fast flaky test filtering
- Slashed spurious test failures by **80%** and cut CI effort by **40%** across a large monolithic repository improving **decision reliability**
- Spearheaded department-wide **rollout** of the engine for the Fixed Income Group and presented the work to key leadership
- Received **Pre-Internship Offer** (continuing internship) from the firm for distinguished work and key **SDLC** contributions

Major Research Projects

Uncertainty Modelling for Open Domain Generalization [May 2023 – Jun 2024]

M.Tech Thesis – Prof. Amit Sethi (Medical Deep Learning and Artificial Intelligence Lab)

- Investigated **Uncertainty Estimation** under domain shift to enable robust open-set recognition in generalized **vision** models
- Characterized the behaviors of **entropy**, **dropout**, and **ensemble**-based uncertainty metrics across diverse natural datasets
- Modeled the metrics as a function of **noise**, **out-of-distribution** samples, and **unknown labels** by analyzing the distributions
- Benchmarked behavior across SOTA **augmentation**-based, **generative**, **meta-learning**, and **ensemble** domain gen. models
- Revealed **distinct uncertainty signatures** across in-distribution and out-of-distribution samples, and unknown classes
- Studied how uncertainty can improve reliable deployment of domain-generalized vision models for pathology applications
- Designed an uncertainty-based **decision framework** for domain shift-aware **open-set** prediction in digital pathology
- Established a systematic **framework** for evaluating uncertainty under diverse distribution shifts across models and datasets

Greedy Cluster Margin for Active Learning [Mar 2023 – Apr 2023]

Course Project – Prof. Ganesh R. (Optimization in Machine Learning)

- Devised a novel agglomerative clustering algorithm achieving **150%** enhanced time-efficient preprocessing of training data
- Executed cluster **sampling** on training corpus, identifying a **diverse** range of **low-confidence** images to reduce labelling costs
- Contributed (preliminary) to the open-source **DISTIL** library, an active learning toolkit, competing with the benchmarks

Cross-domain Semi-supervised Multi-task Learning [Aug 2022 – Dec 2022]

Semester Research Exposition – Prof. Biplab Banerjee

- Built a multi-domain, multi-task learning architecture using **domain-specific batch normalization** for semantic segmentation
- Established baselines using **VGG-16** Encoder with a shared **SegNet** decoder, trained using **cross-entropy** and **MSE** objectives
- Achieved cross-domain IoUs of **0.59** & **0.51** for cross-domain segmentation over Synscapes and Cityscapes respectively

Multi-task Learning Using Attention Networks [May 2021 – Dec 2021]

Self-learning – Prof. Biplab Banerjee

- Implemented a multi-task attention network with a shared core **SegNet** backbone and task-specific **soft-attention** modules
- Jointly trained semantic segmentation, depth estimation, and surface normal estimation using **Dynamic Weight Averaging**
- Achieved segmentation accuracies of **91.08%** on Cityscapes and **55.29%** on NYUv2 benchmarks, validating the backbone

Patents and Publications

1. US Patent: "*ESTIMATING GREENHOUSE GAS EMISSIONS USING SATELLITE DATA AND MACHINE LEARNING*"; Status: Filed
2. ICASSP 2026: "*Time Series Attributes Guided Pretraining Data Selection for Time Series Foundation Models*"; Status: Published
3. CODS 2026: "*Close the Gap: Targeted Synthetic Data to Augment Pretraining Corpus of Time Series Foundation Models*"; Status: Under review
4. NeurIPS CCAI 2025: "*Using Time Series Foundation Models for Atmospheric CO2 Concentration Forecasting*"; Status: Published
5. IGARSS 2025: "*Advancing Flood Mapping with Geo-Spatial Models and Indian Satellites*"; Status: Published

Scholastic Achievements

- Scored **705 / 805** (99th percentile) in the GMAT Focus Edition [2024]
- Awarded a **Branch Change** to Electrical Engineering Department based on **academic excellence** [2020]
- Achieved **AIR 522** (99.96 percentile) in JEE Main out of 1.14 million candidates [2019]
- Bagged **1st position** in school in the National Level Science Talent Search Examination (NSTSE) [2020]
- Certified with **High Distinction** (1st in school) in the Australian National Chemistry Quiz (ANCQ) [2020]

Important Coursework

Optimization in ML, Advanced ML, Foundations of Intelligent and Learning Agents (RL), ML for Remote Sensing, Stochastic Optimization, Probability and Stochastic Processes, Markov Chains and Queuing Systems, Linear Algebra, Game Theory and Algorithmic Mechanism Design

Skills

Python, Time Series Analysis, Foundation Models, Machine Learning, Remote Sensing, Agentic AI, Probability and Statistics

Research Interests

Foundation Models, Time Series Analysis, Representation Learning, Multi-modal ML, CV, Reinforcement Learning, AI Alignment

Positions of Responsibility

- Institute Student Mentorship Program (ISMP) – Subgroup Head and Department Academic Mentor [Jun 2023 – Jun 2024]
- Teaching Assistantship – Intro to (EE 769) and Advanced (EE 782) Machine Learning [Jul 2023 – Apr 2024]
- FinSearch Mentor - Deep Reinforcement Learning for Strategy Optimization [Jun 2023 – Dec 2023]