# Sourav Karmakar

Email: sourakarmakar@gmail.com
Github: github.com/Sourav89068
Mobile: +91-890-6855-327
Linkedin: /in/sourav-karmakar-693b1a1bb/

## Skills Summary

- **Languages**: Python, SQL, Bash, HTML, CSS, JavaScript, R, C
- **Frameworks**: Pytorch, Pytorch Geometric, Tensorflow, Pandas, Numpy, Hadoop, Pyspark, Scikit-learn, Matplotlib, Seaborn, Django, Flask, Selenium, BeautifulSoup, Scipy
- **Tools**: Docker, GIT, PostgreSQL, MongoDB
- **Platforms**: Linux
- **Soft Skills**: Prompt Engineering

## Experience

- **Sravathi AI Technology Pvt Ltd** — On-site
  *Data Scientist (Full-time)* — *Aug 2022 - Present*
  - **Classification with Abstention**:
    1. The classification approach leverages a Graph Convolutional Network (GCN) and a Siamese network with triplet loss to enhance GCN embeddings, facilitating improved understanding of data distribution and confident predictions; additionally, an XGBoost Classifier is utilized for multi-class classification and generating confidence scores for predicting classes beyond the supervised classes.
    2. Automated evaluation metrics are generated to compare different model versions and track improvements.
    3. The Pipeline has been deployed using a Docker container, with django as backend providing a consistent and portable environment.

    **Tech:** Python, Pytorch, Pytorch Geometric, Pandas, Numpy, Rdkit, Scikit-learn, Django, Docker.
  - **One-Step Retro-synthesis with Graph Attention Network**:
    1. A Graph Attention Network is utilized to improve the accuracy of retro-synthetic routes for a molecule.
    2. The data is pre-processed using a scoring model to assess the relevance and feasibility of potential routes.
    3. Django is employed as the backend framework for deploying the system, enabling user interaction and model reusability.
    4. User feedback on each prediction of the model is collected using mongoDB.
    5. The system is encapsulated in a container using Docker, ensuring portability and reproducibility.

    **Tech:** Python, PyTorch, Pandas, Numpy, Django, MongoDB, Docker.
  - **Multi-Step Retro-synthesis**:
    1. A multi-step retro-synthesis model was constructed using Neural Guided A* Search, expanding on a one-step model for predicting multi-step chemical synthesis routes.
    2. The one-step model incorporates graph-based neural networks to guide the search and generates leafs at each step.
    3. The Neural Guided A* Search algorithm is employed to extend predictions to multiple steps by exploring the reaction tree and considering various reaction pathways.
    4. The model assesses promising nodes, produces ranked candidate reactions, and selects the most favorable reaction at each step.
    5. The developed code automates the process, allowing users to input a target molecule and receive the output as a PDF report.
    6. This approach simplifies synthesis planning, enabling researchers to efficiently explore intricate chemical reactions and design multi-step synthesis routes for desired target molecules.
    7. Django is utilized as the backend framework, and the model is encapsulated in a Docker container for operating system level isolation.

    **Tech:** Python, Pytorch, Pandas, Numpy, Rdkit, ReportLab, MongoDB, Django, Docker.

- **Crediwatch Information Analytics Pvt Ltd** — Remote
  *Machine Learning Intern (Full-time)* — *Feb 2022 - Jun 2022*
  - **Streamlining In-House Modelling with custom configuration**:
    1. Implemented a common modeling framework to streamline in-house modeling and reduce model building time for internal stakeholders.
    2. Incorporated data drift modules to monitor changes in tabular data, ensuring model accuracy and reliability.
    3. Introduced a centralized logging module to consolidate logs from various modules within the framework.
    4. Developed a web-based YAML editor using Flask, allowing easy configuration of YAML files for the customized inputs in the generalized framework.

    **Tech:** Python, Scipy, Flask, HTML, Javascript.
  - **Document Image Denoising Model**:
    1. Created an Encoder-Decoder model using the U-Net architecture to address document image denoising.
    2. The model aimed to improve the quality of document images captured by laptop or phone cameras.
    3. By reducing noise in the images, the resulting model achieved clearer and visually enhanced document representations.

## Education

- **Ramakrishna Mission Vivekananda Educational and Research Institute** — Howrah, India
  *Master of Science - Big Data Analytics; GPA: 7.55* — *2020 - 2022*
- **Banwarilal Bhalotia College** — Asansol, India
  *Bachelor of Science - Mathematics; GPA: 7.85* — *2017 - 2020*
- **Asansol Old Station High School** — Asansol, India
  *Higher Secondary - Science; Percentage: 81.8* — *2015-2016*
- **Asansol Old Station High School** — Asansol, India
  *Secondary; Percentage: 85.4* — *2013-2014*

## Projects

- **Sustainability AI Engine (under Debabrota Basu,Faculty at (ISFP),Équipe Scool, Inria Lille- Nord Europe Bât A, France)**:

  1. Gathered data from various consultancy websites, relevant websites, and Twitter to construct datasets for a sustainability AI engine.
  2. Utilized available pretrained language models to obtain embeddings and compared their performance.
  3. Conducted sentiment analysis on social media data to capture the sentiments of end-users.
  4. Implemented the page-rank algorithm to rank keywords and perform text summarization based on the similarity matrix, improving readability and conciseness.
  5. Utilized extractive summarization techniques to generate concise summaries while preserving the essence of the original content.
  6. Devised and implemented alignment checking algorithms.
  7. Employed prompt engineering by leveraging the API from OpenAI to create prompts for summarization, sentiment extraction, and actionable sentence extraction.

  **Tech:** Python, Linux, Pytorch, GCP, Flask, Docker.

- **Voice Cloning, Text-to-Speech (under Purnendu Mukherjee, Founder and CEO of convai)**:

  1. The research emphasizes the significance of real-time voice cloning with limited data.
  2. The process involves recording the voice of a new speaker, cleaning the recorded data, and training the spectrogram generator and vocoder models.
  3. Due to the shared speaking characteristics among different speakers in a language, it is desirable to leverage the knowledge of a pre-trained text-to-speech model when synthesizing the voice of a new speaker.
  4. A text-to-speech model was fine-tuned to clone the voices of gaming characters based on users' requirements.
  5. The deployed text-to-speech model is accessible through Flask in the cloud.

  **Tech:** Python, Linux, Pytorch, GCP, Flask, Docker.

- **Modelling and Prediction of Oil Prices using News (under Gopal Krishna Basak, Professor at Indian Statistical Institute)**:

  1. The data preparation involved scraping data from a news website that provided WTI and Brent oil prices.
  2. News features were extracted using FinBert embeddings.
  3. A comparison was made between the model using news features and a baseline model that did not incorporate news data.

  **Tech:** Python,Selenium, Beautifulsoup, Tensorflow

- **Forecasting of NO2 monthly emission data of 29 Indian States and Union Territories(under Dr. Sudipta Das, Assistant Professor at RKMVERI)**:

  1. Employed the SARIMA model to forecast NO2 emission data, utilizing the same set of parameters for all states.
  2. Examined the co-integration between NO2 data and yearly GDP data for each state.

  **Tech:** R

- **Facial Key-points Detection (under Sujoy Kumar Biswas, Director and Principal Scientist AIMP LABS, Kolkata)**:

  1. Conducted non-linear regression using a Convolutional Neural Network (CNN) model.
  2. Optimized the model's performance by fine-tuning hyperparameters.
  3. Deployed the model locally using Flask for hosting and serving the predictions.

  **Tech:** Python, Pytorch, Opencv