ANUSHRI BHANSALI

anushribhansali.14@gmail.com | https://www.linkedin.com/in/anushri-bhansali/ | https://github.com/AnushriBhansali

EDUCATION

North Carolina State University, Raleigh   **August 2023 to May 2025**
Master of Computer Science   **GPA:3.81/4**
Design Analysis of Algorithms, Automated Learning and Data Analysis, Efficient Deep Learning, Neural Networks, Real Time Machine Learning & AI, Database Management Systems, Software Engineering, Programmer Centered Design & Research

Charotar University of Science & Technology   **August 2019 to June 2023**
Bachelor of Technology in Computer Science   **GPA:9.58/10**

TECHNICAL SKILLS

AI/ML: Pytorch, Hugging Face Transformers, NumPy, Matplotlib, NumPy, Scikit-learn, CNNs, Apache TVM, Machine Learning Compilation Code, TensorFlow **Programming Languages**: Python, C++, TypeScript, Java, **Frameworks/Tools**: Flask, React, Node.js, Azure DevOps, AWS, Databricks **Database**: SQL, NoSQL (MongoDB)

WORK EXPERIENCE

Software Engineering Research Assistant, WISER Lab, North Carolina State University, United States   **Jan 2025 to May 2025**
- **Designed and developed a modular microservices-based software system** using **Spring Boot**, featuring independent services for user authentication, content management, and service discovery. This system was created as a **learning project** to support students in understanding the architecture and operations of real-world distributed systems.
- Implemented **Apache Kafka** for asynchronous messaging between services, **REST APIs** for synchronous interaction, and **Eureka service registry** for dynamic service discovery—demonstrating essential microservice communication patterns.
- Created **extensive documentation and walkthroughs** tailored to educational use, including architectural explanations, code-level insights, and setup instructions to facilitate hands-on learning through NCSU's **Virtual Computing Lab (VCL)**.
- Set up and tested **CI/CD pipelines** to simulate industry-grade software delivery practices, ensuring reproducibility and modular extensibility of the system for iterative student experimentation.
- The project aimed to **bridge theoretical coursework with practical system design**, offering students an immersive environment to engage with core principles like service orchestration, inter-process communication, and containerized deployment.

Software and AI Engineering Intern, Lexis Nexis, Raleigh, NC, United States   **June 2024 to August 2024**
- Performed the fine-tuning of **Meta Llama** and **Mistral** large language models for semantic search enhancement, achieving a **60% performance boost** in precision and response accuracy through **few-shot learning**, **custom synthetic prompt generation**, and dataset curation aligned to legal domain queries.
- Architected and deployed **LLM-powered inference pipelines** within distributed **Databricks workflows**, resulting in a **35% reduction in API response latency** and improved throughput under high-load production scenarios.
- Designed and implemented **scalable RESTful APIs** for seamless integration between LLM components and existing microservices architecture, automating complex data processing tasks and **saving over 15 man-hours per week**.
- Collaborated closely with **cross-functional teams** including product managers, data engineers, and legal domain experts to define use cases, refine model behavior, and ensure end-to-end delivery of production-ready features within agile sprints.
- Contributed to internal documentation and sprint planning, proactively identifying bottlenecks and proposing model performance metrics to optimize business outcomes in content retrieval and knowledge augmentation.

Software Engineering Research Assistant, Hints Lab, North Carolina State University, United States   **March 2024 to June 2024**
- Designed and developed a **web-based interactive learning portal** to support machine learning education, combining **visual model explanations**, **interactive quizzes**, and **real-time feedback** using **React.js** for the frontend and **Flask** for the backend.
- Architected a **PostgreSQL-backed backend system** to manage user authentication, session tracking, quiz results, and ML model data, ensuring seamless RESTful integration and secure data handling across components.
- Applied **Human-Computer Interaction (HCI) design principles** to create an intuitive user experience tailored to the cognitive needs of students, leading to a **40% improvement in student engagement** and a **35% increase in conceptual comprehension**, as measured by pre- and post-activity assessments.
- Collaborated with faculty and researchers to iteratively refine portal features based on student feedback, integrating **data visualizations** and **interactive walkthroughs** that demystify complex ML concepts like gradient descent and decision boundaries.
- The project served as a **learning enhancement tool** within the curriculum, effectively bridging traditional lectures with experiential learning through active, exploratory engagement.

Software and AI Engineering Intern, Netweb Software, India   **January 2023 to June 2023**
- **Led the design and development** of a full-stack **intelligent resume parsing and shortlisting platform**, built with **Java (Spring Boot)** on the backend, **React.js** for the frontend, and **PostgreSQL** as the core database. The system delivered **real-time data updates** and **semantic filtering**, enabling recruiters to reduce manual workload by **15+ hours per week**.

- Engineered a **vector-based semantic ranking engine** utilizing **transformer-based embeddings** and **Retrieval-Augmented Generation (RAG)** techniques, boosting the precision of candidate-job matching by **45%**. Implemented advanced similarity scoring mechanisms to surface the most relevant resumes dynamically.
- Developed and integrated **asynchronous backend APIs** for handling large-scale resume ingestion, coupled with a **responsive modular frontend**. This design scaled efficiently to support **50+ resumes per day**, contributing to a **30% increase in interview conversion rates**.
- Closely collaborated with **product stakeholders and recruiters** to refine matching criteria, gather feedback for iterative model improvements, and deliver business-aligned features under tight deadlines.
- Actively contributed to sprint planning, CI/CD workflows, and documentation, ensuring smooth handover and extensibility of the platform for future ML model integration and feature upgrades.

Machine Learning Intern, BISAG-N, Government of India, India   **May 2022 to July 2022**
- Built a robust **malware detection system** capable of identifying malicious executables using **TF-IDF vectorization**, logistic regression, and other classical machine learning models—achieving **96% classification accuracy** on real-world datasets.
- Integrated **model interpretability techniques** such as **SHAP (SHapley Additive Explanations)** and **feature importance plots**, empowering security analysts to trace threat signatures and decision boundaries in high-risk scenarios.
- Enabled **30% faster incident response times** by providing analysts with clear, model-driven insights into threat classification, thereby supporting real-time triaging and proactive security alerts.
- Collaborated with a cross-disciplinary team of researchers and cybersecurity experts to align model development with practical threat detection requirements.

Software Engineering Intern, Cygnet Infotech, India   **June 2021 to August 2021**
- Designed and developed a **production-ready data visualization dashboard** leveraging **Angular** for the frontend, **Flask** for backend logic, and **PostgreSQL** for structured data management. Delivered the entire solution within **1.5 months**, ahead of schedule.
- Implemented advanced **filtering, charting, and reporting features** to provide real-time business insights, integrating RESTful APIs and asynchronous data fetching for seamless UI responsiveness.
- Achieved **100% feature completion with zero critical bugs**, exceeding delivery expectations and earning **company-wide recognition** for exceptional adaptability and technical execution under tight timelines.
- Quickly ramped up on unfamiliar technologies, independently researching and applying best practices in UI/UX and backend development, leading to reduced QA cycles and expedited deployment.

PROJECTS

License Plate Recognition Model Optimization
- Optimized **LPRNet**, a license plate recognition model, for **edge device deployment**, reducing **inference time from 0.25s to 0.028s** and compressing model size from **1.11MB to 0.51MB** using **unstructured pruning** and **quantization**.
- Applied **post-training quantization techniques** to minimize latency while maintaining competitive accuracy, achieving **82.3% accuracy with a 0.71MB model**, and further accelerating inference using compiler-level optimizations.
- Leveraged **Apache TVM AutoScheduler**, ONNX model conversion, and **CUDA backend tuning**, achieving **89.4% accuracy** with optimized vectorized and parallelized compute kernels—enabling near real-time performance on low-resource environments.
- Demonstrated deep integration of **ML model design with systems-level optimization**, aligning with real-world constraints of embedded AI systems.

LLM-Driven Healthcare Support System
- Designed a complete **AI-powered healthcare platform** combining **disease prediction models** and a **question-answering system**, built using **RoBERTa**, **Flask**, and **Python**, aimed at improving accessibility to medical information.
- Developed a secure **React.js frontend** and structured **PostgreSQL database** to manage patient data and model responses, ensuring HIPAA-aligned separation of data layers.
- Engineered robust **RESTful APIs** to streamline communication between frontend, backend, and ML components, resulting in a **20% reduction in data retrieval time** and improved overall responsiveness.
- Positioned as a **proof-of-concept** for leveraging **LLMs in digital health**, demonstrating potential applications in clinical decision support and patient triage systems.

Data Driven Movie Recommendation System
- Built a personalized **movie recommendation system** using **matrix factorization (SVD)** on the **Netflix Prize dataset**, achieving **18% improvement in NDCG** through **Bayesian optimization** and hyperparameter tuning.
- Conducted in-depth **exploratory data analysis (EDA)** to identify behavioral patterns and preference clusters, which guided model refinement and enhanced overall recommendation quality.
- Incorporated **evaluation metrics such as precision, recall, RMSE**, and **coverage** to measure performance across diverse user profiles and sparsity conditions.
- Demonstrated the ability to build **scalable recommender pipelines**, with potential application in user-centric content platforms, e-commerce, and personalization engines.

RESEARCH & PUBLICATIONS

LLM-Guided Architectural Refactoring in Large-Scale Software System.

- Led a **pioneering investigation into the use of Large Language Models (LLMs)** such as **GPT-4o** to support **automated software architecture analysis and refactoring** across large, highly interconnected codebases—addressing a key gap in modern software engineering tooling.
- Designed a rigorous methodology to evaluate LLM performance in identifying and classifying **14 architectural red flags** (e.g., Information Leakage, Shallow Modules, Temporal Decomposition), using over **300 real-world commits** from complex, modular systems including **Apache Kafka, Hadoop, Flink, Airflow, Pulsar, and Camel**.
- Engineered a multi-stage evaluation pipeline combining **LLM inference with post-correction validation**, revealing both the **capabilities and limitations** of LLMs in reasoning over architectural smells, semantic anti-patterns, and ripple effects across class boundaries.
- Developed a novel framework to **contextualize LLM-based predictions** by augmenting commit-level inputs with **dependency graphs, system design constraints, historical evolution data, and semantic cues**—dramatically improving classification precision and reliability.
- Conducted **comparative analysis** against expert architectural heuristics, demonstrating LLMs' potential to achieve near-expert performance in several cases, particularly when provided with **rich architectural metadata and usage context**.
- Authored a comprehensive error taxonomy highlighting the critical influence of missing structural information, and proposed a hybrid approach integrating **static analysis tooling and repository-level intelligence** with LLMs for enterprise-grade architectural recommendation systems.
- Positioned this work as a **first step toward building intelligent, AI-driven development assistants** capable of system-wide architectural insight—moving beyond code linting to **holistic architectural evolution guidance**.

Interference Classification for Open RAN (O-RAN) xApp.

- Developed a **deep learning-based xApp** for **jammer detection** in O-RAN systems, utilizing **spectrograms from OTA testbed data** on the Near-RT RIC.
- Applied advanced **embedding space augmentation techniques** such as **E-Mixup** and **E-Stitchup** to enhance generalization under varying noise and interference conditions common in open radio networks.
- Trained and benchmarked a range of models including **ResNet, DenseNet, MobileNetV2, ViT**, and **ConvNeXt**, followed by comprehensive evaluation through **t-SNE latent space analysis** to visualize and interpret model decision boundaries.
- The xApp was designed for real-time deployment within the **Near-RT RIC architecture**, enabling proactive spectrum interference detection in 5G/6G networks.

Language Identification Using Machine Learning & Vectorization Techniques

- Designed and implemented a **multi-language identification system** using classical machine learning approaches, trained on **vectorized phonetic inputs** derived from spoken language samples.
- Explored a variety of vectorization and feature extraction techniques to build language-aware models capable of distinguishing between phonetically similar languages with high precision.
- **PUBLISHED under State Government of Gujarat SSIP research initiative**.

Melanoma Skin Cancer Cell Detection

- Conducted a **systematic review of machine learning and image processing approaches** for the detection of **melanoma skin cancer**, a highly fatal but curable disease if diagnosed early.
- Analyzed over a dozen peer-reviewed models and proposed a performance comparison framework based on **accuracy, sensitivity, specificity, and precision**, across classical ML methods, SVM-based pipelines, and deep learning models including CNNs and transfer learning architectures.
- Identified critical preprocessing steps such as **hair and noise removal, RGB-to-grayscale conversion**, and **adaptive threshold segmentation**, and highlighted their impact on improving lesion detection from **dermoscopic images**.
- Emphasized the role of **GLCM texture features**, **color asymmetry**, and **shape irregularity** in accurate feature extraction, and compared classification efficacy using SVM, Random Forest, and hybrid CNN techniques.
- Co-authored the paper titled "A Systematic Review on Melanoma Skin Cancer Cell Detection," which was selected and **published by IEEE** as part of ICECAA 2022, underscoring the research's relevance to the **intersection of healthcare and AI**.