# DANISH FAYAZ

DATA SCIENTIST

## CONTACT


9149574409

Danishfayaz09@gmail.com

Baramulla, J&K, India


---

## SKILLS

**Programming Languages:**
- Python, SQL

**Machine Learning & Deep Learning:**
- TensorFlow, PyTorch, Scikit-learn

**Natural Language Processing:**
- Huggingface, Langchain, Haystack, Ragas, Crewai, Langgraph

**Computer Vision:**
- Detectron2, Ultralytics, Supervision

**Cloud Services:**
- AWS Lambda, S3, Textract, Lex,

**Development & Deployment:**
- Docker, Django, Flask

**Data Visualization & Monitoring:**
- Tableau, Grafana

**Experiment Tracking:**
- Wandb

**Annotation Tools:**
- Roboflow, Label Studio

---

## EDUCATION

Bachelors of Computer Science and Engineering

### National Institute of Technology Srinagar

2018-2022 **CGPA: 8.72**

---

## COURSES

**AI Data Engineering**
- Data Engineering Lifecycle,
- Batch & Streaming Pipelines, Storage Architectures
- ETL Pipelines, DataOps

**Multi AI Agent Systems with crewAI**
- Agents, Tools, Crew, Manager, Memory, Process

**Machine Learning Engineering for Production(MLOPS)** .
- Data pipelines, ML system pipelines, Conceft drift,
- Model analysis and Interpretability, NAS
- Shadow , Blue Green, Canary Deployment.

**AWS Cloud Technical Essentials**
- AWS services such as EC2, Lambda functions, DynamoDB, S3, CloudWatch and Sagemaker

## PROFILE

A results-driven Data Scientist with expertise in designing and deploying end-to-end machine learning and natural language processing solutions. Adept at tackling complex challenges across healthcare, finance, and enterprise automation domains by leveraging cutting-edge techniques in deep learning, NLP, and computer vision. Skilled in implementing scalable systems for data extraction, analysis, and visualization, utilizing frameworks such as TensorFlow, PyTorch, and AWS cloud services. Proficient in Python and SQL, with a strong foundation in building robust ETL pipelines, fine-tuning advanced models.

---

## WORK EXPERIENCE

### Deep Learning Engineer

StarSoftware — Dec 2022-Present

#### Automated Document Processing & Extraction System

**NLP, OCR, Computer Vision | Incremental Learning | Transformers & YOLO**

- Designed and deployed an end-to-end automation pipeline to extract structured data from documents such as invoices, bank statements, transcripts, bills of lading, COAs, and MTRs.
- Built and fine-tuned 54+ NER labels using transformer models including BERT, RoBERTa, and XLM-RoBERTa, enabling accurate extraction of fields like invoice numbers, dates, and company details.
- Implemented incremental learning with mixed sampling of old and new data to reduce forgetting, and integrated Elastic Weight Consolidation (EWC) to support continual learning.
- Combined OCR tools (AWS Textract, PyTesseract) with LayoutLM and YOLOv5/v8 for layout-aware extraction of tables and key-value pairs.
- Ran extensive evaluations with Detectron2 and Ultralytics, optimizing object detection across varied document structures using ResNet and ResNeXt backbones.
- Applied LoRA and QLoRA to reduce training time and memory usage in both NER and object detection workflows.
- Developed reusable ETL and validation pipelines with human-in-the-loop (HITL) review, boosting precision across formats.
- Achieved 10x cost reduction and significant accuracy gains, leading to faster delivery cycles and new client acquisition.

#### Intelligent Business Document Insight & Analytics Platform

**LLM Integration | LangGraph Pipelines | Postgres | ERP & Invoice Analysis**

- Building an advanced AI-driven platform to extract insights from complex business data sources such as invoices, COAs, Material Test Reports (MTRs), and ERP systems.
- Integrated LLMs (OpenAI, Qwen) for natural language analysis, enabling dynamic SQL query generation, query rewriting, error recovery, and response-type classification.
- Architected a LangGraph-based modular workflow, allowing for fine-grained control over stages like validation, error handling, and output post-processing.
- Connected the system to a PostgreSQL backend to support real-time querying and seamless frontend API integration.
- Actively enhancing the system for better performance, optimization, and scalability, with a focus on accuracy, speed, and maintainability.
- Designed the architecture to support natural language–driven data exploration, empowering business users to generate custom reports and dashboards without SQL knowledge.

**Generative Adversarial Networks specialization (GAN)**

- Discriminator, Generator , Adversarial Trianining.
- DGAN, WCGAN.

**Deep Learning specialization**

- Neural Network, Hyperparameters tuning, Optimisers.
- CNN,RNN,LSTM,GRU,Transformer,Bert

## PROJECTS

### RAG-Based QA System

- Integrated Retrieval-Augmented Generation (RAG) architecture within Langchain framework, incorporating advanced components:
- Implemented efficient retrieval mechanisms, parent document retriever, hybrid search, RAG fusion.
- Employed sophisticated techniques to decompose user queries for better understanding and translation.
- Designed prompt-based templates for answer generation using GPT-3.5, GPT-4, Llama 7B, and Cohere.
- Conducted rigorous testing and validation,using RAGAS scores for optimizing each component for enhanced system performance.
- Currently in progress of deploying the system on cloud infrastructure for scalability and reliability, focusing on seamless handling of diverse query loads and user satisfaction.

### Facial Emotion Recognition

- Developed a deep learning model for emotion recognition.
- Focused on a 7-class classification problem, distinguishing Happy, Sad, Neutral, Surprise, Angry, Disgust, and Fear emotions.
- Leveraged advanced neural network architectures and image processing techniques to achieve high classification accuracy.
- Results demonstrate the ability to accurately classify emotions with potential applications in fields such as healthcare, marketing, and entertainment.

### Data Scientist

TheTransmogrify — July 2022- Nov 2022

- Diabetes Insight Tracker
- Leveraged Named Entity Recognition (NER) to extract valuable content from textual posts within applications, focusing on posts related to the health of Diabetes patients on the website.
- Applied Optical Character Recognition (OCR) to detect and analyze glucose levels from image posts within the application, specifically targeting posts related to the health status of users managing diabetes.
- Stored the extracted content efficiently in a database (PostgresSQL), ensuring organized and accessible data for further analysis and insights.
- Utilized a visualization tool (Grafana) to create visualizations for trends and events within the database, enhancing data interpretation and decision-making for monitoring health-related patterns.
- Integrated analytics (Google Analytics) to capture and analyze user events on the website, gaining valuable insights into user interactions and engagement, and contributing to the improvement of the overall platform.

## INTERNSHIPS

### Natural Language Processing Intern

XivTech — Dec -Mar 2022

- Utilizing AWS Textract, Comprehend and OCR for extracting invoices.
- CI/CD pipeline for Invoice Extraction using Lambda Function and S3
- Employing a neural network and BERT for document classification into invoices, bills, etc

### Research Intern

Indian Institute of Science, Bangalore — July -Sep 2021

- Detection of floating plastic through the utilization of remote sensing and deep learning techniques.
- Employing a U-Net model for the segmentation of floating plastic.

### Machine Learning Intern

Indian Institute of Technology,Jammu — Mar-June 2021

- Applied Machine learning and Deep learning models on CICIDS datasets.
- Comparing the effectiveness of random selection and column subset selection methods.
- Few Shot learning using Convolutional Neural Network

### Deep Learning Intern

Indian Institute of Technology, Bhubaneswar — Jan-March 2021

- Object detection and classification using deep learning models.
- Employing a Unet model for image segmentation.