Akash Kumar Singh

+91-9955223293 | Gmail | Webpage | Github | LinkedIn | Scholar

Education

Indian Institute Of Technology (IIT) Delhi — Delhi, India
M.S.(R) in Machine Intelligence & Data Science — *Jan 2024 - Dec 2025 (Expected)*

Indian Institute Of Science Education & Research (IISER) Bhopal — Bhopal, MP, India
B.S in Data Science & Engineering — *Aug 2019 - May 2023*

- CPI: 8.10/10
- **BS Thesis:**
  - **Title:** *CLSNet: Improving Food Classification using Contrastive Learning with Layer Selection*
  - **Advisor**: Prof. Akshay Agarwal

Sr. Secondary Examination (TSBIE) – (May 2016 - April 2018) — Hyderabad, Telangana, India

- Aggregate: 96.2%
- Ranked 4^{th} in the school & the top 4 percentile in the Board

Secondary Examination (CBSE) – (May 2014 - May 2016) — Saharsa, Bihar, India

- CGPA: 9.4/10

Internships

Data Scientist Intern - Trademo | Aug 2023 - Jan 2024
Location: Gurgaon, India

- **Worked** on **Recommender System** for **Supply Chain** based on **LLM** & **Information Retrieval** approaches.
- **Increased** the top 3-hit percentage by **16%**.
- **Worked** on developing a **Chatbot** using Langchain.

Research Internship - Indian Institute of Science (IISc) Bangalore | May 2023 - July 2023
Mentor: Prof. Punit Rathore, CiSTUP, IISc Bangalore, India

- **Developed** a real-time warning system for collision avoidance in a driving simulator using **YOLO** for **ADAS** (Advanced Driver Assistance Systems).
- Worked on **Smartphone Data Analytics** for road anomaly detection system using Machine Learning on **Time Series Data**.
  - Performed **Systematic Literature Review** (SLR) and used **web scraping** and other pre-processing steps to get relevant papers information in CSV.
  - **Collected** speed bump data using two smartphones placed on a motorcycle.
  - Performed data pre-processing to match both the time-series data and performed data analysis on them.
- Worked on **Self-Supervised** framework for cluster assessment in complex image datasets.

Text Analytics Internship - Indian Institute of Management (IIM) Ranchi | Jan 2023 - May 2023
Mentor: Prof. Sasadhar Bera, Department of Operations Management, IIM Ranchi, India — *Github*

- The main aim of this internship is to perform **Topic Modelling** & other **Data Analysis** methods on web-scrapped data related to different web series genres stored in Excel files.
- Performed various data cleaning and pre-processing steps. Generated **WordCloud** for different N-grams.
- Created a **Knowledge Graph** from correlation coefficient on bigrams using **NetworkX** library.
- Applied various Topic Modelling techniques like **Latent Dirichlet Allocation (LDA)** using Gensim and Visualize the topics using **pyLDAvis**. And later applied **Contextualized Topic Modeling (CTM)**.
- Then performed Sentiment Analysis using NLTK library and then performed classification.
- Performed **Multilabel regression** using **XGB Regressor**, **Feed Forward NN**, **RNN**, **LSTM**, and **Bi-LSTM** on another given regression dataset and evaluated using **Root Mean Squared Error (RMSE)**. XGB Regressor outperformed other models.
- Performed **missing value imputation** using KNN and MissForest on different dataset.

Projects

Representative Forgery Mining for Fake Face Detection | *Github* Oct 2022 – Nov 2022

- Tried for novelty in **Deepfake** detection using advanced augmentation technique following this paper. The datasets used are **DFFD, Celeb-DF, & FaceForencics++**. After extracting key frames from these video datasets, the dataset has a **size of 370 GB of images**. Implemented the paper & reproduced some important base results.
- Implemented and measured the performance of the presented technique on other latest advanced datasets, **Faceshifter & Neural Texture**, based on a two-stage face swapping method and found that it has performed better than **Xception** (base model) on these new datasets as well.
- Implemented a method where Instead of using random values to compose erasing blocks while occluding the image, used **Gaussian value** obtained after **Gaussian smoothing** of each image. And it achieved an **accuracy of 99.20%**, whereas the proposed technique in the paper has an accuracy of 99.59%.

Automatic sarcasm detection over social media | *Github* Sept 2022 – Nov 2022

- The **dataset** used is Self-Annotated Reddit Corpus **(SARC)** having **10 lacs+ rows** & **two classes** : 'Sarcastic' & 'Non-Sarcastic'. Applied various pre-processing techniques like **Tokenization, Stopwords removal, Lemmatization**, etc.
- Performed various Feature Extraction techniques like **Bag of Words, Tf-IDF, Word embeddings like SPACY** (pre-trained), and our **custom-trained embeddings**.
- Applied χ^2 Feature Selection and various Classification Models like Logistic Regression, Decision Tree, and Random Forest, with Hyper-Parameter tuning.
- **Logistic Regression** was the **best** model with BoW, with an **Accuracy** of **64.81%** respectively.
- Later applied Deep Learning based architectures like **FeedForward Neural Network, Stacked RNN, Stacked LSTM & Transformers** custom trained with appropriate pre-processing.
- **Transformer** achieves the highest **Accuracy** of **66.48%**.

Mobile Price Prediction | *Github* Feb 2022 – April 2022

- This is a **multi-class classification** problem. Performed **Exploratory Data Analysis** and various pre-processing techniques on the dataset having various mobile features data and its price.
- Applied various Feature Engineering and Feature Selection techniques like **Mutual Information & ANOVA**.
- Applied Classification Models like **Logistic Regression, KNN, Random Forest, Support Vector Machine, & Ada Boost**. And performed model tuning, i.e., Hyper-Parameter tuning, using **GridsearchCV**.
- Logistic Regression was the best model, with an **f1-macro-avg.** score of **98%** with only **8 features** out of 20.

COVID-19 Vaccines Twitter Sentiment in India: Hotspot Mapping | *Github* Sept 2021 - Nov 2021

- **Extracted tweets** from India related to different covid19 Vaccines.
- Applied various NLP methods such as Tokenization, Part-of-Speech (POS) tagging, Stemming, etc., using **NLTK**.
- Did **Sentiment Analysis** using **TextBlob** to get the sentiments of the Indian people about Covid19 Vaccines.
- Created a **hotspot map** of India using **Folium** based on the respective sentiments of different locations in India.

Technical Skills

Languages: Python (proficient), Matlab, Java
Platforms: Microsoft Windows
Databases: MySQL
Cloud Technologies: AWS (including SageMaker)
Other Tools & Platforms: Jupyter Notebook, VS Code, Git, Overleaf, Tableau, MS Office
Libraries/Frameworks: Pandas, NumPy, Matplotlib, Scikit-Learn, OpenCV, Scikit-Image, YOLO v7&v8, PyTorch, Tensorflow, Keras, NLTK, Gensim, Hugging Face, BeautifulSoup, Selenium, FastAPI, Langchain

Publications

- Mazumder, A., Baruah, T., **Singh, A. K.**, Murthy, P. K., Pattanaik, V., Rathore, P. (2023). **DeepVAT: A Self-Supervised Technique for Cluster Assessment in Image Datasets** (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2306.00011, **IEEE ICCVW 2023**

Note: The above information provided by me is true, and I have all the relevant documents to authenticate the same.