

Sooraj Sreekumar

sooraj12sreekumar@gmail.com | +91 8122152267 | [LinkedIn](#) | [GitHub](#)

Summary

Data Science & AI engineer with hands-on experience in NLP, Computer Vision, OCR, and Generative AI. Built production-ready ML and LLM-powered applications using Transformers, PyTorch, and Streamlit, backed by industry internship experience.

Education

Vellore Institute of Technology- MTech Integrated in Computer Science and Engineering specialization in Computational and Data Science

Sept 2021 – May 2026

CGPA: 8.37/10

Technical Skills

Programming: Python, SQL.

ML / AI: Scikit-learn, TensorFlow, HuggingFace Transformers.

NLP & CV: NLTK, OpenCV, OCR (Tesseract, PaddleOCR).

Data & Visualization: Pandas, NumPy, Excel, Google Sheets, Tableau, Power BI.

Tools: Git, Streamlit, LabelStudio, LabelImg.

Experience

Data Analyst Intern, Kumaran Systems, Chennai

Sept 2024 – Jan 2025

- Built and curated high-quality datasets by annotating **5,000+ images**, improving data quality and downstream analytical reliability for OCR-based systems. extraction efficiency.
- Collected, cleaned, and standardized **10,000+ images records**, establishing a scalable data preprocessing and quality assurance workflow.
- Automated text extraction from noisy, real-world images using OCR pipelines, enabling structured data generation from unstructured sources.
- Tech Stack:** Python, OpenCV, Pillow, PyTesseract, PaddleOCR, PyOCR, LabelStudio, LabelImg.

Projects

Resume Classification System Using BERT [\[GitHub-link\]](#)

June 2025 – Aug 2025

- Built an automated resume classification system using **DistilBERT**, trained on **14,000+ resumes**, to categorize resumes across multiple job roles with **78% accuracy** on unseen data.
- Implemented a modular training and evaluation workflow with saved tokenizers, label encoders, and model checkpoints, enabling reproducible and scalable inference.
- Tech Stack:** Python, HuggingFace Transformers, DistilBERT, Pandas, Numpy, Scikit-learn, Git.

AI-Powered Study Assistant [\[GitHub-link\]](#)

July 2025 – Aug 2025

- Built a document-aware AI study assistant using **Google GenAI (Gemini)**, enabling users to upload PDFs and generate **summaries, quizzes, and context-aware Q&A**.
- Designed a modular backend architecture for PDF parsing, **prompt-driven LLM interaction**, and dynamic quiz generation, integrated with a Streamlit frontend.
- Tech Stack:** Python, Streamlit, Google Gemini API, PyMuPDF, dotenv, GitHub.

Multi-Sensor Carbon Monoxide Prognosis System [\[GitHub-link\]](#)

May 2025 – June 2025

- Developed and validated a Carbon Monoxide prediction system using multi-sensor IoT data, achieving **$R^2=0.9963$** for accurate environmental risk forecasting.
- Deployed a **Streamlit-based web application** for real-time CO concentration prediction, featuring user-driven inputs and clear risk-level interpretations for actionable safety decisions.
- Tech Stack:** Python, Pandas, NumPy, Scikit-learn, Streamlit, Matplotlib, Seaborn.

Certifications and Trainings

Generative AI Experience Certification, **Finlatics**

July 2025 - Sept 2025

Statistics for Data Science and Business Analysis, **Udemy**

July 2023 - Aug 2023

The Data Scientist Toolbox, **Coursera** (Johns Hopkins University)

Nov 2022 - Dec 2022