VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"Jnana Sangama", Belagavi-590018.

A Project Report on

"Privacy Preservation of Cluster Integrity on Web-Scraped Hospital Data

Extended Version of the Paper Presented at ICAECT 2025"

Submitted in the partial fulfillment of the requirements for the award of the degree of

Bachelor of Engineering in Computer Science and Engineering

Submitted by

Hrishik Sai Bojnal 1RF21CS052
Dharmisht SVK 1RF21CS035
J Krishna Kaarthik 1RF21CS055
Dhyaan Kotian 1RF21CS039

Under the Guidance of
Dr. Shashidhar V,
Assistant Professor, CSE, RVITM,

Department of Computer Science and Engineering
RV INSTITUTE OF TECHNOLOGY AND MANAGEMENT
(Affiliated to Visvesvaraya Technological University, Belagavi & Approved by AICTE, New Delhi)
JP Nagar 8th Phase, Kothanur, Bengaluru-560076

2024-2025

RV INSTITUTE OF TECHNOLOGY AND MANAGEMENT

(Affiliated to Visvesvaraya Technological University, Belagavi & Approved by AICTE, New Delhi)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the project work titled **'**Privacy Preservation of Cluster Integrity on Web-Scraped Hospital Data**'** is carried out by **Hrishik Sai Bojnal (1RF21CS052), Dharmisht SVK (1RF21CS035), J Krishna Kaarthik (1RF21CS055), and Dhyaan Kotian (1RF21CS039),** who are bonafide students of RV Institute of Technology and Management, Bengaluru, in partial fulfillment for the award of the degree of **Bachelor of Engineering** in **Computer Science and Engineering** of the Visvesvaraya Technological University, Belgaum, during the year 2025. It is certified that all corrections/suggestions indicated for the internal assessment have been incorporated in the report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements for the project work prescribed by the institution for the said degree.

Signature of Guide:	**Signature of Head of the Department:**	**Signature of Principal:**
Dr. Shashidhar V Assistant Professor, Department of CSE, RVITM, Bengaluru-76	**Dr. Malini M Patil** Professor &Head, Department of CSE, RVITM, Bengaluru-76	**Dr. Nagashettappa Biradar** Principal, RVITM,Bengaluru-76

External Viva

Name of Examiners	**Signature with date**
1	
2	

RV INSTITUTE OF TECHNOLOGY AND MANAGEMENT

(Affiliated to Visvesvaraya Technological University, Belagavi & Approved by AICTE, New Delhi)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

We, **Hrishik Sai Bojnal (1RF21CS052), Dharmisht SVK (1RF21CS035), J Krishna Kaarthik (1RF21CS055), and Dhyaan Kotian (1RF21CS039)**, the students of the seventh semester B.E, hereby declare that the project titled **"Privacy Preservation of Cluster Integrity on Web-Scraped Hospital Data"** has been carried out by us and submitted in partial fulfillment for the award of the degree of Bachelor of Engineering in **Computer Science and Engineering.** We declare that this work has not been carried out by any other students for the award of a degree in any other branch.

Place: Bengaluru **Signature**

Date:

1. Hrishik Sai Bojnal (1RF21CS052)

2. Dharmisht SVK (1RF21CS035)

3. J Krishna Kaarthik (1RF21CS055)

4. Dhyaan Kotian (1RF21CS039)

ACKNOWLEDGEMENT

The successful presentation of the **Privacy Preservation of Cluster Integrity on Web-Scraped Hospital Data** would be incomplete without expressing our sincere gratitude to all those who made this journey meaningful and rewarding.

We would like to extend our heartfelt thanks to **RV Institute of Technology and Management**, Bengaluru, and **Dr. Nagashettappa Biradar,** Principal, RV Institute of Technology and Management, Bengaluru, for his leadership and the open, encouraging environment that RVITM provides. We are truly grateful for the space and freedom the institution gives students like us to pursue our educational goals without hindrance.

Our deepest gratitude goes to **Dr. Malini M Patil**, Professor and Head, Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru, for her initiative and encouragement.

We gratefully thank our Project Guide, **Dr. Shashidhar V,** Assistant Professor, Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru. His high standards of academic excellence and his insistence on precision and clarity pushed us to deliver our best work. His deeply knowledgeable reviews, patience, and insightful suggestions guided us throughout the project. We would like to thank all the **Teaching** Staff and **Non-Teaching** Staff of the college for their cooperation.

Finally, we extend our heartfelt gratitude to our **families** for their encouragement and support, without which we would not have come so far. Moreover, we thank all our **friends** for their invaluable support and cooperation.

1. Hrishik Sai Bojnal (1RF21CS052)
2. Dharmisht SVK (1RF21CS035)
3. J Krishna Kaarthik (1RF21CS055)
4. Dhyaan Kotian (1RF21CS039)

Abstract

As India undergoes a digital revolution, a large amount of public health data is becoming available. This data can help researchers and policymakers understand how people across the country access healthcare and design better, fairer health programs. But using this information safely is not simple. Health data often includes location details that can expose personal information. A common way to protect privacy is to add random "noise," or small errors, to the data so individual locations cannot be identified. However, this method often degrades the data beyond usability. It removes important regional patterns that analysts need to study, which makes the dataset far less useful.

This work presents a new cluster-preserving anonymization framework that keeps natural patterns in the data while protecting privacy. Instead of randomly changing the positions of hospitals, our method first finds natural geographic clusters of hospitals that serve nearby areas. It then hides the exact locations by slightly moving each hospital only within the boundaries of its own cluster. This way, the data stays private but still preserves the overall macro structure of the healthcare system. Our experiments show that this approach keeps the data's original shape 3.25 times better than standard methods, based on the Earth Mover's Distance.

To promote further study and application, we have released the fully anonymized and enhanced dataset on Kaggle. Its impact on the data science community in India is already visible: as of January 2025, within two months of release, the dataset has been downloaded over 100 times and is currently trending under the "India" tag.

Retroactive Note: This report was carried out as a group project for the *Final Project Work* (10 credits) and is evaluated at the same level as a bachelor's thesis. It also serves as an extension of our findings published at the *2025 Fifth International Conference on Advances in Electrical, Computing, Communications, and Sustainable Technologies (ICAECT 2025)*.

Table of Contents

List of Figures

Chapter 1

INTRODUCTION

India stands at a fascinating, chaotic, and promising crossroads. The nation's rapid digitalization, a surge of connectivity reaching into the smallest towns and villages, presents a monumental opportunity. For the first time, we have the tools to tackle long-standing, deeply entrenched societal problems at a scale previously unimaginable. Public health is, perhaps, the most important frontier. The ability to map, measure, and model healthcare infrastructure across a subcontinent could revolutionize policy, save lives, and build a healthier future. This requires a lot of data. The digital breadcrumbs left across online platforms, from hospital websites to mapping services, form a vast, untapped reservoir of information about the state of our nation's health. The question is no longer *if* we can collect this data, but *how* we should.

1.1 The Central Conflict: Utility vs. Privacy

This research focuses on the conflict between data utility and data privacy. To be useful for understanding healthcare accessibility, data needs to be granular. We need to know where hospitals are, what services they offer, and how they are perceived by the communities they serve. This geospatial data, however, is deeply sensitive. The precise location of a healthcare facility is a piece of information that, when aggregated, can reveal patterns about communities, supply chains, and even patient demographics. Sharing this data in its raw form is not an option. It opens the door to misuse, raises security concerns, and erodes the very foundation of public trust. For a nation of 1.4 billion people, building and maintaining trust in our burgeoning data-driven institutions is paramount. If citizens fear that their data will be used against them, the entire promise of a data-driven society crumbles. We must find a way to balance the immense analytical power of data with the fundamental right to privacy.

1.2 When Naive Anonymization Fails

The simplest solutions are often the most tempting. One could, for instance, take each

hospital's coordinates and add small quantities of random noise, shifting each point by a few kilometres. This is often called naive anonymization, perhaps using a Gaussian distribution to add the noise. This approach might seem to solve the problem. Individual hospital locations are now incorrect, seemingly protecting them. But this is not the full picture. While this technique does obscure individual data points, it inadvertently destroys the most valuable information in the dataset: the *macro-level spatial patterns*. The natural clustering of hospitals in urban centres, the sparse distribution in rural areas, the very "shape" of the healthcare landscape, is warped and degraded. The data becomes useless for the very purpose it was collected for, which is regional healthcare planning.

1.3 Our Contribution: A Cluster-Preserving Framework

This project introduces a solution to this utility-versus-privacy dilemma. Our contribution is twofold:

A. An Anonymization Algorithm.

We propose and implement an anonymization technique that fundamentally respects the data's structure. Instead of scattering points randomly, our algorithm first identifies natural spatial clusters of healthcare influence. It then anonymizes hospital locations by moving them within the geometric confines of their own cluster. This preserves the integrity of the clusters and allows meaningful analysis of regional healthcare density, while still protecting the exact locations of each facility. Importantly, the algorithm's application extends beyond healthcare; it can be adapted to any geospatial context where an "impact" or "prominence" metric, like a star rating, is available. We publish the algorithm and associated findings in the 2025 Fifth International Conference on Advances in Electrical, Computing, Communications and Sustainable Technologies (ICAECT 2025) [1].

B. A Data Pipeline.

We meticulously document a full, end-to-end workflow for acquiring, cleaning, anonymizing, and enriching a nationwide hospital dataset. This process, detailed from initial PDF extraction to distributed web scraping, is designed to be transparent and replicable, forming a solid foundation for any subsequent analysis. The code for this is

publicly available on GitHub. [2]

1.4 Objectives

To translate these contributions into measurable goals, this project was guided by a set of specific objectives:

a. To propose a cluster-preserving anonymization methodology that is grounded in computational geometry, using concepts like convex hulls and Delaunay triangulation to define and enforce displacement boundaries.

b. To document and justify a full data-collection pipeline, including an automated table extraction of the PDF source, data cleaning, imputation and enrichment with a distributed web scraping effort.

c. To quantitatively evaluate the proposed anonymization method against a best-effort naive-noise baseline. We selected the Earth Mover's Distance (EMD) [3] as our primary metric to measure the preservation of the dataset's distribution.

d. To qualify the impact of the tool for sustainable growth, and progress toward a few UN Sustainable Development Goals [4], specifically SDG 3 (Good Health and Well-being), SDG 10 (Reduced Inequalities), SDG 9 (Industry, Innovation, and Infrastructure), and SDG 16 (Peace, Justice and Strong Institutions).

Chapter 2

LITERATURE SURVEY

In this section, we review the foundational literature that informs the proposed framework. The central challenge outlined in this work is not a single problem but a chain of them. Data must first be acquired ethically, aggregated securely, mined privately, and finally, anonymized in a manner that preserves its core utility.

2.1 Data Acquisition: Ethics, Technique, and Utility

The proposed methodology starts with gathering publicly available data. Although this step is technically straightforward, it raises the legal and ethical issues that this study addresses. Previous research notes a conflict between the usefulness of public data and the potential risks involved in collecting it.

The work of Krotov et al. [5] is a foundational text in the uncertain legal landscape of web scraping. They provide a tutorial that reviews the legal and ethical literature, identifying important questions that practitioners engaged in Web scraping need to address. Their discussion focuses on legal theories such as trespass to chattels and the ambiguity of the Computer Fraud and Abuse Act (CFAA). They emphasize that researchers need to manage the risk of ethical disputes and legal action in advance. This study is informed by their insights by proposing a strong anonymization framework downstream, which ethically justifies the need for upstream data collection.

This legal environment has become much more complicated due to recent large-scale policy changes by major data holders. The 2018 Cambridge Analytica scandal, which exposed the exploitation of Facebook user data, marked the end of the "Data Golden Age". In the resulting "APIcalypse," as termed by Bruns [6], many social platforms drastically reduced or completely cut off free, programmatic access to their data through Application Programming Interfaces (APIs). This shift has fundamentally changed how digital research is conducted.

This shift creates an important cause-and-effect relationship for the current study. The shutdown of official APIs has not, as one might assume, stopped large-scale data collection. Instead, as Brown et al. [7] and Trezza [8] note, it has harmed research by causing a "mass migration" back to the same web scraping methods that Krotov et al. identified as legally uncertain. This shift increases the ethical responsibility of researchers. When the straightforward option of API access is no longer available, researchers must take greater care to justify collecting public data through alternative methods. The strongest justification is being able to prove that individuals' privacy remains protected after the data is published.

In response to this more complex landscape, Brown et al. [7] introduce a modern and comprehensive framework for researchers, especially those in the U.S. Building on but moving past Krotov et al.'s [5] focus on legal aspects, their framework identifies four important areas to consider: legal, ethical, institutional, and scientific. They emphasize that researchers must navigate not only legal requirements but also Institutional Review Boards (IRBs), while ensuring that their data collection methods are both scientifically sound and reproducible.

These ethical concerns are not just theoretical. Research in specific fields shows their real-world importance. For example, Stringam et al. [9], in their study of online hospitality data, emphasize that privacy is a core value. They point out that "marketing intrusiveness and perceived surveillance elicit negative consumer responses," showing how essential privacy is within the broader discussion of digital ethics.

In contrast to the legal risks, the data science community often promotes web scraping for its educational and practical value. Dogucu and Çetinkaya-Rundel [10] recommend including web scraping in statistics and data science courses, arguing that it provides real, relevant data and lets students experience the full data science process. Their perspective is important because it presents scraping as a central and legitimate data collection skill, rather than a marginal practice, supporting realistic and meaningful data work. Our

framework follows this approach, positioning scraping as the valid starting point of a complete, reproducible, and ethically grounded data science workflow.

2.2 Technical Challenges in Large-Scale Geospatial Data Extraction

Apart from ethical and educational considerations, data acquisition also involves technical challenges related to efficiency. Uzun [11] addresses this by introducing UzunExt, a new scraping method designed to increase extraction speed. It relies on string-based techniques and uses extra information available on web pages, instead of depending on traditional and computationally heavy DOM-based extraction. This approach results in significant time savings. Such technical efficiency marks the first major challenge in building any large-scale data acquisition system, including the one developed for this project.

This efficiency issue becomes even more significant when dealing with geospatial data. Modern web mapping services like Google Maps do not deliver static HTML pages. Instead, they depend on dynamic content, asynchronous JavaScript (AJAX), and component-based frameworks to load map tiles, points of interest, and other information interactively.

This technical limitation creates a problem of practicality, not just performance. The DOM-based methods criticized by Uzun [11] are not only slow but often unusable for dynamic geospatial platforms. They cannot access data that loads after the initial page render, and they depend on stable DOM structures that are disrupted by the constantly changing, obscured class names produced by modern web frameworks. Tools like Selenium [12] offer a better solution by automating the entire browser to handle dynamic content.

2.3 Privacy-Preserving Artificial Intelligence (PPAI)

Before anonymization, the framework must detect natural geographic clusters and model influence, both of which are AI and machine learning tasks. The PPAI literature introduces an alternative privacy model that shifts focus from anonymizing and releasing data to keeping the data in place and running code directly on it. This section examines that approach to establish a contrast with the method proposed in this study.

The healthcare field offers a strong setting for examining this topic. Surveys by Khalid et al. [13] and Torkzadehmahani et al. [14] provide detailed reviews of privacy-preserving AI in healthcare and biomedicine, respectively. Both identify a main trend: the increasing use of Federated Learning (FL) and hybrid methods as the most promising approaches. Torkzadehmahani et al. note how FL brings computation to data, allowing models to be trained on sensitive data (like patient or genomic records) without centralizing it.

Several studies demonstrate this approach. Dodda et al. [15] provide a detailed review of federated learning (FL), describing how it enables collaborative model training. Qu et al. [16] suggest a hybrid edge computing setup that combines FL for pre-training with homomorphic encryption to securely split models. Aminifar et al. [17] propose k-PPD-ERT, a distributed algorithm that modifies randomized trees to safely analyse health data stored across different locations.

The dominant view in the PPAI literature is that data should remain in place and not be transferred. Although this provides strong privacy protection, it conflicts with the "data-for-good" objective of this project. Methods like Federated Learning generate trained models rather than datasets that can be openly shared. Advancing sustainable development in India requires wide and equitable access to data among researchers, urban planners, and NGOs, which a single trained model cannot provide. As a result, existing PPAI approaches do not meet this project's requirements. The proposed framework takes an alternative approach: instead of releasing raw data, which would be unethical, or only releasing a federated model, which would be inadequate, it produces a high-utility dataset that is provably anonymized. This brings the project back to the well-established, though complex, approach known as Privacy-Preserving Data Publishing (PPDP).

2.4 Foundational Models for Privacy-Preserving Data Publishing (PPDP)

The framework's rejection of the PPAI model means we need to go back to the PPDP approach. This area focuses on techniques for sharing datasets that remain useful for

analysis while keeping individual information private. The main idea behind PPDP is *k*-anonymity, first introduced by Samarati and Sweeney [18] in 2002.

Samarati and Sweeney's research identified the main issue of re-identifying individuals from so-called "quasi-identifying" attributes, such as ZIP code, date of birth, and gender. A dataset is considered *k*-anonymous if each record cannot be distinguished from at least *k*-1 other records in the dataset based on these attributes. This protection is provided through techniques like generalization, which replaces specific values with broader categories (for example, using an age range instead of an exact age), and suppression, which removes certain data entirely.

This general approach was soon adapted for the specific case of location data. Gedik and Liu [19] created a scalable framework to protect location privacy in Location-Based Services (LBSs). Their work introduced a personalized *k*-anonymity model, letting each user choose their own desired level of privacy, defined by their *k* value.

At the heart of Gedik and Liu's [19] framework is their "CliqueCloak" algorithm. This algorithm generates spatio-temporal cloaking boxes, which are three-dimensional regions defined by *(x, y, time)*, that group at least *k* users together before sending a query to a service provider. This process effectively hides an individual's exact location and time within a larger, generalized group.

Gedik and Liu's [19] work serves as a direct intellectual predecessor to the framework proposed in this study. Their spatio-temporal cloaking box functions as an early form of clustering for anonymization. By grouping users in space and time to meet a *k*-anonymity requirement, they set the precedent for using spatial aggregation to protect privacy. The framework presented here can be seen as a modern and more robust extension of this established line of research.

2.5 The Great Debate: Differential Privacy vs. Anonymization in Geospatial Contexts

Any modern PPDP framework must address the current "gold standard" in data privacy: Differential Privacy (DP). This section examines DP and explains why a high-utility anonymization model was chosen instead.

DP was created to address the weaknesses of *k*-anonymity, which can be broken by homogeneity attacks (where all *k* individuals in a group share the same sensitive attribute) and background knowledge attacks. DP offers a stronger, mathematically defined guarantee: the result of a query is nearly the same whether or not any single individual's data is included. This is usually achieved by adding carefully calibrated statistical noise, such as through the Laplace mechanism, to the query results or the dataset itself.

Applying DP to high-dimensional, detailed geospatial data, however, leads to a severe and often unmanageable trade-off between privacy and utility. The recent survey by Lorestani et al. [20], provides a detailed taxonomy of "geomasking" techniques, which are the main ways to protect privacy in geospatial datasets. These techniques include affine transformations, weighted random masks, and point rotations.

This creates a fundamental methodological conflict for the current study. The project aims to identify and preserve natural geographic clusters for urban planning and social analysis. Differential Privacy, when applied through geomasking, adds noise to points or otherwise obscures their true locations, explicitly aiming to hide or destroy fine-grained spatial patterns. Adding enough noise to meet a strong DP guarantee often makes the data unusable for the very cluster analysis it is meant to support.

For this reason, the proposed framework does not reject DP for privacy reasons, which are strong, but for utility reasons specific to this task. The "data-for-good" goal requires maintaining spatial accuracy. Consequently, the study returns to the anonymization paradigm [18] [19] and proposes a modern, robust method that uses natural spatial patterns as the foundation for anonymization rather than obscuring them.

2.6 Privacy-Preserving Clustering and Anonymization

Because PPAI is limited by utility issues and DP reduces the quality of spatial patterns, this framework falls under Clustering-Based Anonymization Mechanisms (CAMs). This section looks at the most relevant research connected to our new approach, which combines three areas: (1) grid-based density clustering, (2) perturbation with geometric constraints, and (3) spatial analysis that keeps data highly useful.

The core background for this framework comes from research on CAMs. A detailed review by Majeed et al. [21] organizes these mechanisms and explains why they work better than traditional anonymization methods. CAMs are described as a practical approach for responsible data science because they can keep data useful in ways that methods like DP cannot. The review also shows the importance of CAMs for privacy in location-based systems, which is the specific area this project focuses on.

Majeed et al. [21] point out an important gap in research: the need for anonymization methods that are practical, easy to verify, and efficient. Our framework addresses this need by using fast, vector-based grid calculations and reliable geometric checks with Delaunay triangulation [22] to offer a workable solution.

2.6.1 Influence based Grid Clustering

Before data can be anonymized, clusters need to be found. Most research on spatial clustering focuses on versions of DBSCAN, which are effective but can be slow and require careful adjustment of parameters.

A faster and more practical option is grid-based clustering. For instance, the GRIDBSCAN algorithm divides data into a grid, combines cells with similar densities, and then performs a final clustering step. This shows that using a grid to simplify space can be an efficient way to analyze density.

Our framework adopts this grid-based principle but implements it more directly. Instead of the complexity of GRIDBSCAN, our method first creates an *influence map*—a common technique in spatial modeling for representing phenomena that radiate outwards (like heat,

noise, or, in this case, hospital prominence). The use of percentile-thresholding on this influence grid, followed by a standard scipy.ndimage.label (8-adjacency) scan, is a highly efficient, non-iterative method for identifying "islands" of high influence. This approach effectively identifies arbitrarily-shaped clusters in a single pass, directly answering the call from Majeed et al. [21] for an *efficient* mechanism.

2.6.2 The "Gap": Perturbation vs. Geometric Boundary Analysis

The final and most original part of our framework is its method for anonymization: cluster-preserving spatial perturbation. Current research is split on how to handle this.

One approach is Geomasking, which is the standard way to anonymize individual points. As Lorestani et al. [20] review, this includes techniques like affine transformations, weighted random masks, or adding noise. The main issue with this approach is the unavoidable trade-off between privacy and usefulness: adding noise protects privacy but reduces data accuracy and can destroy the spatial patterns, like clusters, that analysts want to study.

The other approach is Geospatial Cluster Analysis. Here, computational geometry is used to study clusters. For example, Kapanski et al. [23] look at smart city water pressure sensors, using clustering and cluster boundaries to visualize and optimize the network. These boundaries improve analysis but are not used for anonymization.

This shows the gap our framework targets. Geomasking protects privacy but harms clusters. Geospatial analysis preserves clusters for utility but ignores privacy.

Our framework combines these ideas. It uses the geometric outline of a cluster (its convex hull) not for display, but as a restriction for making changes. Points are moved only within this outline, which is checked quickly using Delaunay triangulation [22]. The size of each movement is reduced according to how much influence the point has. With this design, the method changes individual points while still keeping the cluster’s shape, position, and internal influence pattern. It provides a practical, efficient, and testable CAM that solves the main usefulness problem found in standard geomasking.

Chapter 3

THE DATA PIPELINE

3.1 Phase 1: Initial Dataset Curation from PDF

3.1.1 Provenance

The source dataset of this project was a single, publicly available document: a PDF file titled "List of Hospitals - Pan India," published online by the Dr. B. R. Ambedkar National Institute of Technology (NIT), Jalandhar, containing 4,396 hospitals, their full address, city, and state [24]. This document, while comprehensive, was a PDF file, not anything computer-friendly. Moreover, it was curiously biased towards eye hospitals (which we had to eliminate) and also lacked geospatial data. We needed a way to:

a. Extract the table from this PDF, and
b. Extract global coordinates from the address.

3.1.2 Automated PDF Table Extraction

Extracting structured data from PDFs is notoriously difficult. To tackle this, we chose the `camelot` library in Python [25], a tool specifically designed to extract complex table layouts from PDFs. We faced a few practical challenges. The PDF's first page had a different table structure than the rest. Our solution was to process the first page separately and then programmatically concatenate the remaining tables. The column headers from the first page's data frame were then manually assigned to the unified dataset.

3.1.3 Initial Data Filtering Logic

The initial dataset showed a significant bias towards specialized eye care centres. To prevent this from skewing our analysis of general healthcare accessibility, we applied a filter to remove any entry where the name contained "eye," using `re.compile('eye', re.IGNORECASE)`. Also, to focus specifically on hospitals rather than smaller clinics or nursing homes, we filtered the dataset to only include entries containing the word

"Hospital." These filtering steps were deliberate acts of methodological definition.

3.2 Phase 2: Data Enrichment via Distributed Web Scraping

3.2.1 Technical Stack and Rationale

To extract coordinates and star ratings from addresses, we turned to web scraping. We chose the Selenium framework [12] because the target data, such as ratings and review counts on Google Maps, is loaded dynamically with JavaScript. A static scraper would fail. The Microsoft Edge browser, controlled via its corresponding WebDriver, was selected as our automation tool.

3.2.2 The Scraping Workflow

The process was broken down into a series of logical steps, encapsulated in Python functions.

First, we defined a `getSearchTerm(rowIndex)` function to create a high-precision search query. We searched the the hospital's name, city, and state concatenated to one string. This simple heuristic proved remarkably effective, often leading the browser directly to the correct location page on the mapping service.

The core of the operation was `scrape_data(driver)`. This is where the interaction with the webpage happened. We used `WebDriverWait` for dynamic JavaScript content. Instead of telling the script to simply "wait 5 seconds", we instructed it to wait *until a specific element was present on the page*. We targeted specific CSS selectors (`div.F7nice` for the ratings block and `button.CsEnBe` for the Plus Code, which is a location-encoding system). The extracted text was then carefully parsed, e.g., turning the web content `"1,168 reviews"` into the integer `1168`.

3.2.3 Error Handling & Robustness

Real-world web scraping is never clean. We had to use various workarounds to handle some unexpected conditions during the scraping process.

a. *Search Page instead of Hospital's page:* Sometimes, a search would lead to a list of results instead of a direct place page. Our code was built to handle this by automatically clicking the first result, identified by the CSS selector `a.hfpxzc`.

b. *The "Directions" Bug:* Occasionally, the automated search would trigger a navigation or directions panel. Seemingly, this is a bug in Google Maps itself. The script was designed to identify and click the appropriate close button, whether its label was "Close" or "Close directions."

c. *Backups from a crash:* A long-running scrape is vulnerable to crashes or network errors. To mitigate data loss, the script was programmed to save its progress to a CSV file every 10 hospitals. This created a checkpoint system, ensuring that hours of work wouldn't be lost due to a single unforeseen error.

3.2.4 Distributed Task Management

The scraping process was time-consuming. Hence, the initial, filtered hospital list was split into four separate CSV files, one for each author. We then each ran the same scraping script, each assigned to one of the four files. Once all four scraping tasks were complete, a final script was used to concatenate the four resulting CSVs back into a single dataset.

3.3 Phase 3: Geospatial Data Acquisition

We had erroneously assumed that the scraped "Plus Code" could be easily converted to latitude and longitude coordinates later on. We discovered that doing so programmatically and in bulk required using a paid Google Maps API, which was outside the scope of this project's resources. This necessitated a second scraping phase, a workaround born of necessity. The fix: search for a Plus Code on the map, which causes the browser to center on that location. Fortunately, the resulting page URL itself contained the coordinates (e.g., `.../@lat,lon,...`). The script was designed to trigger this URL change, close any informational panel that appeared, and then parse the latitude and longitude directly from the browser's current URL string using

```
driver.current_url.split('@')[1].split(',')[0:2].
```

3.4 Phase 4: Data Cleaning and Final Preparation

3.4.1 Imputation Strategies.

No scraped dataset is perfect; missing values are a certainty. We made deliberate methodological choices for how to handle these gaps.

a. For missing Latitude and Longitude values, we used a backward-fill method (`df['Latitude'].bfill()`). Hospitals were often listed in a geographically clustered manner. Therefore, it was reasonable to assume that a hospital with missing coordinates was located near the next hospital in the list.

b. For missing Rating and Number of Reviews, we chose to fill the gaps with the median value of the respective column. The median was chosen over the mean because it is far more robust to outliers.

3.4.2 Merging with Administrative & Population Data

The final and perhaps most important preparation step was to integrate our hospital dataset with official administrative and demographic data to demonstrate medical inequity

An unexpected ad-hoc task that occurred here was reconciling city names. The hospital data used common names ("Bangalore," "Calicut"), while the official district data used administrative names ("Bengaluru Urban," "Kozhikode"). A comprehensive dictionary, was manually created to map dozens of these variations. LLMs made this part of the process less labour-intensive.

The cleaned hospital data was merged with a district-density.csv file. This helped complement each hospital entry with the population density of the administrative district in which it was located.

Chapter 4

METHODOLOGY

The core of the project lies in the custom algorithm for anonymizing geospatial data. The algorithm must have two core features to be effective:

a) Preserve the geographical influence of each hospital while still anonymizing individual locations.
b) Be generalizable to any other geographical entity, not just hospitals, provided a definition of "influence".

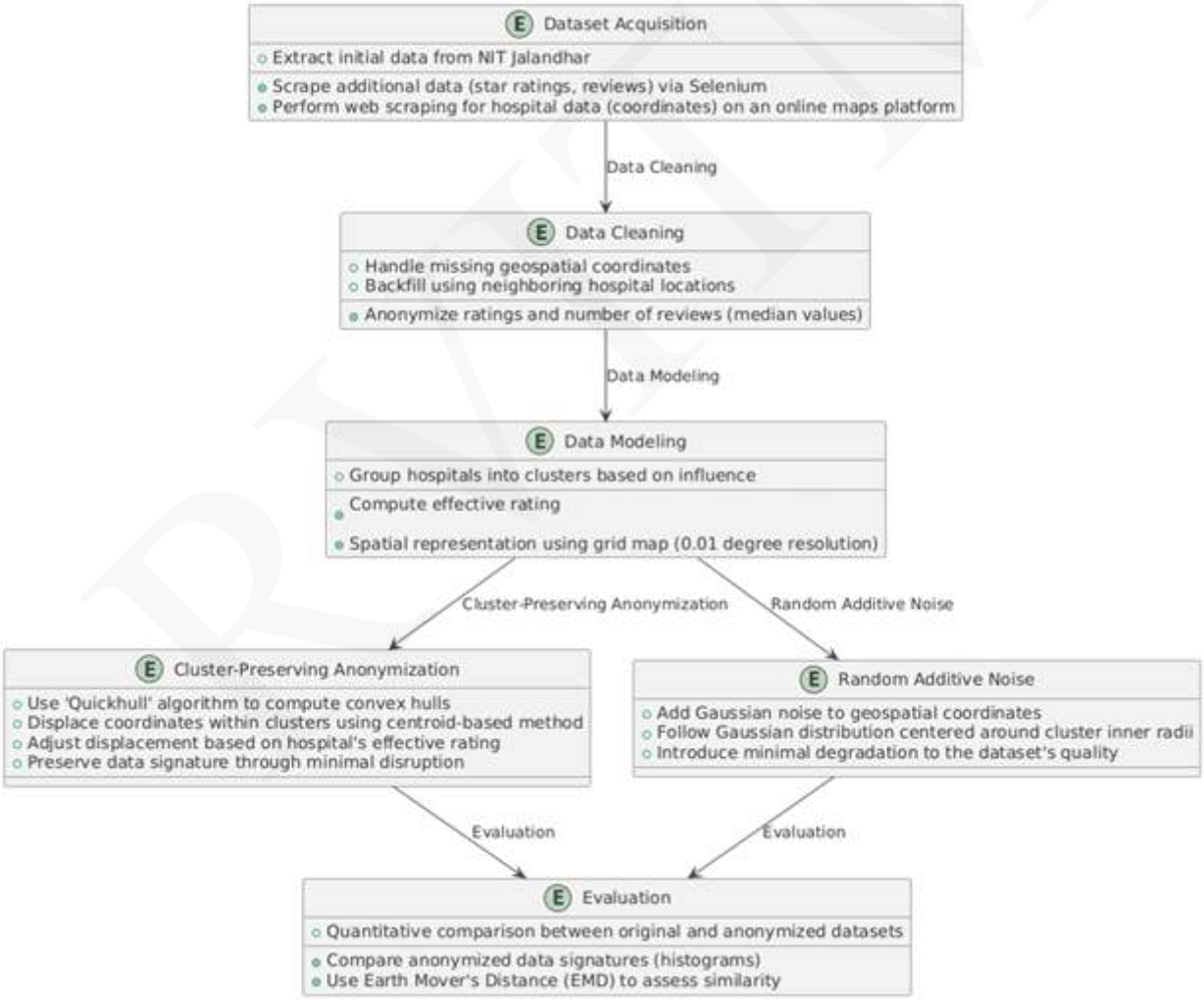

Fig. 4.1. Methodology of the process as a flowchart

4.1 Defining "Healthcare Influence"

4.1.1 The Problem with Raw Ratings

A simple 1-to-5-star rating is obviously an insufficient metric for a hospital's true prominence. For example, a new facility with a single 5-star review could misleadingly appear more reputable than an established hospital with a 4.5-star rating averaged over thousands of reviews. Raw ratings lack context and weight.

4.1.2 The "Effective Rating" (η)

To create a more nuanced metric, we created the "Effective Rating," denoted by η. This is defined by:

$$\eta = \alpha \cdot R \cdot \log(N)$$

where *R* is the star rating and *N* is the number of reviews. The use of the natural logarithm, log(N) models the concept of diminishing marginal value: the difference in credibility between 10 and 100 reviews is massive, while the difference between 1000 and 1100 reviews is far less significant. The constant α is an arbitrary coefficient used for scaling.

4.1.3 Normalization and Influence Radius

After calculating η for all hospitals, the values were normalized to a 0-5 scale. This brought the new metric back to the human-interpretable scale of five stars. This final, normalized Effective Rating was then used to define a "Radius of Influence" for each hospital, by multiplying it by a constant `RADIUS_FACTOR`. This radius would become an important parameter in the next stage of spatial modelling.

4.2 Spatial Modelling: The Influence Map

4.2.1 Discretizing Geospatial Coordinates:

To analyse spatial patterns, we first had to transform continuous latitude/longitude coordinates into a discrete grid. We defined a grid covering the Indian mainland with a

`STEP` of 0.01 degrees. This resolution was chosen because it corresponds to an approximate 1x1 square kilometre grid, an appropriate scale for regional analysis. We implemented helper functions, `get_idx` and `get_lat_lon`, to seamlessly convert between real-world coordinates and their corresponding grid indices.

4.2.2 The Influence Decay Function

A hospital's influence is not confined to its own 1x1 km cell. Instead, it radiates outwards. However, the farther a hospital is from the public, the less influential it becomes. We modelled this using an influence decay function, specified as

$$i = ae^{-\lambda r^2}$$

The choice of an exponential function is to model a real-world phenomenon where a hospital's influence is strongest at its center and fades rapidly with distance.

4.2.3 Vectorized Map Generation

The `get_influence_map` function generates the final influence map. A naive implementation might loop through every hospital and then loop through every cell on the grid, which would be computationally expensive. Instead, we used a vectorized approach. For each hospital, we defined a bounding box around its radius of influence. Within this box, NumPy's meshgrid was used to create coordinate grids, allowing us to calculate the distance from the hospital to all surrounding cells in a single operation. A mask was then applied to ensure that influence was only added to cells actually within the circular radius. This method is orders of magnitude much faster than nested loops.

4.3 Cluster Identification and Signature Generation

4.3.1 Clustering via Percentile Thresholding

The resulting influence map is a continuous landscape of values. To identify distinct clusters, we needed to apply a threshold. The cluster function implements this by first calculating the 50th percentile of all *non-zero* influence values. Any cell with an influence

value below this threshold was set to zero. This acts as a powerful noise filter, removing areas of low influence and allowing the significant healthcare hubs to emerge as distinct islands.

4.3.2 Cluster Labelling

Once the map was binarized, we used the `scipy.ndimage.label` function to identify contiguous regions of non-zero cells. This function scans the grid and assigns a unique integer ID to each distinct cluster, using 8-adjacency. Each contiguous region is essentially a cluster.

4.3.3 The "Dataset Signature"

After identifying the clusters, we calculated the magnitude of each cluster by summing the original influence values of all cells within it. We then took the logarithm of these magnitude values and plotted them as a histogram. This distribution represents the "Dataset Signature." It is a unique fingerprint of the healthcare landscape, showing the distribution of cluster sizes: many small clusters, a few medium ones, and perhaps one or two very large ones. This signature becomes our ground truth, the essential pattern that our anonymization algorithm must preserve.

4.4 The Anonymization Algorithm (Cluster-Preserving)

4.4.1 Geometric Boundary Definition

The first step in our anonymization is to understand the precise geometric shape of each cluster. For this, we used the convex hull algorithm, specifically, the QuickHull implementation [22]. For a given set of points (the grid cells of a cluster), the convex hull is the smallest convex polygon that contains all the points.

4.4.2 Defining "Safe" Displacement: The Inner Radius (r)

Simply knowing the outer boundary isn't enough. We need to define a "safe" region *inside* which we can move points. The logic is as follows: first, calculate the geometric centroid (the average coordinate) of all points in the cluster. Then, calculate the distance from this

centroid to every point on the convex hull. The smallest of these distances is the radius of the largest circle that can be drawn inside the hull, centred at the centroid. This is the "inner

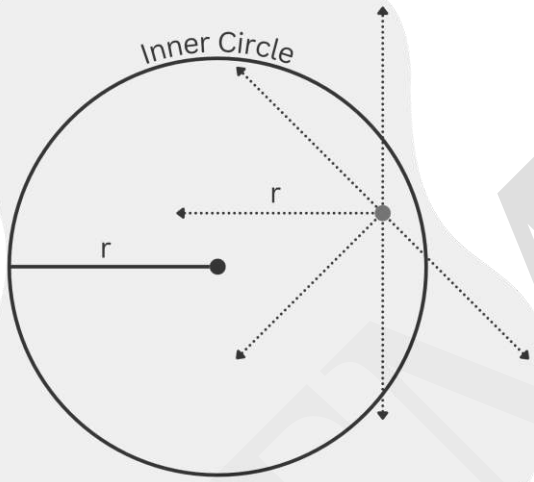

Fig. 4.2. A typical cluster is shown here, along with its incircle of radius r. The dotted lines, each also measuring r, represent the valid displacements of a sample point within the cluster.

radius" (r), as depicted in Fig. 4.2. It defines a guaranteed-safe zone for displacement.

4.4.3 Ensuring In-Cluster-Anonymization: Delaunay Triangulation

To ensure that a displaced point, even if moved beyond the inner circle, still remains within the cluster's original footprint, we use another concept from computational geometry: Delaunay Triangulation [22]. By creating a Delaunay triangulation of the hull points, we create a mesh that perfectly covers the area of the cluster. The check `delaunay_region.find_simplex(new_coords) >= 0` is then a highly efficient way to ask: "Is the new coordinate point located inside one of the triangles of this mesh?" If the answer is yes, the point is guaranteed to be inside the original cluster's convex hull.

4.4.4 The Overall Algorithm

To summarize the algorithm is as follows:

1. For a given hospital, identify the cluster it belongs to.
2. Generate a random noise vector (a random direction).
3. Scale the magnitude of this vector. Calculate as `noise_magnitude = np.random.uniform(0,min_radius)/effective_rating**2`. The displacement is inversely proportional to the square of the hospital's Effective Rating. This means that highly influential hospitals (high η) are moved *less*, preserving their prominence and the overall structure of the influence map. Less influential hospitals are moved *more*, maximizing their anonymization.
4. Apply the scaled noise to the hospital's coordinate.
5. Check to verify if the new point is still inside the cluster's hull.
6. If it falls outside, discard it and try again with a new random vector, up to 256 times.

4.5 The Control Algorithm (Naive Gaussian Noise)

4.5.1 A "Best-Effort" Naive Approach:

We implemented a naive anonymization method using a Gaussian. However, instead of choosing arbitrary parameters, we framed it as a "best-effort" naive approach. The mean and standard deviation of the Gaussian noise were derived directly from the statistics of the inner radii of all clusters found in the original dataset. This gives the naive method the *best possible chance* to succeed, as its displacement parameters are informed by the actual spatial structure of the data.

4.6 Final Anonymization Steps

4.6.1 Obfuscating Non-Spatial Data

We obfuscated identifying information by replacing hospital names with generic IDs like id = "Hospital #1234".

4.6.2 Perturbing Ratings and Reviews:

We couldn't simply add random noise to both Ratings and Reviews, as this would destroy the carefully calculated Effective Rating (η) that underpins our entire spatial model. Instead, we first added a tiny amount of noise to the Rating (e.g., `np.random.normal(0, 0.05)`). Then, we used the formula for η to *back-calculate* the new Number of Reviews required to preserve the original η value as closely as possible. This ensures that while the individual metrics are perturbed for privacy, the hospital's overall influence metric, which is essential for the dataset's usability, remains intact.

Chapter 5

RESULTS

5.1 Visualizing the Dataset

5.1.1 Geospatial Distribution (Fig. 5.1)

The initial scatter plot of hospital coordinates immediately revealed a non-uniform distribution. Clear, dense clusters were visible, corresponding directly to India's major metropolitan areas: Delhi, Mumbai, Kolkata, Chennai, Bengaluru, and Hyderabad. This visualization confirmed the inherent clustering of the data and provided a first look at the urban-centric nature of healthcare infrastructure in the country.

5.1.2 Effective Rating Bubble Plot (Fig. 5.2)

The bubble plot provided a better view. Here, the size and colour intensity of each bubble represented its "Effective Rating." This moved beyond simple location to visualize influence in a crude way before implementing the influence map.

5.1.3 Influence Map (Fig. 5.3)

The influence map of the dataset brings out the same metropolitan clusters, but now shows the "hotspots" of high-influence facilities within them. More importantly, it also visualized the inverse: the vast areas with smaller, fainter bubbles, or no bubbles at all. These are the potential healthcare "deserts." This visualization asks the questions that drive public policy: Where is high-quality healthcare concentrated, and which regions are being left behind?

5.2 Quantitative Evaluation: Earth Mover's Distance (EMD)

To move beyond subjective visual comparison, we turn to the quantitative analysis of the "Dataset Signature" histograms. These plots show the distribution of the logarithm of cluster magnitudes.

5.2.1 Ground Truth (Fig. 5.4 (a))

This histogram is the fingerprint of the original dataset. It shows a characteristic distribution with a strong central peak, indicating a large number of similarly-sized mid-range clusters, and tails on either side for the few very small and very large clusters.

5.2.2 Cluster-Preserving (Fig. 5.4 (b))

The signature of our anonymized dataset is strikingly similar to the original. It retains the same general shape, spread, and the height and position of the central peak. This visual similarity is quantified by a low Earth Mover's Distance (EMD) of **0.4002**. EMD [3] can be thought of as the minimum "work" required to transform one distribution into another; a low value signifies high similarity.

5.2.3 Naive Noise (Fig. 5.4 (c))

The signature from the naively anonymized data is visibly distorted. The central peak is reduced, and the distribution is flattened and more irregular. It has lost the characteristic shape of the original. This visual dissimilarity is quantified by its much higher EMD of 1.3015.

5.3 Visualizing the Cluster-Based Signatures

5.3.1 Ground Truth: Original Clusters (Fig. 5.5 (a)):

This plot shows the clusters identified from the original, unaltered influence map, using the 50th percentile cutoff. Each distinct colour represents a unique cluster of healthcare influence. This map is our baseline, the "ground truth" spatial distribution that a successful anonymization technique must preserve in essence, if not in exact detail.

5.3.2 Result 1: Cluster-Preserving Anonymization (Fig. 5.5 (b))

This figure displays the clusters after applying our proposed anonymization algorithm. A qualitative visual analysis shows an extremely high degree of fidelity to the original. The clusters largely maintain their original shape, size, geographic location, and separation

from one another. While the individual points within have been shifted, the macro-structure, which is essential for regional analysis, remains remarkably intact. The change is minimal and subtle.

5.3.3 Result 2: Naive Gaussian Noise (Fig. 5.5 (c))

The result of applying the naive Gaussian noise method stands in stark contrast. The visual deterioration is immediately obvious. The once-crisp clusters are now smeared and fragmented. Some clusters have bled into one another, while others have been broken apart. The structure in the "Deccan region," for example, is visibly distorted and scattered compared to the ground truth in Fig. 5.5(a). This qualitative evidence strongly suggests that naive noise addition, even when carefully parameterized, fails to preserve the essential spatial characteristics of the dataset.

5.4 Figures

5.4.1 The Dataset

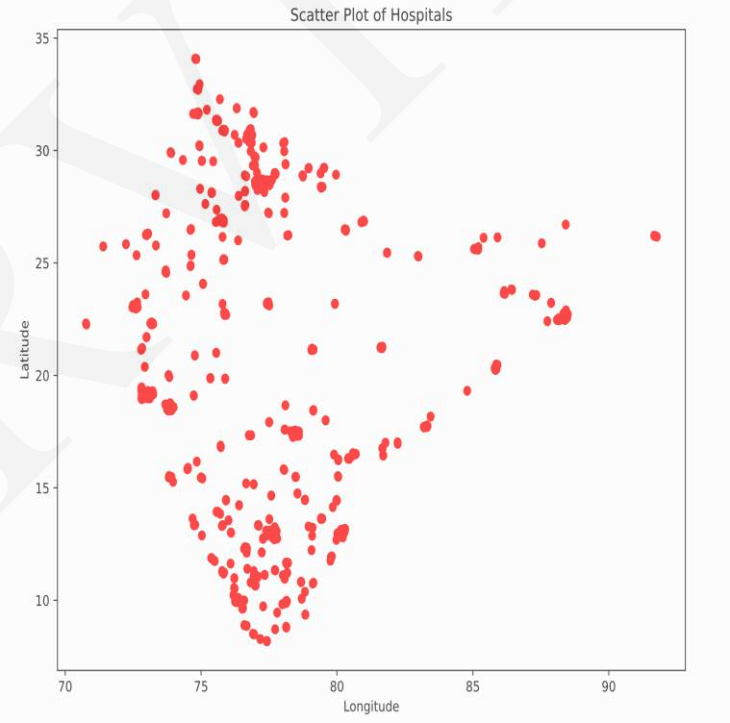

Fig. 5.1. Scatterplot of hospitals in India.

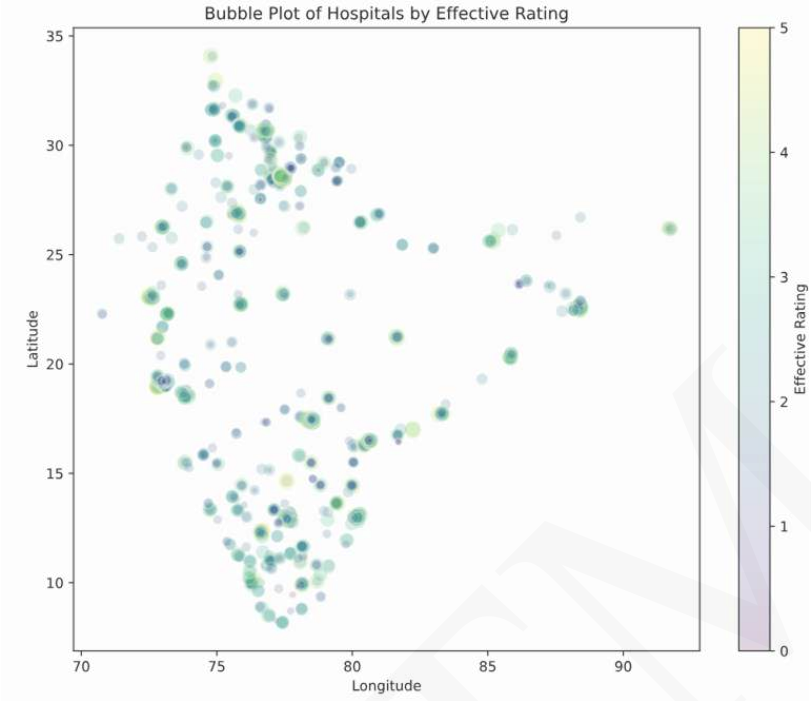

Fig. 5.2. Bubble plot of hospitals by the calculated effective rating

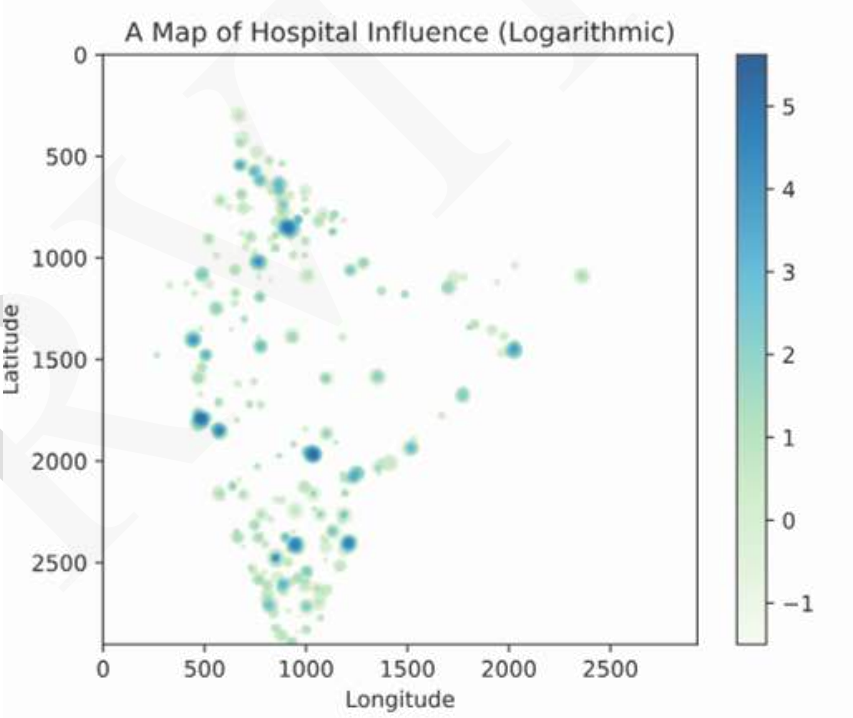

Fig. 5.3. Influence Map after exponential decay

5.4.2 Histogram Comparisons

A histogram of cluster magnitudes in the original dataset, serving as a baseline for comparison with anonymized datasets, compared with both the cluster magnitudes after cluster-preserving anonymization and naive anonymization. Minimal differences are observed after cluster-preserving anonymization while the naive anonymization significantly diverges.

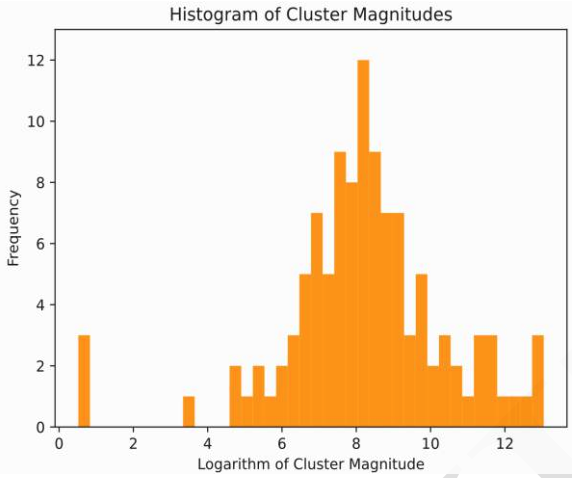

Fig. 5.4 (a). Signature of the original dataset, depicting the distribution of hospital influence based on effective ratings.

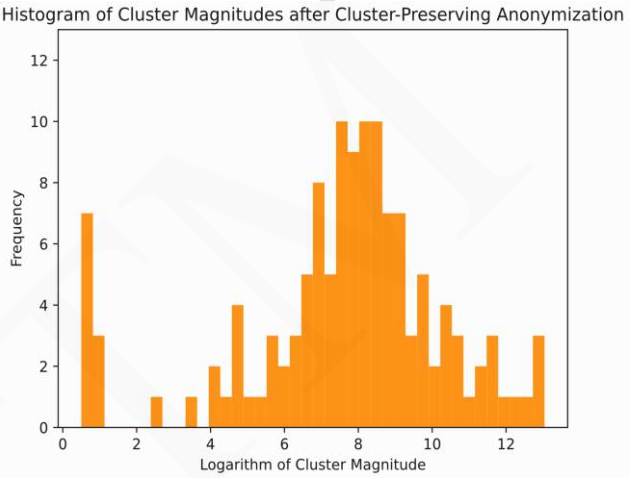

Fig. 5.4 (b). Signature of the dataset anonymized using clustering-based methods, exhibiting similar spread, central peak height and overall shape to the original dataset.

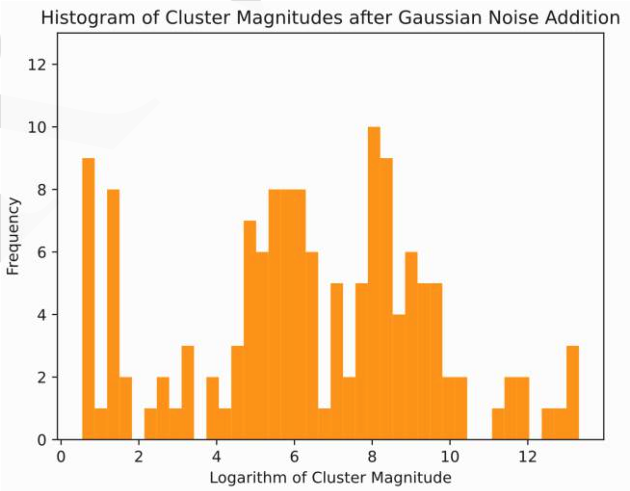

Fig. 5.4 (c). Signature of the naively anonymized dataset, showing irregular patterns and reduced central peak compared to the original dataset.

5.4.3 Clusters

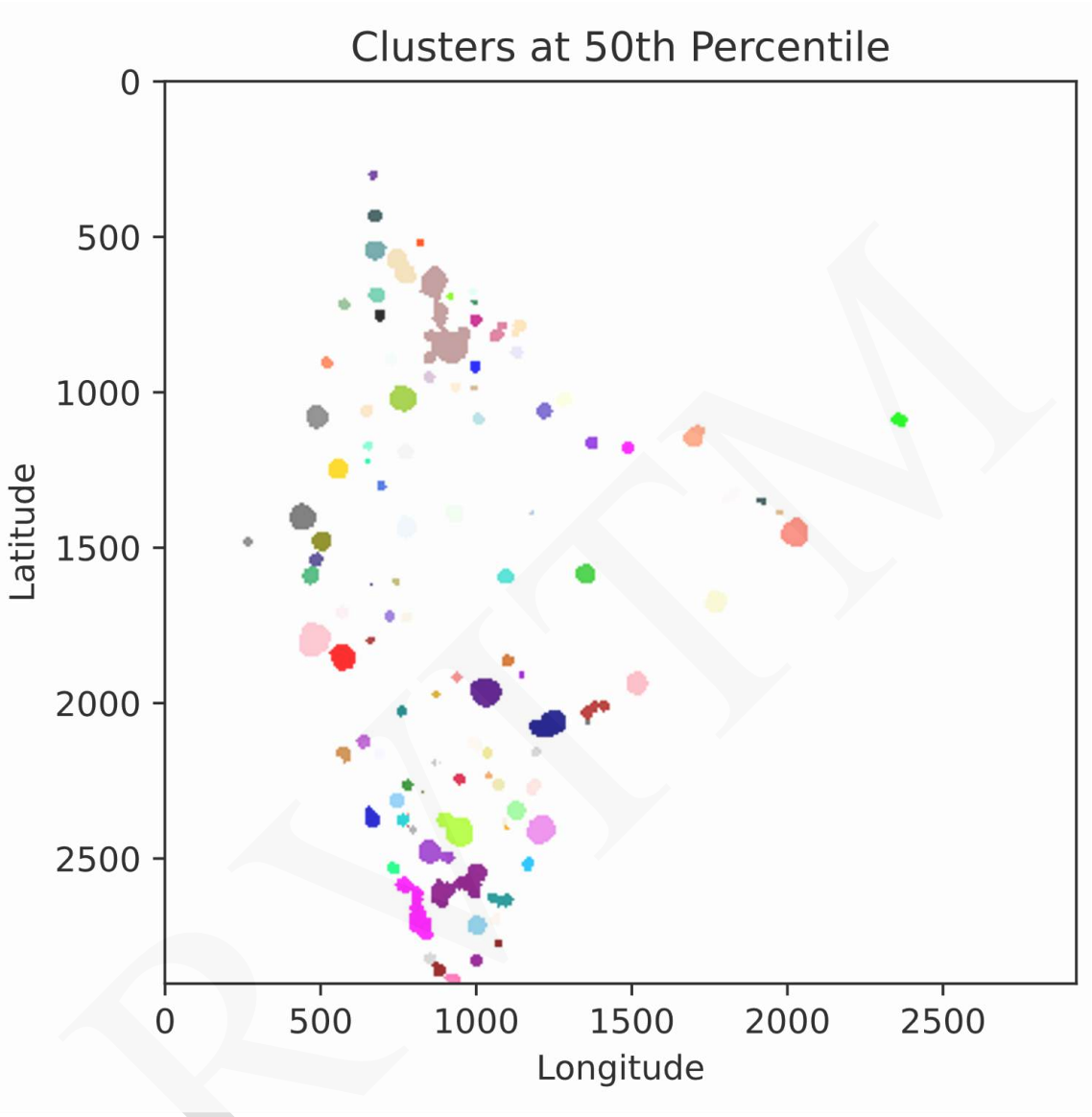

Fig. 5.5. (a) Clusters representing the top 50% of hospital influences by effective rating.

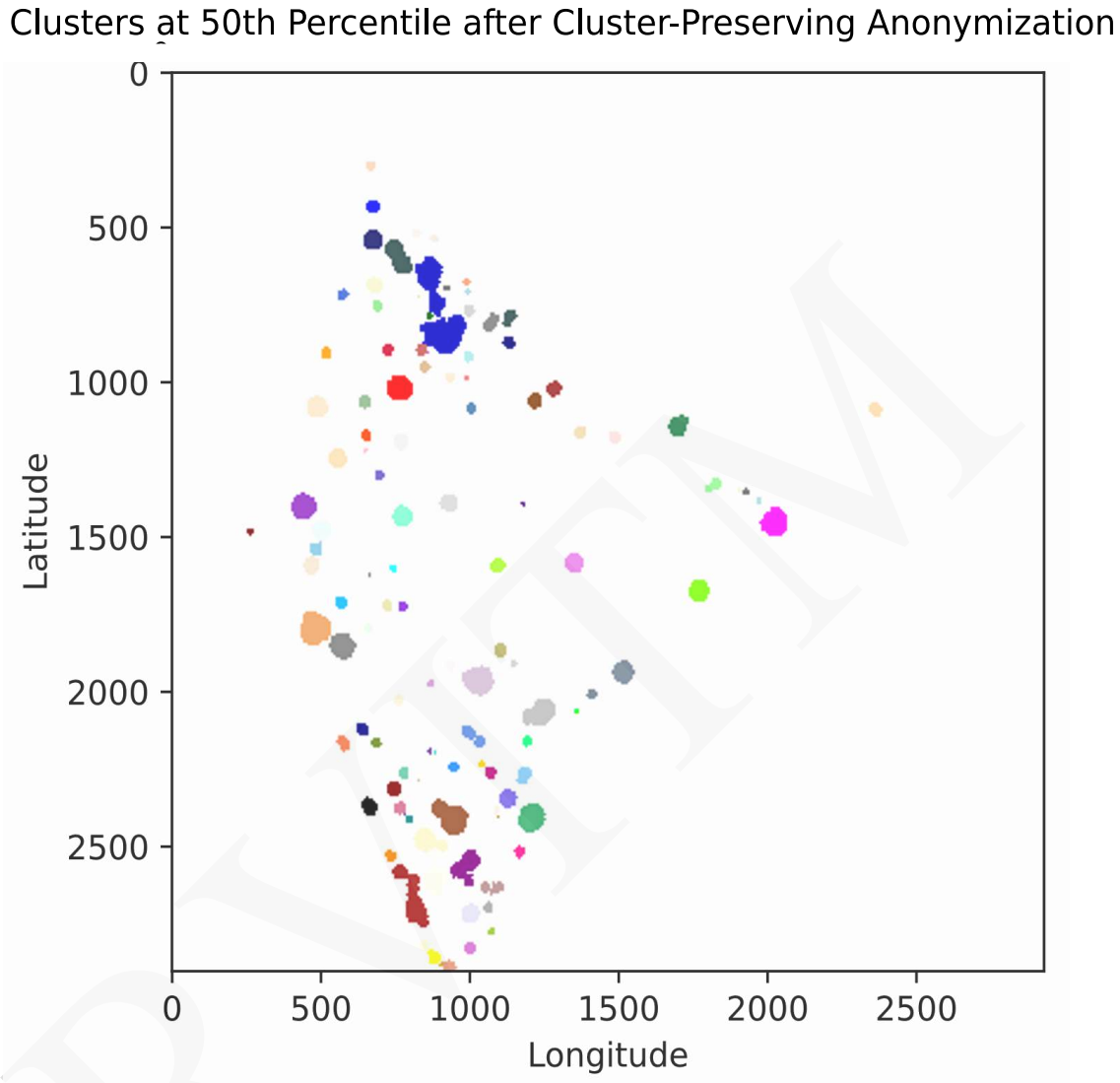

Fig. 5.5. (b) Hospital-influence clusters after applying clustered anonymization, where cells are displaced within each cluster's incircle radius. Minimal differences are observed when compared to the original data.

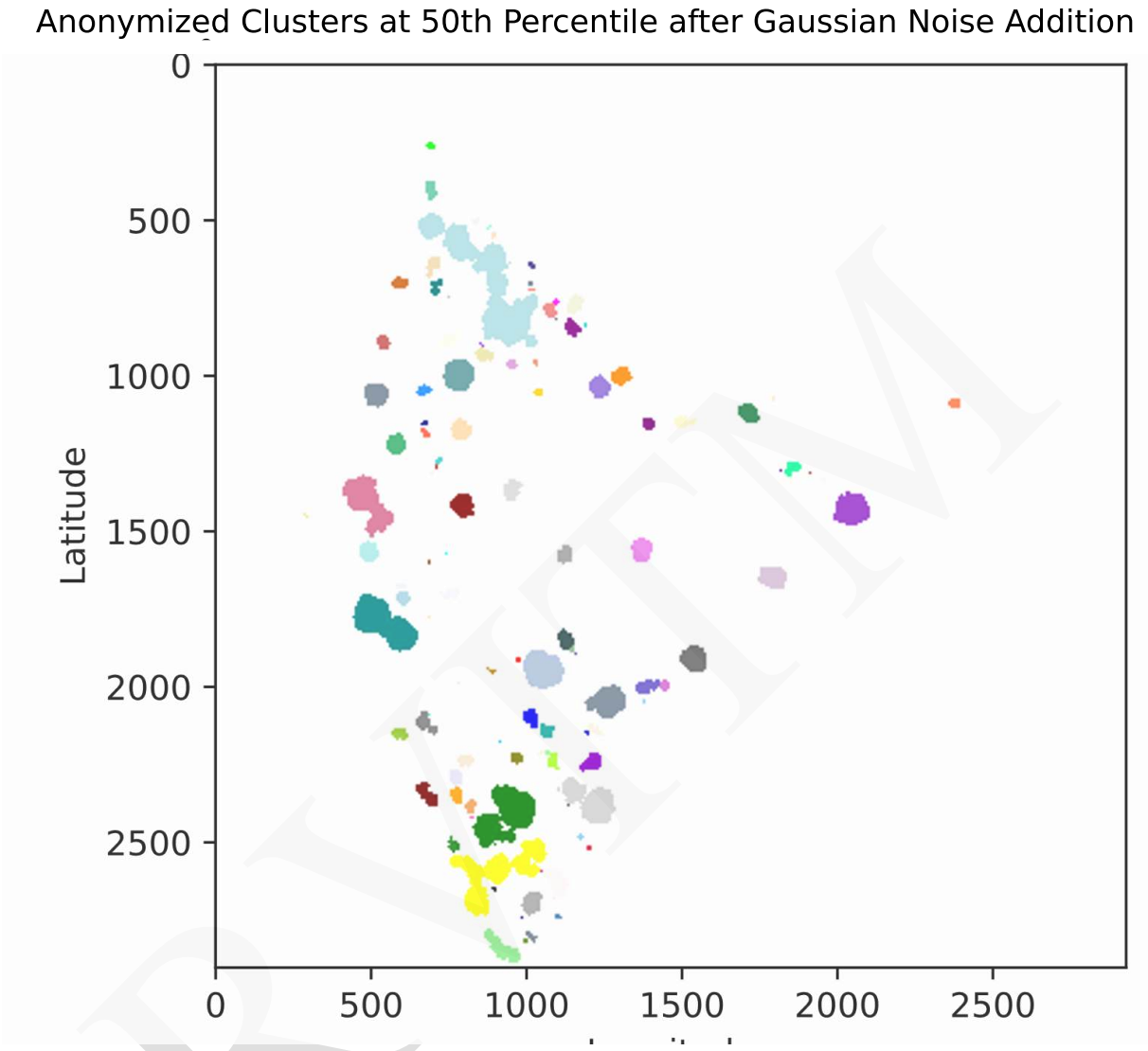

Fig. 5.5 (c). Hospital-influence clusters after random noise addition, showing noticeable changes in cluster structure, scale, and distribution in contrast to the original dataset. and clustered anonymization results.

Chapter 6

IMPACT AND SUSTAINABILITY

6.1 Answering the Central Question: Utility *and* Privacy

The core of this research was to find a solution to the utility-versus-privacy paradox. The results from the previous chapter demonstrate that we have successfully done so. The anonymized dataset produced by our algorithm, is *safe* to share and analyse because the precise locations of individual hospitals have been obfuscated. It simultaneously retains the same *usability*, evidenced by the low EMD score of 0.4002 and the well-preserved signature in Fig. 5.4 (b). We have developed a method that resolves the central conflict, proving that (in this use-case), it is possible to protect privacy without destroying the value of the data. This is the foundation of using data for good, and doing it right.

6.2 Contextualizing the Research (ICAECT 2025)

The publication tied to this work was presented in the Fifth International Conference on Advances in Electrical, Computing, Communications and Sustainable Technologies (ICAECT 2025) [1]. The conference celebrates contributions to the broader theme of technological advancement for a sustainable future in India. Other papers presented in this conference include those that rethink Machine Learning for air quality prediction, evaluating the cost of generating renewable energy with artificial neural networks, and more traditional optimization algorithms for minimizing energy mismatch in Electric Vehicle Batteries.

6.3 Sustainability

In this section, we would like to discuss the broader impact of our work in a developing nation like India. Specifically, we use to the United Nations Sustainable Development Goals [4] to gauge beneficial, sustainable progress.

6.3.1 Impact on Public Health (SDG 3)

The most direct application of this work is in public health, addressing SDG 3 (Good Health and Well-being). The preserved cluster map (Fig. 5.5 (a)) is a directly usable tool for policymakers, researchers, and NGOs. Because the macro-structure is intact, it allows them to perform the exact kind of analysis the raw data was intended for: to understand the "distribution of quality healthcare" across the nation. They can identify regions of high influence and, more importantly, regions of "sparsity." This provides a data-driven basis for resource allocation, helping to decide where new hospitals are needed most, where to invest in upgrading existing facilities, and how to ensure a more equitable distribution of healthcare services for all citizens.

6.3.2 A Tool for Justice (SDG 10: Reduced Inequalities)

This is where the true power of the full data pipeline, specifically the merging with population density data in Section 3.4.2, comes to fruition. The *anonymized* dataset can now be used to conduct a powerful and safe secondary analysis directly targeting SDG 10 (Reduced Inequalities). Researchers can now correlate the healthcare influence map with the population density map. This allows us to move beyond simple observation and ask pointed, data-driven questions for social justice: "Which high-density urban slums have low healthcare influence?" or "Are there densely populated rural districts that are effectively healthcare deserts?" By enabling this analysis on a privacy-preserved dataset, our work provides an evidence-based tool to identify and address systemic inequalities in healthcare access. It helps shift the conversation to concrete data, giving legislators an idea of medical inequity in the nation.

6.4 Impact in Kaggle

The fully anonymized dataset generated from this project was published on Kaggle for researchers, policymakers, healthcare analysts or data science hobbyists. The final dataset consisted of 2557 hospitals located across the country and includes location information, ratings, and the number of reviews. As of January 2025, the dataset has over 100 downloads [26].

Chapter 7

CONCLUSION AND FUTURE WORK

7.1 Summary of Contributions

This research project set out to solve the challenge of balancing data utility with privacy in the context of geospatial healthcare analysis in India. We have successfully delivered on this goal through three primary contributions. First, we engineered and meticulously documented a complete, end-to-end data pipeline, demonstrating a reproducible method for transforming a static public document into a rich, spatially-aware dataset. Second, we designed and implemented a novel cluster-preserving anonymization algorithm that leverages computational geometry to obscure individual locations while maintaining the vital macro-level structure of the data. Third, we provided conclusive quantitative proof of our method's superiority, showing that its ability to preserve the dataset's "signature" is 3.25 times better than a best-effort naive approach, as measured by Earth Mover's Distance. We have framed this technical work within a "data for good" narrative and see it as an essential tool for advancing India's Sustainable Development Goals.

7.2 Limitations of the Study

a) Our work began with a single, static PDF file. This source, while valuable, is inherently incomplete and may not represent a fully comprehensive or up-to-date list of all hospitals in India.

b) Our influence heuristic, which decays exponentially with distance, does not account for situations where physical proximity does not mean practical accessibility. For instance, patients who live on the road directly behind the hospital may need to travel several times the actual distance to reach the hospital’s front entrance.

c) The web scraping process, reliant on Selenium, is inherently brittle. It depends on the stability of third-party website structures, which can change without warning, potentially breaking the data collection scripts.

d) The retry loop (up to 256 attempts) to ensure a point lands within the cluster's convex hull is perhaps inelegant, even though it proved extremely effective in practice. There definitely exists a more deterministic geometric method that could be used instead.

7.3 Future Work

The most immediate next step is to use the final, anonymized, and population-enriched dataset to *perform* the analysis discussed in Chapter 6. A dedicated study correlating healthcare influence with population density could yield powerful insights into healthcare inequality in India.

The brute-force retry loop could be replaced with a more sophisticated displacement method. For instance, an algorithm could be developed to project any point that falls outside the convex hull back to the nearest point on the hull's boundary, guaranteeing a valid displacement in a single step.

The data pipeline can be re-run on an annual basis. This would create a powerful *temporal* dataset, allowing researchers to track changes in healthcare accessibility and influence over time. Such a longitudinal study could be used to measure the real-world impact of public health policies and investments.

REFERENCES

[1] H. S. Bojnal, D. SVK, J. K. Kaarthik, D. Kotian and S. Virupaksha, "Privacy Preservation of Cluster Integrity on Web-Scraped Hospital Data," in *2025 Fifth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, Bhilai, India, 2025.

[2] H. S. Bojnal, "GitHub - fringewidth/towardsMedicalEquity," 2024. [Online]. Available: https://github.com/fringewidth/towardsMedicalEquity. [Accessed January 2025].

[3] "A metric for distributions with applications to image databases," in *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, Mumbai, India, 1998.

[4] "THE 17 GOALS | Sustainable Development," United Nations Department of Economic and Social Affairs, Sustainable Development, [Online]. Available: https://sdgs.un.org/goals. [Accessed January 2025].

[5] V. Krotov, L. Johnson and S. Leiser, "Legality and Ethics of Web Scraping," *Communications of the Association for Information Systems,* vol. 47, pp. 539-563, 01 2020.

[6] A. Bruns, "After the 'APIcalypse': social media platforms and their fight against critical scholarly research," *Information, Communication & Society,* vol. 22, no. 11, pp. 1544-1566, 2019.

[7] M. A. Brown, A. Gruen, G. Maldoff, S. Messing, Z. Sanderson and M. Zimmer, *Web Scraping for Research: Legal, Ethical, Institutional, and Scientific Considerations. arXiv Preprint,* arXiv 2410.23432, 2024.

[8] D. Trezza, "To scrape or not to scrape, this is dilemma. The post-API scenario and implications on digital research," *Frontiers in Sociology,* vol. 8, 2023.

[9] B. Stringam, J. Gerdes and C. Anderson, "Legal and Ethical Issues of Collecting and Using Online Hospitality Data," *Cornell Hospitality Quarterly,* vol. 64, 2021.

[10] M. Dogucu and M. Cetinkaya, "Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities," *Journal of Statistics Education,* vol. 29, pp. 1-24, 2020.

[11] E. Uzun, "A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages," *IEEE Access,* vol. 8, pp. 61726-61740, 2020.

[12] "Selenium ver 4.31," The Selenium Browser Automation Project, [Online]. Available: https://www.selenium.dev/.

[13] N. Khalid, A. Qayyum, M. Bilal, A. Al-Fuqaha and J. Qadir, "Privacy-preserving artificial intelligence in healthcare: Techniques and applications," *Computers in Biology and Medicine,* vol. 158, 2023.

[14] R. Torkzadehmahani, R. Nasirigerdeh, D. Blumenthal, T. Kacprowski, M. List, J. Matschinske, J. Spaeth, N. Wenke and J. Baumbach, "Privacy-Preserving Artificial

Intelligence Techniques in Biomedicine," *Privacy-Preserving Artificial Intelligence Techniques in Biomedicine,* vol. 61, pp. e12-e27, 2022.

[15] S. B. Dodda, S. Maruthi, R. R. Yellu, P. Thuniki and Byrapu, "Federated Learning for Privacy-Preserving Collaborative AI: Exploring Federated Learning Techniques for Training AI Models Collaboratively While Preserving Data Privacy," *Australian Journal of Machine Learning Research & Applications,* vol. 2, 2022.

[16] X. Qu, Q. Hu and S. Wang, "Privacy-Preserving Model Training Architecture for Intelligent Edge Computing," *Computer Communications,* vol. 162, p. 94–101, 2020.

[17] A. Aminifar, M. Shokri, F. Rabbi, V. K. I. Pun and Y. Lamo, "Extremely Randomized Trees With Privacy Preservation for Distributed Structured Health Data," *IEEE Access,* vol. 10, pp. 6010-6027, 2022.

[18] P. Samarati and L. Sweeney, "Protecting Privacy When Disclosing Information: k-Anonymity and Its Enforcement Through Generalization and Suppression," 1998.

[19] B. Gedik and L. Liu, "Protecting Location Privacy with Personalized k-Anonymity: Architecture and Algorithms," *IEEE Transactions on Mobile Computing,* vol. 7, pp. 1-18, 2008.

[20] M. A. Lorestani, T. Ranbaduge and T. Rakotoarivelo, *Privacy Risk in GeoData: A Survey. arXiv Preprint,* arXiv 2402.03612.

[21] A. Majeed, S. Khan and S. O. Hwang, "Toward Privacy Preservation Using Clustering Based Anonymization: Recent Advances and Future Research Outlook," *IEEE Access,* vol. 10, pp. 53066-53097, 2021.

[22] F. P. Preparata and M. I. Shamos, "Computational Geometry - An Introduction," Springer-Verlag, 1985.

[23] A. A. Kapanski, R. V. Klyuev, A. E. Boltrushevich, S. N. Sorokova, E. A. Efremenkov, A. Y. Demin and N. V. Martyushev, "Geospatial Clustering in Smart City Resource Management: An Initial Step in the Optimisation of Complex Technical Supply Systems," *Smart Cities,* vol. 8, 2025.

[24] Dr. BR Ambedkar National Institute, "List of Hospitals - Pan India," 22 October 2019. [Online]. Available: https://v1.nitj.ac.in/nitj_files/links/List_of_Hospital_-_Pan_India_28496.pdf. [Accessed 25 October 2024].

[25] "Camelot: PDF Table Extraction for Humans ver 1.0," Camelot Developers, 2024. [Online]. Available: https://camelot-py.readthedocs.io/en/master/.

[26] H. S. Bojnal, D. SVK, J. K. Kaarthik and D. Kotian, *Hospitals In India,* Kaggle, 2024.