# Shrenik Jain

+1-(858)-241-1904 | shrenik-jain.github.io | shrenikkjain81@gmail.com | linkedin/shrenik-jain9 | github/shrenik-jain

## Education

**University of California San Diego** **Present**
Master of Science, Electrical and Computer Engineering - Machine Learning and Data Science
***Coursework:*** *Recommender Systems & Web Mining, Probability & Statistics, Statistical Learning*

**Vishwakarma Institute of Information Technology** **May 2022**
Bachelor of Technology, Electrical Engineering - GPA: 3.98/4.0
***Coursework:*** *Data Structures, Design & Analysis of Algorithms, Machine Learning, Deep Learning, Image & Video Processing, Cloud Computing, Internet of Things (IoT)*

## Experience

**Applied Researcher**, Cosman Lab, UC San Diego **Oct 2024 - Present**
- Spearheading the development of a novel region-of-interest-based video compression method by integrating human attention on top of **Google's** baseline **Video Compression Transformer (VCT)**.

**Machine Learning Engineer**, Pivotchain Solutions **Jul 2022 - Aug 2024**
- Led a cross-functional team to design an in-house AI System, developing **Computer Vision** algorithms to provide a scalable and real-time interface for autonomous event monitoring mitigating security risks and incident response time by 70%.
- Implemented a **Spatio-Temporal Autoencoder** model to validate AI-generated video clips, precisely classifying **15,000+** daily security events, reducing false positives by 30% and saving **$250K** annually in operational costs.
- Optimized **MongoDB** queries to efficiently handle over **10 million+** records of video surveillance JSON data, reducing data retrieval times by 55% and improving overall system performance for end-users across multiple regions.

**Software Development Intern**, Qualys Inc. **Jan 2022 - July 2022**
- Designed multi-stage **CI/CD** pipelines using Groovy-based declarative pipelines and containerized builds, to streamline workflows and cut average deployment time from 30 minutes to 10 minutes.
- Led the deployment orchestration of policy-compliant microservices across 3 major environments using Kubernetes, ensuring 95% uptime (equating to less than 2 hours of downtime per quarter).

**Applied Researcher**, Vishwakarma Institute of Information Technology **Jul 2021 - Dec 2021**
- Led the development of a research paper summarization system using a **BERT**-based encoder, improving ROUGE-1 scores from 0.35 to 0.46 compared to baseline extractive summarization methods.
- Conducted research on **transformers** and multi-head self-attention mechanisms, for enhanced language modeling.

**Machine Learning Intern**, Validus Analytics LLP **Feb 2021 - Dec 2021**
- Analyzed Vector-Quantized VAEs (VQ-VAEs) and Convolutional VAEs (Conv-VAEs) for unsupervised learning of complex data via latent representations and **generative** modeling.
- Implemented **ConvVAE**-based generative modeling for dataset enhancement, expanding a critical training dataset from 50,000 to 150,000 samples while maintaining high perceptual integrity (SSIM $> 0.85$).

## Consulting Experience

**Machine Learning Consultant**, Pixstory **Aug 2023 - Mar 2024**
- Contributed to building a **RAG**-based Conversational Search System using **Large Language Models (LLMs)**, improving search relevance and increasing average user session duration by 2 minutes.
- Optimized system throughput and hardware efficiency by **3x** by introducing asynchronous requests and parallelized execution, reducing average query response time from 2 seconds to 600 milliseconds.

**Software Development Consultant**, AI For Rural **Sep 2021 - Nov 2021**
- Implemented data preprocessing and visualization pipelines for insightful data handling and exploration.
- Developed REST APIs and integrated them with various data sources, enabling real-time data updates and reducing data retrieval time by 40% for critical information.

## Technical Skills

**Languages**: Python, Java, JAX, JavaScript, C++, Shell Scripting, Bash, SQL, HTML, CSS
**Machine Learning**: Tensorflow, PyTorch, TFLite, Keras, LangChain, CUDA, Scikit-learn, OpenCV, NLTK, SpaCy, ONNX Runtime, TorchServe, TritonServer, TF-Serving, Hugging Face Transformers
**Frameworks & Technologies**: Flask, SpringBoot, PySpark, Git, FFmpeg, Docker, Kubernetes, Jenkins, REST, Linux/Unix, ONVIF, Apache Spark, Kafka
**Cloud & Databases**: AWS, GCP, Azure, MongoDB, SQL, Elasticsearch, Milvus, Vector Stores, Cassandra, RabbitMQ, MQTT

## Projects

**Face Physiognomy** | *Python, Flask, Docker, Git*
- Designed an emotion recognition system, integrating Haar cascades for real-time face detection and a CNN-based classifier for emotion analysis, achieving a 92% accuracy rate across 7 distinct emotion categories.
- Optimized the model's inference speed to handle large-scale input, reducing classification latency by 30%.