

Shib Kumar Saraf

+91 7001486011 — contact@shibkumar.dev — shibkumarsaraf — shib111111 — shibkumar.dev — Orcid Profile

AI Engineer with 2.5+ years of experience delivering production-grade AI/ML solutions in **Agentic Workflows, Enterprise Gen AI , RAG Architectures & Computer Vision** to drive automation, insights, and scalable business impact.

Skills

- **Generative AI & LLMs:** Agentic AI, RAG Architectures, Fine-Tuning, LangChain, LlamaIndex, Hugging Face, OpenAI API, Vector Databases (Pinecone, ChromaDB, Faiss), MLOps, Vertex AI, Gemini, Model Deployment (Ollama, vLLM)
- **Traditional AI & ML:** Machine Learning, Deep Learning, NLP, Computer Vision, Model Training, Feature Engineering, Predictive Modeling, Time-Series Forecasting, OpenCV, Diffusion models
- **Data Engineering & Analytics:** Exploratory Data Analysis (EDA), Data Cleaning, Data Pipelines, ETL, PySpark, ELK Stack, Pandas, NumPy, SQL, BigQuery, Large-Scale Data Processing
- **Programming & Microservices:** Python, Bash, FastAPI, Flask, REST APIs, Asynchronous Processing, API Orchestration
- **Cloud Platforms, Systems Automation:** Google Cloud Platform (GCP), Linux (Ubuntu, Debian), Cloud-Native Architectures, Docker, Jenkins, CI/CD, Workflow Automation

Experience

- **EXL Service (India) Private Limited** Oct 2025 – Present
Senior Executive (Business Analyst)
 - Conducted **actuarial factor analysis** (age, gender, geo-area) to evaluate customer churn patterns for CVS Health (Aetna) insurance products, supporting data-driven retention and pricing strategies.
 - Executed data operations and advanced analytics to present **churn insights** to plan sponsors, contributing to an increase in customer retention from 49% to 60%.
- **Centre for Development of Advanced Computing (C-DAC), Chennai** Jul 2024 – Oct 2025
Project Associate (Lead AI Engineer)
 - Designed and deployed AI-driven **chatbot platforms** (**ChatDB, FileQAExpert**) leveraging LLMs and vector-based search, serving 10K+ monthly queries across web, Telegram, and WhatsApp.
 - Developed **real-time analytics systems**, incorporating anomaly detection, suspicious behavioral profiling, & collusion detection, for high-stakes government examinations, enhancing monitoring efficiency by 40% and reducing manual effort by 60% under the **AI-powered Secure Exam Analytics** project.
 - Engineered AI-driven secure end-to-end **financial workflow applications** for claims processing and reporting automation, ensuring data integrity and compliance within the **QDAS project**.
 - Led **data migration and automated reporting** initiatives, enabling faster decision-making and enhanced operational efficiency for **RailTel**.
- **Ministry of Electronics and Information Technology (Govt. of India)** Jan 2024 – Jun 2024
Big Data & Data Analytics Intern
 - Designed and developed **Social Sentinel**, a real-time NLP-powered system for monitoring, classifying, and flagging offensive content on Telegram, incorporating scalable data ingestion pipelines and analytics dashboards.
 - Conducted applied NLP and ML research on clinical risk prediction under **Health Analytics Project**, leveraging transformer models and topic modeling techniques on real-world Chinese hospital data.
- **Solvitude Solutions Private Limited** May 2021 – Nov 2021
Subject Matter Expert & Quality Assurance Analyst
 - Automated **reporting workflows** and developed **structured question banks and analytical reports** to streamline **product evaluation, training, and QA processes**, while testing, validating, and documenting hardware and software systems in collaboration with **cross-functional teams** to improve **reporting accuracy and operational efficiency**.

Research Publications

- **Multifaceted Feature Extraction and Machine Learning for Mortality Risk Prediction in Pediatric ICUs**
DOI: 10.1109/ICSTSDG61998.2024.11026232 **IEEE ICSTSDG 2024, Chennai, India**
Developed an ML-driven clinical risk prediction framework using advanced feature engineering to improve mortality assessment in pediatric intensive care units.
- **Hybrid QLSTM Framework for Time-Series Forecasting in Dynamic Financial Markets**
DOI: 10.1109/ICIMIA67127.2025.11200988 **IEEE ICIMIA 2025, Tirupur, India**
Designed a hybrid Quantum-inspired LSTM architecture to improve forecasting accuracy under highly volatile financial market conditions.

- **Real-Time Vigilance: Advanced Video Analysis for Violence Detection in Public Spaces**

DOI: 10.1109/ICRITO66076.2025.11241501

IEEE ICRITO 2025, Noida NCR, India

Built a real-time computer vision pipeline leveraging deep learning models for automated violence detection to support public safety monitoring systems.

Personal Projects

- **QueryCortex AI** — An Agentic AI Reasoning & QnA Platform  Oct 2025 – Dec 2025
 - Built a secure **Agentic AI Orchestration Platform** using FastAPI with JWT authentication, enabling safe multi-tenant enterprise data access.
 - Engineered a hybrid reasoning pipeline combining **NL2SQL** and **vector search** to answer natural language queries across **databases and documents with explainable outputs**.
- **CodeGuard AI** — AI-Powered Secure Code Review Platform  Jan 2025 – Mar 2025
 - Developed a full-stack **MERN application** integrating **Google Generative AI** to automatically analyze code for quality, security vulnerabilities, performance issues, and best practices.
 - Designed **REST APIs** for AI-driven reviews and built a **real-time React-based code editor** with syntax highlighting and structured Markdown feedback.
- **MindMetrics** — Mental Health Insights & Behavioral Analytics  Feb 2024 – Apr 2024
 - Built an **LLM-driven NLP system** leveraging transformer models (**BERT, DistilBERT, XLNet, GPT, Gemini**), zero-shot classification, and **topic modeling (LDA, NMF)** to detect mental health signals from high-volume social media streams.
 - Designed scalable data pipelines for text preprocessing, embedding generation, sentiment analysis, and hybrid rule-based modeling to uncover behavioral patterns and emerging mental health trends.
- **SereneSounds** — Personalized Music Curation Engine  Sep 2023 – Dec 2023
 - Designed a hybrid recommendation engine using **Collaborative Filtering(based on genre, artist, lyrics, etc.), Alternating Least Squares(ALS), and Graph-based similarity models** to personalize music discovery.
 - Improved recommendation relevance by integrating lyrical embeddings and audio feature vectors, deployed via an interactive Streamlit interface.

Education

- **Pondicherry University**
Master of Computer Applications Nov 2022 – Jun 2024
CGPA: 9.35/10
- **Indian Institute of Technology Madras**
Diploma in Data Science Aug 2021 – May 2023
CGPA: 8.8/10
- **Panskura Banamali College, Vidyasagar University (Affiliated)**
B.Sc. in Physics (Honours) Jul 2019 – Aug 2022
CGPA: 9.08/10

Certifications

- **Google Digital Leader** — Coursera (2024)
- **Data Science and Machine Learning using Python** — NIELIT (2023)
- **IoT Software Analyst** — NIELIT (2023)
- **Deep Learning for Computer Vision** — IIT Hyderabad (NPTEL, 2023)
- **Artificial Intelligence for Economics** — IIT Kharagpur (NPTEL, 2025)
- **Introduction to MongoDB** — MongoDB University (2022)
- **Advanced Excel Training** — Udemy (2021)

Achievements

- **Best Paper Award** — Received for the paper titled "*Real-Time Vigilance: Advanced Video Analysis for Identifying Violence in Public Spaces*" presented at **IEEE ICRITO 2025**.
- **Best Project Award** — Awarded for the project in the *Modern Application Development* course at **IIT Madras**.
- **Top 3 Rank Holder** — Secured a position among the top 3 students in the MCA program at **Pondicherry University** with a CGPA of 9.35.
- **Swami Vivekananda Merit cum Means Scholarship (SVMCM)** (Mar 2015 – Aug 2022) — Awarded by the *Government of West Bengal* for consistent academic excellence from Class XI to B.Sc. Physics.
- **Multiple Badges** — Earned top course grades and finisher badges in various AI, ML, and data science programs, demonstrating commitment to continuous learning and technical excellence.