Manish Gawali

Portfolio Google Scholar (10 research papers, 140+ citations)

Email : gawali@usc.edu
Mobile : +1-(213)-574-6099

Education

- **University of Southern California** — Los Angeles, CA
  Master of Science in Computer Science; GPA: 3.86/4.0 — *May 2022 – Dec 2023*
- **Pune Institute of Computer Technology** — Pune, India
  Bachelor of Engineering in Computer Science; GPA: 9.20/10.0 — *Aug. 2014 – May 2018*

Full-Time Work Experience (4+ Years)

- **Amazon** — Seattle, WA
  Applied Scientist II (Domain: Large Language Models, Gen AI, Agentic AI, Recommeder Systems) — *Jan 2024 – Aug 2025*
  - Advanced LLM robustness research by leading data creation and experiment pipelines for **SetLexSem** (**NeurIPS 2024**, **20%** acceptance). The evaluation methodology was leveraged partners such as **Anthropic** to improve foundation LLMs.
  - Reduced seller listing time by **94% (6h → 20min)** by designing and deploying **Generative AI into QuickList**, used across **21 global marketplaces**, accelerating onboarding for **750 million** monthly listings.
  - Increased attribute coverage **17× (3 → 51)** with **87.5%** seller acceptance, directly achieving goal of auto-generating $> 50\%$ of required attributes.
  - Pioneered **agentic planning** for Selling Assistant with 100K active users by building **DAG-based adaptive workflows** for multi-agent orchestration, enabling context-aware task execution and laying the foundation for a **conversational e-commerce chatbot**.
  - Automated evaluation of the Seller Assistant's knowledge retrieval system via a **science-driven framework**: created a **synthetic ground-truth datastore** and introduced **retrieval metrics**, enabling scalable, repeatable benchmarking.
  - Developed a **PoC transforming retrieval** in sponsored ads recommendation, leveraging **generative transduction (HSTU)**—High-Performance Self-Attention Encoder for Generative Recommendations, arXiv:2402.17152. Improved **retrieval quality** and set the foundation for **next-gen ad personalization** at Amazon scale.
  - Strengthened Amazon's **research visibility** with multiple publications (**SetLexSem at NeurIPS 2024**; **TopoSem at KDD 2025**) and **2 internal publications** ; led efforts to **open-source SetLexSem** via Amazon Science GitHub.
- **DeepTek Medical Imaging Pvt. Ltd.** — Pune, India
  Senior Data Scientist (Tech Lead) — *Aug 2020 – Mar 2022*
  - Led **AI-driven clinical deployments** across **50+ hospitals and diagnostic centers**, productionizing **10+ deep learning models** (**UNet**, **Mask-RCNN**, **YOLO**, **Autoencoders**, **GANs**) for detecting chest, CT, and MRI pathologies — **reducing radiologist reporting time by ∼30% per study** and improving throughput in high-volume settings impacting **tens of thousands of patient scans monthly**.
  - Advanced **medical NLP capabilities** by fine-tuning **BioBERT** on multi-hospital datasets, extracting structured clinical entities (**symptoms**, **anatomical sites**, **diseases**) from millions of unstructured notes — enabling faster triage and improved diagnostic decision support.
  - Published **7 peer-reviewed research papers** (Springer, IEEE, JMIR), establishing DeepTek as a scientific leader in AI for radiology. Key contributions included: **SplitFedv3** — a novel federated learning method reducing communication overhead and training time while preserving privacy; and **Alternate mini-batch training** — a distributed strategy improving convergence stability in hospital networks.
  - Mentored **junior scientists and engineers**, driving project delivery timelines and ensuring **reproducibility standards** for deployments in regulated clinical environments.
- **AlgoAnalytics Pvt. Ltd.** — Pune, India
  Associate Consultant (Deep Learning) (Domain: Computer Vision, Natural Language Processing, LLMs) — *Jan 2020 - Jul 2020*
  - **Signature Generation**: Developed a novel variant of DCGAN (Deep Convolutional Generative Adversarial Network) architecture for generating handwritten signatures using the GPDR-960 handwritten dataset.
  - **Virtual Try-On**: Implemented a multi-stage CNN-based deep learning method for Virtual Try-On and deployed it.
  - **Title Generation for Research Papers**: Leveraged the Transformer-based GPT-2 architecture and trained it on the arXiv dataset containing academic papers across various fields and deployed it.
- **Veritas Technologies LLC** — Pune, India
  Associate Software Engineer (Domain: Full-Stack Development) — *Jul 2018 - Jul 2019*
  - **Enterprise Vault & Compliance/Discovery Accelerator**: Developed multiple features like Auditing Enhancements, Cloud-SQL, Hotword Statistics, Hotword, and Hotword Set Facet.
  - **POCs, Bug Fixes, Optimization**: Contributed to 8 winning POC's and rectified 30+ bug fixes.Optimized legacy code to increase the speed of operations like archiving e-mails and file shares by 20%

Internship and University Research Experience (2.5 YEARS)

- **Amazon** — New York City, NY
  Applied Scientist Intern (Domain: Natural Language Processing, Large Language Models, Statistics) — *Sept 2023 - Dec 2023*
  - **Product Relationship**: Trained a production Language Model XLM-RoBERTa to determine for classifying pair of products for variation relationship.
  - **Synthetic Data Generation for Cold Start Problem**: Implemented LLM as Optimizers from scratch to automatically learn best prompt needed to generate missing data, pseudo labels, and new product information using LLMs.
  - **Noisy Label Learning**: Developing a novel method based on reweighing samples to learn from noisy synthetic data and data based on outdated guidelines.
  - **LLM Confidence Scoring Framework**: Developed an automated framework to output soft labels/confidence scores for labels and synthetic products generated by LLMs.
  - **Assessment of Quality of Synthetic Products**: Used a CLIP Text Encoder to embed real products and synthetic products and designed statistical tests to determine the quality of synthetic products.
- **Amazon** — Denver, CO
  Software Development Engineer Intern (Domain: Full-Stack Development) — *May 2023 - Aug 2023*
  - **UI Page**: Created a Forecast History Page which enables the operations team to view history of uploaded forecast files along with the statistics and metadata about that forecast. This involved changing the database schema, creation of a new step function, modifying five existing lambdas, and creation of new APIs.
- **vITAL Lab and Radiomics Lab, University of Southern California** — Los Angeles, CA
  Graduate Research Assistant (Domain: Federated Learning, Medical Imaging, Computer Vision) — *May 2022 - May 2023*
  - **Effective Aggregation in Federated Learning (FL)**: Developed adaptive weighing scheme which uses information from client model updates and determines aggregation weight for that client. Used nnUnet architecture on KiTs 19 dataset.
  - **Semi-supervised Federated Learning**: Research on FL where only few silos have ground truths.
  - **Bone Metastasis**: Leveraged ViT (Vision Transformer) for segmentation of bone lesions from NiFTi CT scans.
- **GSLab** — Pune, India
  Machine Learning Intern (Domain: Natural Language Processing) — *Aug 2017 - Mar 2018*
  - **Chatbot for Automated Customer Service**: Built the backend machine learning & natural language processing pipeline using concepts like TF-IDF, Clustering, and Cosine similarity to help agents deliver better customer support achieving an accuracy of 82%. Deployed ML model as REST API using Flask as a backend service.

Projects

- **Paper Grading using Transformers** | **Technologies: Python, Google Cloud**: Created a framework to grade handwritten student answers using Google's Cloud Vision API for handwriting recognition and an NLP pipeline with BERT encoder to calculate grades on semantic similarity to model answers.
- **eYantra** | **Technologies: C++, Embedded C, Python**: Co-ordinated the synchronous motion of two robots by building path planning algorithms, to match the musical notes extracted from audio files by striking steel rods at specific nodes of the graph.

Programming Skills

- **Languages**: Python, Java, Javascript, Typescript, C, C++, C#, Embedded C, SQL
- **Libraries**: Tensorflow, Keras, PyTorch, OpenCV, Pandas, NumPy, Scikit-Learn, Gensim, spaCy, NLTK, p5.js, React.js, Redux
- **Frameworks**: Langchain, Hugging Face, Flask, .NET, AWS, Azure, PySpark
- **Developer Tools**: Git, Docker, GCP, AWS, IntelliJ, VS Code, Visual Studio, PyCharm, Eclipse, Android Studio

Publications (8 External + 2 Internal)

- **Advances in Neural Information Processing Systems 37**: "SETLEXSEM CHALLENGE: Using Set Operations to Evaluate the Lexical and Semantic Robustness of Language Models". Presented the paper at the NeurIPS 2024 conference.
- **KDD 2025 Workshop Publication**: "TopoSem: In-context planning with semantically-informed tooling graph similarity".
- **Springer Publication**: "Comparison of Privacy-Preserving Distributed Deep Learning Methods in Healthcare". Presented the paper at the MIUA 2021 conference.
- **IEEE Publication**: "Deep Learning Models for Calculation of Cardiothoracic Ratio from Chest Radiographs for Assisted Diagnosis of Cardiomegaly", icABCD 2021 conference.
- **JMIR Publication**: "Key Technology Considerations in Developing and Deploying Machine Learning Models in Clinical Radiology Practice", JMIR Medical Informatics Journal.
- **Springer Publication**: "Vulnerability Due to Training Order in Split Learning". ICT4SD 2021 conference.
- **Springer Publication**: "A deep learning approach for automated diagnosis of pulmonary embolism on computed tomographic pulmonary angiography".
- **Arxiv Pre-Print**: "Application of Federated Learning in Building a Robust COVID-19 Chest X-ray Classification Model".
- **Medrxiv Pre-Print**: "Automated assessment of chest CT severity scores in patients suspected of COVID-19 infection".

Research Services

- **Reviewer**: ML4H: Machine Learning For Health Care (MLHC) 2025.
- **Reviewer**: ML4H: Machine Learning for Health 2021.
- **Reviewer**: Privacy-Preserving Machine Learning for Medical Imaging: MICCAI 2021 Workshop and Tutorial (Springer).
- **Reviewer**: 2021 International Conference on Artificial Intelligence and its Applications (icARTI 2021) (ACM).