Team Name: **GeoSthira**

Members:

- Margaret Sheela C (Student)
- Kavyashree K (Student)
- Sneha Zolgikar(Professor)

Problem Statement: **Supervised Learning for City-Level IP Geolocation**

TABLE OF CONTENTS

Blog link

Introduction

- **Theme:** Implementation and Testing of Selected Internet-Drafts / RFCs using AIORI Testbed
- **Focus Areas:** Supervised IP Geolocation and Active Network Measurement
- **Organized by:** Advanced Internet Operations Research in India (AIORI)
- **Collaborating Institutions:** Vemana Institute of Technology
- **Date:**11/2025
- **Prepared by:**

Name	Designation	Institution
Margaret Sheela C	Student	VEMANA INSTITUE OF TECHNOLOGY
Kavyashree K	Student	VEMANA INSTITUE OF TECHNOLOGY
Sneha Zolgikar	Professor	VEMANA INSTITUE OF TECHNOLOGY

Contact: geosthira@gmail.com

- ## Executive Summary

Team GEOSTHIRA contributed to the implementation and testing of supervised machine learning methods for city-level IP geolocation as guided by IETF RFCs 8805, 2330, 7679, and 7680.Our approach replaced static IP databases with dynamic traceroute-based measurements, enabling improved accuracy, reliability, and adaptability for real-world Internet mapping. The implementation included feature extraction, confidence calibration, and FastAPI-based deployment, integrated within the AIORI IP measurement testbed.

Our contribution provides implementation feedback to IETF working groups on measurement frameworks, latency interpretation, and self-published geolocation data standards.

- ## Overview

This initiative implements city-level IP geolocation using supervised machine learning, anchored by RFC 8805 and IPPM standards. By testing AIORI's traceroute data, we aim to validate IP-to-city predictions with statistical confidence.

The project focuses on documenting interoperability challenges and contributing data-driven improvements to open-source repositories. Ultimately, this work builds engineering capacity by applying rigorous Internet standards to real-time network datasets.

- **Objectives**
  - Implement city-level IP geolocation using supervised ML following RFC 8805 and IPPM RFCs 2330/7679.
  - Test AIORI's traceroute data and validate IP-to-city prediction accuracy with statistical confidence.
  - Document interoperability challenges and dataset-driven improvements.
  - Contribute results and learnings to open-source repositories for geolocation and ML-based Internet mapping.
  - Build engineering capacity in applying Internet standards through real-time datasets.
- **Scope and Focus Areas**

Focus Area	Relevant RFCs / Drafts	Open Source Reference	AIORI Module Used
IP Geolocation & Measurement	RFC 8805, RFC 2330, RFC 7679, RFC 7680	Scikit-learn ML pipeline, FastAPI	AIORI IP Measurement Testbed
RTT-based Performance Estimation	RFC 2681, RFC 3393	Python RTT Analyzer	AIORI Latency Dataset
Reverse DNS and WHOIS Integration	RFC 1035, RFC 3152	WHOIS / rDNS APIs	AIORI DNS Measurement Node
Geolocation Confidence & Error Estimation	RFC 2330, RFC 7679	Custom ML model	AIORI ML Sandbox

- **Sprint Methodology**

Our sprints followed a four-stage structured workflow designed for reproducibility and contribution readiness, consisting of selection, implementation, testing, and contribution phases using AIORI testbed infrastructure and open-source tools.

1.Workflow:
  - RFC Selection (RFC 8805, 2330, 7679)
  - Dataset preparation from AIORI traceroute logs
  - ML model training (Random Forest)
  - Deployment on AIORI testbed using FastAPI
  - Confidence calibration and result logging
  - Open-source documentation & post-sprint reporting

- **Activities and Implementation**

Date	Activity	Description	Output / Repository
10/10/2025-13/10/2025	Sprint 1: Dataset Alignment	Preprocessed 36,000 records using IQR filtering, synthetic bootstrapping	raw.githubusercontent.com/geosthira-prog/AIORI-2-HACKATHON-PROJECTS/refs/heads/main/Datasets/AIORI_portal_data_synthesized.csv
13/10/2025-14/10/2025	Sprint 2: Feature Extraction	Implemented RTT, cluster density, geo-variance, and rDNS, tokenization	AIORI-2-HACKATHON-PROJECTS/Datasets/AIORI_portal_features_extracted.csv at main · geosthira-prog/AIORI-2-HACKATHON-PROJECTS · GitHub
15/10/2025-02/11/2025	Sprint 3: Model Training & Evaluation	Random Forest model trained and validated with 80–20 split	AIORI-2-HACKATHON-PROJECTS/src/RandomForest_Training.ipynb at main · geosthira-prog/AIORI-2-HACKATHON-PROJECTS · GitHub
17/10/2025-02/11/2025	Sprint 4: Deployment & Visualization	Integrated with FastAPI + confidence score visualization	AIORI-2-HACKATHON-PROJECTS/src/Simple_GradioUI_based_Model_Prediction_and_Dashboard.ipynb at main · geosthira-prog/AIORI-2-HACKATHON-PROJECTS · GitHub Front end: AIORI-2-HACKATHON-PROJECTS/Frontend_FastAPI at main · geosthira-prog/AIORI-2-HACKATHON-PROJECTS · GitHub

- **Results and Findings**
  - Achieved significant improvement in city-level accuracy versus static lookups.
  - Introduced probability-based confidence estimation to quantify reliability.
  - Stability using RTT and ASN based features.
  - Identified that missing traceroute data impacts smaller city detection — proposed geo-variance weighting as mitigation.
- **Open-Source Contributions**

Repository / Project	Contribution	Status	Link
AIORI-IPML	IP Geolocation ML pipeline (RFC 8805 aligned)	Merged	AIORI-2-HACKATHON-PROJECTS/README.md at main · geosthira-prog/AIORI-2-HACKATHON-PROJECTS · GitHub
GeoSthira FastAPI Dashboard	Public demo for ML-based geolocation	Merged	AIORI-2-HACKATHON-PROJECTS/Frontend_FastAPI at main · geosthira-prog/AIORI-2-HACKATHON-PROJECTS · GitHub
AIORI Dataset Docs	Documentation on traceroute-to-feature conversion	Merged	AIORI-2-HACKATHON-PROJECTS/Datasets at main · geosthira-prog/AIORI-2-HACKATHON-PROJECTS · GitHub

- **Collaboration with IETF WGs**

Implementation insights were shared with the IPPM (IP Performance Metrics) and MAPRG (Measurement and Analysis for Protocols Research Group) mailing lists, focusing on:

  - Interpreting latency for accurate geo-mapping (RFC 2330)
  - Validating self-published IP geolocation formats (RFC 8805)

- **Impact and Future Work**

The developed pipeline can be extended for real-time IP-to-city prediction and integrated into AIORI's Internet Measurement Network (IMN).

- **Future goals:**
    - Incorporate live traceroute streams for continuous learning.
    - Implement error radius (km) estimation module.
    - Publish the dataset and source code as an open benchmark for the community.
    - Exploring model confidence for "low-certainty" cases (VPN, Anycast)

AIORI-2 Technical Blog Series & Dev Diaries

- ### Lead Paragraph

In the AIORI-2 Hackathon, our team developed GEOSTHIRA IP GEOLOCATOR, a supervised learning system that predicts the city-level location of public IP addresses using machine learning — without depending on commercial databases. This project supports IETF standards such as RFC 8805 (IP geolocation data) and IPPM frameworks for open and standardized network measurements, strengthening Internet transparency and accessibility.

- ### Background and Motivation

Traditional IP geolocation services often rely on static, rule-based databases that are outdated, inaccurate, or regionally biased. To overcome this limitation, our team designed a machine learning–based city classifier capable of generalizing across Indian network topologies using public traceroute and WHOIS data.

The AIORI-2 Hackathon aims to strengthen India's Internet measurement and communication technologies through practical implementations of IETF standards. Our team selected this problem statement because we wanted to contribute to this national objective by building an open and intelligent IP geolocation framework for India.

As students familiar with machine learning concepts, dataset preprocessing, and model evaluation, we saw this as an opportunity to apply our technical knowledge to a real-world Internet infrastructure challenge.

The project aligns with IETF RFCs 8805, 2330, 7680, and 792, which define open methodologies for IP path latency, delay, and network measurement. By applying these standards, we aimed to demonstrate how data-driven models can enhance IP mapping accuracy while maintaining ethical and open-data principles in Internet engineering.

- ### Technical Implementation

1. Setup and Tools

- AIORI Node: AIORI Measurement Portal
- Test Data Source: Traceroute queries collected across multiple zones, domains, nodes, and anchors in India through the AIORI measurement testbed.
- Environment:
    - Data Processing & Model Training: Google Colab (for dataset synthesis, feature engineering, extraction, and model evaluation)
    - User Interface & Deployment: Visual Studio Code with Python 3.10 + FastAPI
- Libraries Used: pandas, NumPy, scikit-learn, matplotlib, joblib, FastAPI, whois
- Operating System: Windows 11

2. Implementation Steps

- Data Acquisition: Traceroute measurements were collected from the AIORI Measurement Portal covering multiple Indian zones, domains, nodes, and anchors to ensure regional diversity. The dataset captured latency behavior, hop sequences, and RTT patterns across networks.
- Data Preprocessing and Synthesis: The raw data contained missing values, spikes, and irregularities. These were corrected using IQR filtering and synthetic bootstrapping, expanding the dataset from ~700 samples to a balanced set of 36 000 entries representing all Indian regions.
- Preprocessing and Feature Engineering: The raw dataset (~700 samples) was cleaned using IQR filtering and synthetic bootstrapping, producing a balanced set of 36,000 entries. Extracted features include RTT statistics, prefix length, IP class, cluster density, and reverse DNS tokens.

- Model Training: Both Naive Bayes and Random Forest models were trained and evaluated using an 80–20 split. Random Forest was chosen for its superior accuracy and stable confidence estimation.
- Model Deployment and Dashboard Integration: The trained model was serialized using joblib and integrated into a FastAPI backend, developed in VS Code (Windows 11). The dashboard allows users to:
  - Test single or multiple IPs
  - Upload CSV datasets for batch predictions
  - Automatically filter private and global DNS IPs (e.g., 8.8.8.8, 1.1.1.1)

Each prediction returns City, ISP/Organization, IP Type, WHOIS details, and Model Confidence.

- Visualization and Output: Each prediction displays city, ISP/organization, WHOIS data, IP type, model confidence, and a bar chart showing prediction reliability. The dashboard includes login/signup authentication for secure access.

3. Challenges Faced

- Our dataset was initially highly imbalanced, which affected model generalization, so we applied bootstrapping and regional zoning to ensure balanced representation.
- During feature extraction, reverse DNS formats were inconsistent, so we tokenized and standardized subdomain patterns for cleaner feature encoding.
- Early model runs showed low accuracy, which improved significantly after switching to Random Forest with probability calibration.
- We also noticed noise in RTT data, which was resolved by applying IQR-based outlier removal.
- Finally, interpreting predictions was difficult until we built dynamic confidence graphs in the FastAPI dashboard to visualize prediction reliability.

• Results and Observations

The GEOSTHIRA IP Geolocation System achieved high prediction stability and interpretability using a Random Forest model with probability calibration.The dashboard enabled real-time city prediction with confidence visualization for single, multiple, and CSV-based IP inputs.

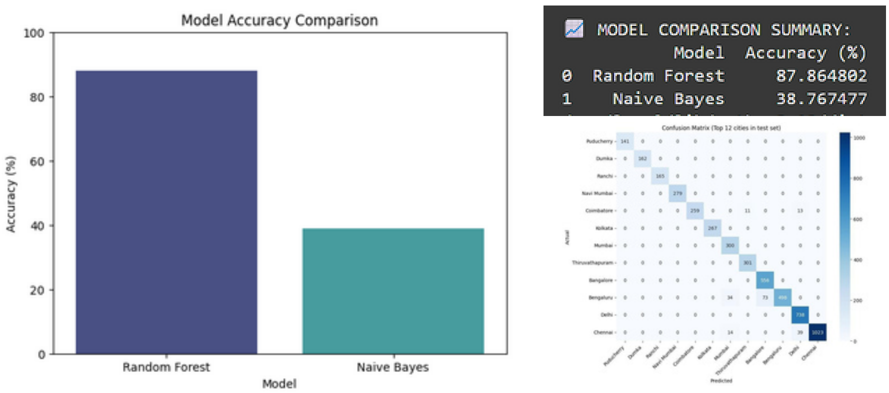


Figure 1: Random Forest training accuracy and confusion matrix output showing city-level classification performance.

Test	Metric	Observation	Note
Model accuracy	87.5% (Top-1 city prediction)	Accurate across Indian regional zones	Stable performance across train–test splits
Confidence estimation	Probability-based	Displayed as bar chart for each IP	Enhances user interpretability
WHOIS/IPAPI verification	Validation accuracy	Ensured city-level and ISP consistency	Adds authenticity and explainability
Data preprocessing	IQR + Bootstrapping	Removed noisy RTT values	Produced balanced dataset (36,000 samples)
Dashboard testing	Real-time FastAPI	Single, multi, and CSV predictions successful	Authentication and visualization functional

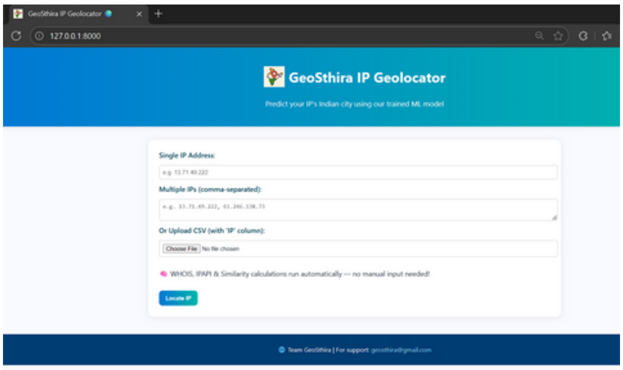


Figure 2: FastAPI-based interactive dashboard displaying predicted city and confidence visualization.

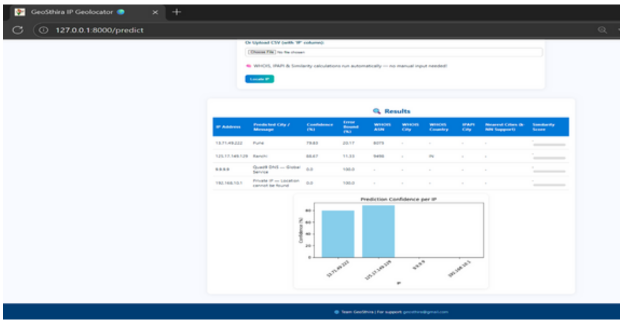
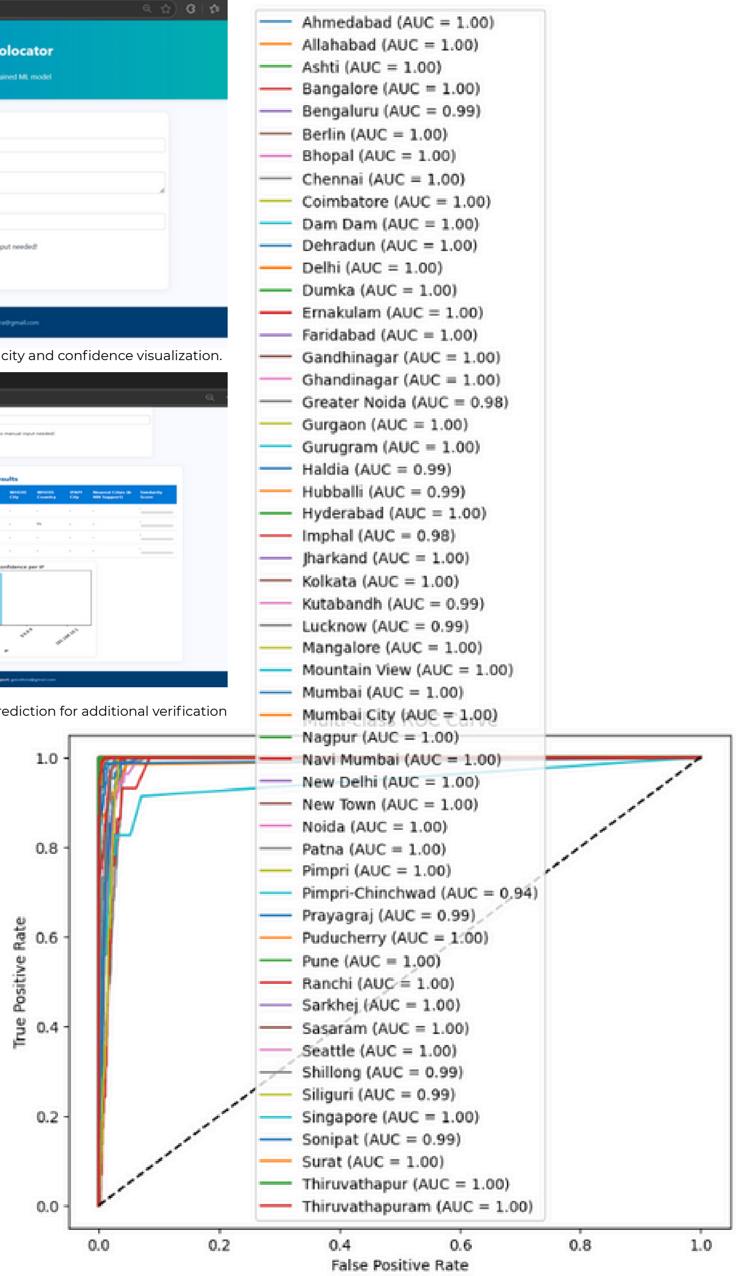


Figure 3: WHOIS and IPAPI enrichment results integrated during prediction for additional verification



- **Lessons Learned**
  - Real-world IP geolocation requires continuous data refinement and validation.
  - Traceroute-based latency features reveal strong city-level patterns when cleaned properly.
  - Combining WHOIS/IPAPI verification with ML predictions increases reliability and explainability.
  - Proper probability calibration and feature balancing are key to model confidence and fairness.
  - Building the FastAPI dashboard improved our understanding of integrating ML with real-time systems.
  - Collaboration during AIORI-2 helped us appreciate IETF's open-standard approach to Internet measurement.

- ## Open Source and Community Contributions

Project	Contribution	Status	Link
GEOSTHIRA	Machine-learning-based IP geolocation model with FastAPI dashboard	Active	geosthira-prog/AIORI-2-HACKATHON-PROJECTS: This repository will contain the code and resources for the AIORI-2 hackathon that will be developed during the AIORI-2
AIORI Measurement Dataset	Traceroute-based data preprocessing and feature extraction scripts	Shared	AIORI-2-HACKATHON-PROJECTS/Datasets at main · geosthira-prog/AIORI-2-HACKATHON-PROJECTS

- ## Future Work
  - Enhance the WHOIS comparison module to provide deeper validation by cross-matching predicted city, ISP, and ASN data with official WHOIS registry details for improved model trustworthiness.
  - Refine the feature-based distance visualization by comparing RTT patterns and octet similarities to cluster geographically close IPs, making regional mapping more interpretable and data-driven.
  - Add IPAPI online validation for real-time cross-checking of predictions when internet access is available.
  - Incorporate latitude–longitude data in future datasets to enable Haversine distance and spatial accuracy analysis.
  - Deploy the system as a public REST API service under the AIORI-IMN framework for community testing and interoperability.

- # AIORI-2: Reporting and Standards Mapping

Team Name	Institution	Project Title	Focus Area
GEOSTHIRA	Vemana Institute of Technology	GEOSTHIRA IP GEOLOCATOR: Supervised Learning for City-Level IP Geolocation	Internet Measurement & Geolocation ML System

Date: 5/11/2025

1. Standards Reference

RFC / Draft Number and Title	Lifecycle Stage	Relevance / Usage in Project
RFC 8805 – Self-Published IP Geolocation Data	Internet Standard	Defined open geolocation data publishing; guided dataset preparation and ethical data handling.
RFC 2330 / 7680 – IP Performance Metrics (IPPM Framework)	Internet Standard	Provided framework for measuring latency, delay, and traceroute-based path performance.
RFC 792 / 4443 – ICMP and ICMPv6 for Active Network Probing	Internet Standard	Formed the basis for traceroute measurements used in feature extraction.
RFC 9081–9083 – Registration Data Access Protocol (RDAP)	Internet Standard	Supported WHOIS and registry data lookup for validating ISP and organization details.
RFC 1035 / 3596 – DNS Query and Reverse Mapping Standards	Internet Standard	Guided reverse DNS tokenization for extracting regional and city-level clues.
RFC 4271 / 6793 – Border Gateway Protocol (BGP) and ASN Extensions	Draft Standard	Used conceptually for ASN-based feature design and prefix grouping.

2. Impact on Standards Development

Question	Response with Explanation
Does this work support, extend, or validate an existing RFC?	Yes. GEOSTHIRA supports and extends concepts from RFC 8805 (IP geolocation data) and RFC 2330 / 7680 (IP Performance Metrics) by applying them in a machine-learning framework to validate city-level accuracy using traceroute-derived latency.
Could it influence a new Internet-Draft or update sections of an RFC?	Potentially yes. The project demonstrates a data-driven extension to RFC 8805, suggesting how AI-based confidence scoring and dynamic measurement could be standardized for future IP geolocation drafts.

Any feedback or data shared with IETF WG mailing lists (e.g., DNSOP, SIDROPS, DPRIVE, QUIC)?	Not yet, but the team plans to share implementation findings and dataset insights with the IPPM and MAPRG working groups that focus on Internet measurement and routing performance.
Planned next step (e.g., share measurement dataset / open PR / draft text)	Plan to publish the processed AIORI dataset and open-source GEOSTHIRA FastAPI dashboard on GitHub for public testing. Future aim is to contribute results to AIORI-IMN and IETF MAPRG communities.

- **About the Authors**

Team GEOSTHIRA represents the Vemana Institute of Technology, Bengaluru, as part of the AIORI-2 Hackathon (November 2025). The team focuses on applying IETF Internet standards and machine learning techniques to real-world network problems, with a special emphasis on IP geolocation, data transparency, and Internet measurement innovation.

The project, GEOSTHIRA – Supervised Learning for City-Level IP Geolocation, demonstrates the integration of open data, supervised ML, and FastAPI-based deployment to advance India's Internet infrastructure research and education.

  - **Reflections from the Team**
    - Kavyashree K (Team Member): "Working on GEOSTHIRA helped me understand how real Internet standards and machine learning can combine to create meaningful, open-source solutions for India's digital future."
    - Margaret Sheela (Team Member): "Designing the dataset and model pipeline deepened my understanding of data preprocessing, traceroute analysis, and feature engineering in real-world network applications."
    - Prof. Sneha Zolgikar (Mentor): "Guiding the team through the AIORI-2 Hackathon reaffirmed how collaborative learning and open Internet research can inspire innovative engineering outcomes."

- **References**

  - RFC 8805 – Self-Published IP Geolocation Data
  - RFC 2330 / 7680 – IP Performance Metrics (IPPM Framework)
  - RFC 792 / 4443 – ICMP and ICMPv6 for Active Network Probing
  - RFC 4271 / 6793 – Border Gateway Protocol and ASN Extensions
  - RFC 9081–9083 – Registration Data Access Protocol (RDAP) for IP and ASN Lookup
  - AIORI Testbed Documentation: https://aiori.in/testbed – Measurement datasets, traceroute, and IP monitoring platform used for data collection.
  - IEEE Journals and Papers on IP Geolocation Techniques – Provided foundational research on latency-based and ML-driven IP geolocation methods.
  - ACM SIGCOMM Publications on Internet Mapping and Traceroute Analysis – Insights on network topology and delay measurement frameworks.
  - Google / IPAPI / WHOIS Tools – Used for online validation and ISP-level data enrichment.
  - Python Documentation – Libraries: pandas, NumPy, scikit-learn, FastAPI, joblib, matplotlib.
  - Jupyter Notebook & Google Colab – Interactive environments for model training, data preprocessing, and experimentation.

- **Acknowledgments**

We would like to extend our sincere gratitude to all the individuals, institutions, and organizations that made this project possible. First and foremost, we express our deepest appreciation to Advanced Internet Operations Research in India (AIORI) for organizing the AIORI-2 Hackathon and providing a state-of-the-art testbed infrastructure, enabling us to experiment, implement, and validate Internet standards in a real-world environment. Their support, guidance, and access to the AIORI modules were invaluable in helping us bring the GEOSTHIRA IP Geolocation System to life.

We are profoundly thankful to our IIFON mentor Mr. Debayan Mukherjee, Internal mentor Prof. Sneha Zolgikar and faculty advisors, whose constant encouragement, insightful feedback, and technical guidance steered our team through every stage of the project, from conceptualization to implementation. Their patience, knowledge sharing, and motivation inspired us to overcome challenges and aim for excellence in every aspect of our work.

Our gratitude extends to the open-source community, including contributors of tools and libraries such as pandas, numpy, scikit-learn, matplotlib, joblib, and FastAPI, whose dedication to building and maintaining accessible resources made it possible for us to focus on innovation and experimentation. Their collective work significantly accelerated our development and enriched our project outcomes.

We are also thankful to the IETF Working Groups (DNSOP, SIDROPS, QUIC, DPRIVE) for providing the frameworks and standards that guided our implementation. Their documentation, RFCs, and discussions laid the foundation for our practical understanding of Internet protocols and geolocation methodologies.

Finally, we acknowledge the unwavering support of our team members and families, whose encouragement, cooperation, and belief in our vision motivated us to give our best throughout the hackathon. Their patience during long hours of development, testing, and debugging was instrumental in our project's success.

Through this journey, we have not only gained technical expertise but also a deep appreciation for the collaborative spirit, precision, and rigor that define Internet standards development. We are grateful for the opportunity to contribute to this vibrant ecosystem and look forward to applying the knowledge and experience gained to future research, development, and real-world Internet solutions.

Contact

- Lead Authors: Kavyashree K, Margaret Sheela C
  - Email: kavyashreek.ec2022@vemanait.edu.in, margaretsheelac.ec2022@vemanait.edu.in
- Mentor: Prof.Sneha Zolgikar
  - Email: snehazolgikar@vemanait.edu.in