Assessment 2 **Technical Answers to Real World Problems**

Members:
Bedant Patnaik (20BCE0568), Shreedhar Hegde (20BCE0576), Sai Ruthvik Kasi (20BCE0929), Anuj Aghi (20BCE0932), Asmita Pal (20BCE2454), Shreyoshi Kamboj (20BCI0244)

Title: De-Identification of Facial Images and Tabular Data to Preserve Data Privacy using L-Diversity

1. **Project Timeline:**

Project Timeline

Name	Duration	Start	End
Problem Statement Identification	2days	28-08-2023	30-08-2023
Literature Survey	4days	05-09-2023	09-09-2023
Formulating Requirements	5days	09-09-2023	14-09-2023
Finalizing Algorithm for anonymization of facial image	2days	17-09-2023	19-09-2023
Finalizing Algorithm for anonymization of tabular data	2days	25-09-2023	27-09-2023
Coding	30days	28-09-2023	27-10-2023
Testing	10days	28-10-2023	08-11-2023
Improvisation	6days	09-11-2023	15-11-2023
Documentation	2days	16-11-2023	18-11-2023

2. **Introduction:** With video surveillance proving to be an integral part of current security infrastructure, privacy rights are beginning to gain importance. The key concern is the fact that private citizens, who are not suspects, are being recorded and recordings archived through the use of video surveillance systems. Such a record-everything and process-later approach has serious privacy implications.

   The same privacy issues arise when surveillance cameras routinely record highway traffic as vehicle tags are recorded. The solution of removing the identities by blurring/blackening the portions of video is not acceptable to security personnel as they may have legitimate need to review the videos. On the contrary, leaving the videos with identities of people and vehicles public is a breach of privacy. A solution to the problem is selective encryption of portions of the video that reveal identity (e.g., faces, vehicle tags) in surveillance applications. Regions of a video can be encrypted to ensure privacy and still allow decryption for legitimate security needs at any time in the future. The goals of the video surveillance are still met as selective encryption allows monitoring the activities without knowing the identities of those being monitored. When a suspicious activity needs to be investigated, the identities can be uncovered with proper authorization. In this project we will use SAP HANA cloud to do analytics on data in addition to storing and managing images. This will involve executing queries and doing analysis to detect existing patterns or trends in data.

3. **Detailed Literature Review:**
   Katsuhiro Honda, Masahiro Omori, Seiki Ubukata, and Akira Notsu presented their work "A research on fuzzy clustering-based k-anonymization for privacy preserving crowd movement analysis with face recognition" at the 2015 International Conference on Soft Computing and Pattern Recognition. It provides a unique approach to k-anonymization, a technique used in data mining to preserve privacy, especially for crowd movement analysis utilizing facial recognition. The authors compare their proposed method to many different k-anonymization strategies, and the findings reveal that the suggested fuzzy clustering-based method surpasses other methods in terms of both data quality and privacy protection. While choosing a privacy-preserving data mining approach, the study emphasizes the need of evaluating the application's unique requirements.

T. Nakamura, Y. Sakuma, and H. Nishi presented their paper "Face Picture Anonymization as an Application of Multidimensional Data K-Anonymizer" at the 2019 Seventh International Conference on Computing and Networking Workshops. It presents a unique technique to k-anonymization for face picture anonymization, which is critical for maintaining privacy in a variety of applications, including social media and public monitoring. The authors highlight the significance of privacy in face recognition and the difficulties in maintaining privacy while yet allowing meaningful analysis of face photos. The authors compare their suggested approach to various existing k-anonymization approaches, and the findings reveal that the proposed multidimensional data k-anonymizer approach beats the others in terms of data quality and privacy protection. While choosing a privacy-preserving data mining approach, the study emphasizes the need of evaluating the application's unique requirements.

Yan, Herman, Mahmood, et al study "A weighted K-member clustering method for K-anonymization," published in the journal Computing in 2021, presents a unique k-anonymization approach that employs weighted clustering to safeguard sensitive information while keeping data quality. The suggested method separates the data into clusters based on similarity and assigns weights to each record depending on sensitivity. The authors compare the performance of their proposed algorithm to that of various existing k-anonymization algorithms, and the findings reveal that the new algorithm beats the others in terms of data quality and privacy protection. The study emphasizes the necessity of evaluating the application's unique needs when choosing a privacy-preserving data mining approach and argues that the suggested algorithm may be employed in a variety of applications, including healthcare, banking, and social media.

Mehta and Rao's paper "Improved l-diversity: Scalable anonymization approach for Privacy Preserving Big Data Publishing," published in 2022 in the Journal of King Saud University - Computer and Information Sciences, proposes an improved l-diversity technique for scalable anonymization of big data, with the goal of achieving higher levels of privacy protection while minimizing data distortion. The strategy works in two steps: first, the data is partitioned into subgroups based on similarity, and then the sensitive qualities in each subset are varied using a probabilistic approach. The authors evaluate the performance of their proposed anonymization approach to that of many other strategies, and the findings reveal that the suggested technique surpasses the other techniques in terms of both privacy protection and data quality. While choosing a privacy-preserving data publication approach, the study emphasizes the need of evaluating the application's unique requirements.

4. **Overview of the Proposed Work:**
This paper attempts to use the aforementioned l-diversity methodologies to enhance existing fuzzy-clustering-based k-anonymity algorithms. Accordingly, two prototypes shall be prepared to follow the following two algorithms outlined in the diagrams below. The first diagram on the left outlines the workflow of the proposed algorithm for anonymising pictographic medical records, while the second outlines that for anonymising tabular data and patient records.