Shubham Sharma

amnour.rajsubham@gmail.com | shubh8434.github.io | linkedin.com/in/shubh8434

Education

Panjab University, Chandigarh **2020 – 2024**
Bachelor of Engineering (Hons.), Department of Computer Science & Engineering CGPA : 9.23/10
Rank 1 in a class of **65** students

Publications

- S. Ray[1*], **Shubham Sharma**[1*], S. Aditya, P. Goyal, *"EduVidQA: Generating and Evaluating Long-form Answers to Student Questions based on Lecture Videos"*, *EMNLP 2025 (Core A*)*. [**Paper**] [**Code**]
- A. Singh[1*], V. Gangwar[1*], **Shubham Sharma**[2], S. Saha, *"Knowing What and How: A Multi-modal Aspect-Based Framework for Complaint Detection"*, *ECIR 2023 (Core A)*. [**Paper**] [**Code**]
- **Shubham Sharma**[1*], S. Mukherjee[1*], D. Kaplun[2], R. Sarkar, *"Pneumonia Detection in Chest X-Rays using XGBoost based Meta-learner with Deep Feature Extractors"*, *CPAMCS 2023, Springer, Cham.* [**Paper**] [**Code**]
- D. Prakash[1*], Raghvendra K[1*], **Shubham Sharma**[2], S. Saha, *"IndicBART alongside Visual Element: Multimodal Summarization in Diverse Indian Languages"*, *ICDAR 2024 (Core A)*. [**Paper**] [**Code**]
- **Shubham Sharma**[1], G. K. Walia[2*], K. Singh[2*], V. Batra[3], A. K. Sekhon, A. Kumar, K. Rawal, D. Ghai, *"Forecasting of Crop Yield Using Various Machine Learning Approaches: A Comparison"*, *Rural Sustainability Research, Vol. 52 (347), 2024.* [**Paper**] [**Code**]
- S. Bamber[1], A. Katkuri[2*], **Shubham Sharma**[2*], M. Angurala, *"A Hybrid CNN-LSTM Approach for Intelligent Cyber Intrusion Detection System"*, *Computers & Security (Elsevier, IF: 5.2, Q1), 2024.* [**Paper**] [**Code**]

Work Experience

Machine Learning Engineer I, Recursive Softpro (Bengaluru) **Jan 2025 – Present**

- **AI-Driven RAG Framework for Multimodal Retrieval with LLM Generation Response**
  - Built a pipeline that retrieves multimodal data (text, images, audio, and video frames) using CLIP embeddings and ChromaDB, ensuring efficient retrieval based on user queries.
  - Integrated GPT-4o for generating context-aware responses by leveraging retrieved documents and media, enhancing user interaction and relevance.
  - Utilized Base64 encoding to embed visual data (images, frames) into LLM prompts and fine-tuned retrieval thresholds for accuracy and latency trade-offs.
- **Intelligent Multi-Agent Orchestrator with RAG and Tooling (Django, OpenAI, Pinecone)**
  - Engineered a Swarm-driven multi-agent router with OpenAI function calling, dynamic tool resolution (registry + user-defined tools), and safe agent hand-offs.
  - Implemented end-to-end RAG using Pinecone (knowledge-base selection, top-k retrieval) and OpenAI chat completions with configurable parameters, tone, and language.
  - Integrated Django ORM and DRF for chat persistence, analytics endpoints, and conversation-load monitoring for real-time system health.
- **AI Marketplace (GenAI • Agentic AI • Workflow Use Cases)**
  - Building a marketplace for customizable AI workflows powered by LLM agents and composable microservices.
  - Developed a unified LLM integration layer (OpenAI, Anthropic, Llama) with multi-tenant authentication, API-key management, logging, and CI/CD for scalable model onboarding.
  - Implemented an end-to-end RAG pipeline with document ingestion, chunking, embeddings, and vector search (Chroma/Weaviate/FAISS), enabling "Ask-Your-Docs" APIs via LangChain.
  - Designed LangGraph-based multi-agent workflows with reusable agent templates, LangMem-backed memory, and evaluation harnesses for accuracy, latency, and cost.

Language Research Analyst, Indian Institute of Technology–Kharagpur **July 2024 – Dec 2024**
Advisor: Prof. Somak Aditya | LLM Based Educational Videos Question Answering System

- Developed a data generation pipeline integrating manual annotation and LLM-assisted question synthesis (Google Gemini 1.5 Flash), creating over 4,300 high-quality Q&A pairs across lecture domains.
- Benchmarked state-of-the-art multimodal LLMs (Video-LLaMA, MPlug-Owl, etc.) under zero-shot, few-shot, and fine-tuned configurations to evaluate reasoning depth and generalization.
- Performed both quantitative (BLEU, ROUGE, METEOR) and qualitative (student preference) evaluations to assess answer relevance and pedagogical alignment.
- Authored paper on educational multimodal video question answering, accepted for publication at EMNLP 2025.

Research Experience

Carnegie Mellon University (CMU), USA **Sept 2023 – Jul 2024**

Advisor: Prof. Min Xu | 3D Cryo-ET classification using self-supervised learning

- Integrated Momentum Contrast for Unsupervised Visual Representation Learning (MoCo) with Video Vision Transformer (ViVIT) and ViT using weight inflation for 3D Cryo-ET classification.
- Proposed a novel pipeline for processing 3D data, resulting in improved performance and training efficiency for the classification task.
- Achieved an F1 score of 70.28 in open-set environments using ViVIT for classifying real Cryo-ET data and noise, setting a new benchmark in the field.
- Contributed to a research manuscript on self-supervised 3D Cryo-ET classification, involving architecture design, experimentation, and writing of related works.
- Written a book chapter on feature semantic segmentation, including membrane, template matching, and deep segmentation (In Review).

Indian Institute of Technology-Patna (Under ACM Research Internship) **May 2023 – Sept 2023**

Advisor: Prof. Sriparna Saha | IndicBART: Multimodal Summarization in Diverse Indian Languages

- Collected a multimodal dataset (Hindi, Tamil, Bengali, Marathi) with paragraphs, summaries, and images from Large Scale Multi-Lingual Multi-Modal Summarization Dataset (M3LS).
- Extended the BART model with a visual-aware encoder to generate regionally contextualized summaries using both text and image data.
- Achieved a ROUGE-1 score of 0.266 on the Hindi dataset and 44.1% image precision using an image pointer for multi-output prediction.
- Work accepted at ICDAR 2024 based on multimodal summarization research with Indian regional languages.

Jadavpur University, Kolkata **Jan 2023 – May 2023**

Advisor: Prof. Ram Sarkar | Pneumonia Detection Using Meta-learner With Deep Feature Extractors

- Compiled and preprocessed a Mendeley dataset with three classes (Covid-19, Pneumonia, Normal), applying data augmentation techniques to address overfitting.
- Extracted image features using Vision Transformer (ViT) and ResNet50 models.
- Developed an ensemble model by combining features from both ViT and ResNet50, and used XGBoost for classification, achieving 96.19% accuracy.
- Contributed to writing the results section of the paper, which was accepted at CPAMCS-2023.

Indian Institute of Technology-Patna **Aug 2022 – Jan 2023**

Advisor: Prof. Sriparna Saha | Multimodal Aspect Based Complaint Detection

- Collected a multimodal dataset of product images, customer reviews, and aspects from Amazon website.
- Extracted textual embeddings using BERT and visual representations using pre-trained CNN models such as VGG16 and ResNet.
- Developed a multimodal interaction model to learn the relationship between text, image, and product aspects.
- Achieved accurate complaint eligibility classification, with the work published at the European Conference on Information Retrieval (ECIR) 2023.

National University of Singapore **Jan 2022 – Aug 2022**

Advisor: Prof. Luo Wei | Vessel Collision and Trajectory Prediction

- Utilized the U.S. Coast Vessel Traffic System dataset for Automatic Identification System (AIS) data analysis.
- Preprocessed and cleaned key navigational parameters including Latitude, Longitude, Speed Over Ground (SOG), Course Over Ground (COG), and Maritime Mobile Service Identity (MMSI).
- Applied clustering algorithms such as DBSCAN and K-Means to segment vessel trajectories and remove outliers.
- Developed and evaluated sequence-to-sequence models (RNN, LSTM, Bi-LSTM, GRU, Bi-GRU) for predicting vessel paths and assessing collision risks.

Indian Institute of Technology-BHU **Oct 2021 – Jan 2022**

Advisor: Prof. Rajesh Kumar | Concrete Mix Design

- Designed and optimized concrete mix compositions by varying sand fines, cement, and plasticizer ratios to enhance strength, density, and durability.
- Implemented and compared multiple machine learning regression models — Multiple Linear Regression, Support Vector Regression (SVR), Decision Tree Regression (DTR), and Artificial Neural Network (ANN) — to predict compressive strength.
- Achieved the highest accuracy for both 7-day and 28-day strength predictions using the Decision Tree Regression model.

Scholastic Achievements

- Recipient of Medal & Prizes for consistently securing the **highest CGPA** in the department since the **first year**.
- Selected for the prestigious **ACM IKDD Uplink Research Internship 2023** (Acceptance Rate: 2%).
- Accepted to attend **IIIT Hyderabad's CVIT** and **Amazon Machine Learning Summer School** in 2023.
- Achieved **State Rank 1143** out of 1 million students in the Hindustan Olympiad.
- Received a scholarship from the **Macquarie EdX Group**, out of 35,000 applicants.
- Advanced to the Pre-Elimination Round in the **Codechef Smackdown Coding Competition 2021**.

Academic Projects

Multi-Agent LLM System for Automated App Development | Github **Aug 2025 - Oct 2025**

- Built a planner → architect → coder LangGraph pipeline that turns plain-language specs into runnable project scaffolds with controlled recursion.
- Integrated Groq's openai/gpt-oss-120b through LangChain with Pydantic outputs to keep task plans deterministic.
- Hardened the tool layer for sandboxed file and command access, enabling safe, automated generation of multi-file web app prototypes.

AI-Powered Knowledge Graph Builder | Github **Jan 2025 - April 2025**

- Built a knowledge-graph backend combining FastAPI with Ollama/OpenAI LLM pipelines for automatic entity and relationship extraction from unstructured documents.
- Engineered secure, production-grade ingestion (MIME validation, path traversal defense, suspicious-content scanning) with version-controlled graph history and REST APIs for upload, retrieval, and lifecycle management.
- Optimized persistence via SQLAlchemy models, Alembic migrations, and batched database operations, yielding low-latency queries plus detailed stats/monitoring endpoints.

Intrusion Detection System | Github **Aug 2023 - Dec 2023**

- Developed a deep learning-based IDS using the NSL-KDD dataset, optimizing feature selection with RFE and a Decision Tree classifier.
- Achieved 95% accuracy and 0.94 F1-score with the CNN-LSTM model, demonstrating superior performance in intrusion detection.
- Research accepted for publication in Computers & Security (Elsevier, Q1, IF: 5.2), validating the model's effectiveness and industry relevance.

E-Commerce Product Recommendation System | Github **Dec 2021 - Feb 2022**

- Leveraged a pretrained ResNet50 convolutional backbone with Global Max Pooling to encode images into dense feature embeddings.
- Applied L2 normalization on the embeddings and used scikit-learn's brute-force k-Nearest Neighbors (Euclidean distance) search to rank visually similar catalog items.
- Created a Streamlit-based interface that uploads customer images, infers embeddings on the fly, and displays the top recommendations in real time.

Relevant Coursework

- Machine Learning
- Deep Learning
- Database Management
- Data Structures
- Analysis & Design of Algorithms
- Operating System
- OOPS
- Probability & Statistics
- Linear Algebra
- Natural Language Processing
- Calculus
- Digital Image Processing
- Artificial Intelligence
- Computer Graphics

Technical Skills

Languages: Python, C/C++, SQL, HTML, CSS, JavaScript, MATLAB, LaTeX
Frameworks & Library: PyTorch, TensorFlow, Hugging Face, LangChain, LangGraph, FastAPI, Django, Keras, Streamlit, OpenCV, Scikit-learn, NumPy, Pandas, Matplotlib, Seaborn, SQLAlchemy
Environment/Tools: Docker, Git/GitHub, Linux, CUDA, Anaconda, Google Colab, Jupyter, Tableau, MLflow, Weights & Biases
Databases: PostgreSQL, ChromaDB, Pinecone, FAISS, Weaviate

Extracurriculars

Google Developer Student Club | Machine Learning Lead **Jan 2022 - Dec 2023**

- Led a team of **20+ members** in organizing **10+ workshops and events** focused on machine learning and AI, benefiting **500+ students** across the university.
- Guided **10+ technical projects**, mentoring students in applying ML models to real-world problems and providing **hands-on learning** experiences.

National Service Scheme (NSS) | Volunteer **Feb 2023 - Jan 2024**

- Coordinated various **community service projects**, engaging with 100+ students to address **local societal issues** and promote social welfare.
- Organized **awareness campaigns** focused on health, education, and environmental sustainability, reaching over **200 local residents**.