IDENTIFYING VENDORS USING TOPOLOGICAL DATA ANALYSIS METHODS

MSc Data Science Dissertation - Interim Report

Author: Deepika Chandramouli
Supervisor: Dr Mahbub Gani, Dr Tong Xin

1 Introduction

Over the past decades, there have been several advancements in the field of Artificial Intelligence (AI) paving way for numerous innovations in machine learning and deep learning. However, Topological Data Analysis (TDA) remains an under-explored field, with recent developments in applied algebraic topology and computational geometry bringing it to the forefront [1]. While the historical roots of geometric approaches to data analysis are more classical, TDA formally emerged as a distinct field through groundbreaking contributions by Edelsbrunner et al. (2002) and Zomorodian and Carlsson (2005) in the branch of persistent homology [1][2][3]. These contributions focus on identification of significant topological features in data and analysis of complex shapes and structures across scales [2]. TDA encompasses an array of methods that reveal the shape or structure of the data. It is motivated by the principle that topology and geometry offer a strong framework for extracting qualitative and quantitative information about the structure of data [4][5]. Thus, TDA methods aims to offer strong geometrical, mathematical, statistical and algorithmic methodologies for analysing and leveraging intricate topological structures [1].

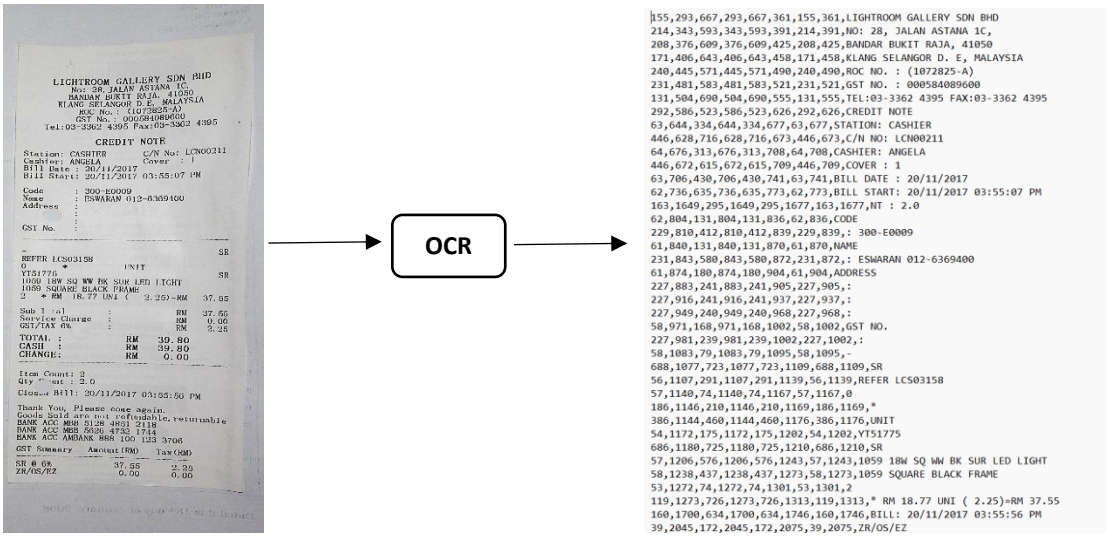
Significant advancement in TDA by Nicolau pinpointed a discrete subclass within breast cancers, highlighting its capacity to reveal unique patterns and traits within intricate datasets [6]. Other areas like material science explored by Kramar et al (2013) investigate structural characteristics in persistence systems particularly in machine behaviours [7]. 3D shape analysis has also seen advancements and successful results were demonstrated by Skraba et al (2010) in computer vision and pattern recognition, particularly in segmenting shapes that undergo distortion. Another motivation to TDA approaches arose when scientific datasets began to grow in size. The large sized datasets make it difficult to efficiently identify the target variables making it ineffective. There are no optimal comparisons after the algorithm's performance indicating the need for advanced data analysis algorithms. Using topological and geometric approaches we get intricate nuances into the data analysis and visualisation enabling concise and thorough capture of the structure of data.

Topology is majorly used in the computer vision tasks particularly with images in identifying the region of interest. *Figure 1* depicts how prior topology or bounding box information helps in efficiently segmenting the images with the point of focus [9].

Figure 1: Topology or bounding box used to efficiently segment an image to the region of interest. Image taken from [9]

The topological structures are considered as data points and are usually represented as rectangular coordinate distances within x and y coordinates, majorly Euclidean distance [1]. Therefore, other areas,

such as the deer and grass depicted in the above figure, are not the main focus, as attention is solely directed towards the region of interest specified by the topology or bounding box rectangle coordinates. This research focuses on using the bounding box topological information and applying data analysis techniques on the topology to efficiently identify the target variables (vendor in this case). The primary aim of this research is to utilize the topological bounding box information from invoices or receipts to identify and forecast vendor companies. Scanned Receipts OCR is a technology used for retrieving the bounding boxes through Optical Character Recognition. This method involves scanning an invoice receipt to extract crucial information such as text elements like company names, addresses, etc., along with their corresponding bounding box coordinates. These extracted details are then utilized in the research's algorithms and analyses to effectively determine the vendor linked with the invoice. Figure 2 illustrates how a Scanned Receipt could have the essential information through OCR technology.



a) b)

Figure 2: Scanned Receipt OCR a) Scanned Receipt before OCR b) Scanned Receipt after OCR with key information such as text and bounding boxes

Scanned Receipts are facing many breakthroughs in many practical tasks such as document-intensive processes such as entity, amount and name recognition and hence it has found a growing area in the financial and accounting sector [10]. It has also seen limitations in terms of performance and efficiency using the Machine Learning (ML) algorithms and hence this research will focus on creating a base line using ML algorithms and applying the underexplored field of topological analysis on scanned receipts data.

2 Aim and Objectives

2.1 Aim

The aim of this research is to revolutionize the vendor identification in invoice processing by leveraging TDA methods. By incorporating advanced machine learning algorithms within topological data, it is aimed to increase the correctness in predicting vendor companies. This approach also provides a robust framework to capture meaningful patterns from complex invoices eliminating the limitations in current methods. Ultimately it uses TDA methods to streamline business processes in efficient vendor identification enabling more informed decision-making in various industries

2.2 Objectives

This work will focus on four primary objectives:

- Develop and implement the baseline model using machine learning algorithms to predict vendor companies based on information relating to topological bounding boxes obtained from scanned receipts.
- Investigate and apply topological data analysis methods to obtain significant features from scanned receipt data for vendor identification.
- Compare the performance, in terms of accuracy and efficiency, between the baseline machine learning model and the proposed topological data analysis approach.
- Provide actionable recommendations to accounting and financial operations by streamlining the vendor identification tasks, minimizing errors in recognition, cost benefit analysis, and deducing practical suggestions for adoption.

Optional objectives time permitting:

- Explore the potential of deep neural network-based classifiers, such as convolutional neural networks (CNN) or recurrent neural networks (RNN), for improving vendor identification tasks.
- Evaluate and compare the effectiveness of machine learning, deep learning and topological data analysis approaches in vendor identification, considering metrics such as accuracy, precision, recall, F1-score, and Mean Intersection over Union (mIoU).

3 Overview of progress

A thorough literature review including application areas of TDA methods has been successfully conducted providing a strong foundation for the project. Invoice and receipt image data study has been diligently undertaken to know the formats and type of data. Data collection, preparation and exploratory data analysis has been undertaken on Scanned Receipts OCR and key Information Extraction (SROIE) dataset v2, which is the publicly available version from Kaggle [11]. The bounding boxes are represented as 8-pixel coordinates forming a rectangle; for instance, 72,25,326,25,326,64,72,64: top left, top right, bottom right and bottom left. Here, top left and bottom right pixels are required to draw bounding box rectangles on the receipt images. Visualisations of the bounding box coordinates indicate the accuracy of the boxes on the receipt images.

Figure 3 indicates invoice samples with bounding box coordinates highlighted as rectangles which are colour coded. The figure indicates the bounding boxes based on entities like company (blue), address (red), total (green) and date (yellow). All other bounding boxes are represented as grey. Several preprocessing checks like normalisation, text orientation angle correction, centroid and area of the bounding boxes arrived to generate a sparse matrix. This matrix serves as the feature vector for our algorithms. Logistic regression and K-nearest neighbours were applied on the feature vector X with vendor company as the target variable y. The results indicate an accuracy of over 61% in the test set indicating possible amendments are required to the feature vector.

There are around 236 unique companies which is a very large number. Hence, class imbalance correction must be undertaken to have a better effectiveness of the model.

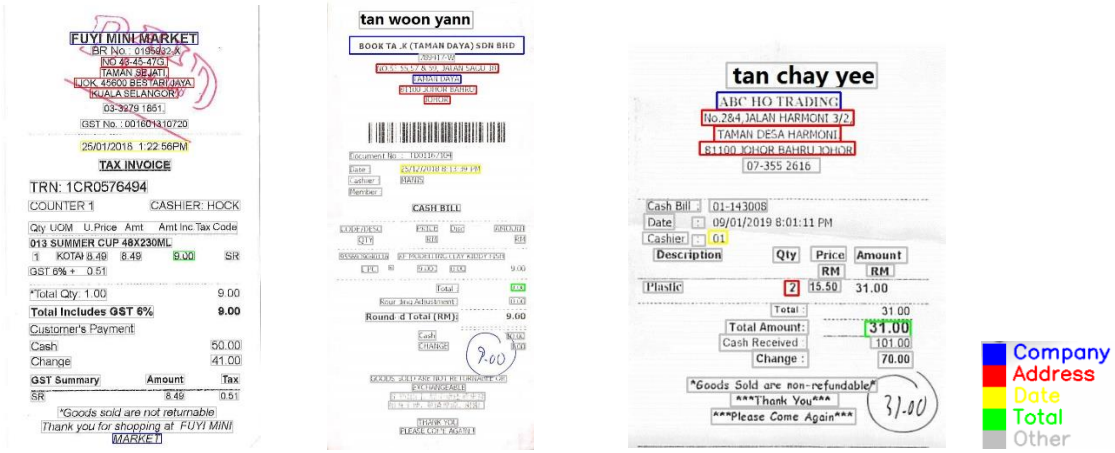


Figure 3: Three samples of SROIE data with exploratory data analysis on the text information

The indicates the necessary areas that need to be corrected to have an effective baseline with the machine learning algorithms. Gradient boosting algorithms could also be explored to have a better precision over the result. Further, deep learning models could be tried if time permits. The baseline will be compared over topological data analysis methods using the python libraries dedicated to topology, which is in the early stages. This approach will contribute to a more practical and efficient nature of this research, leading to improved adoption and implementation.

4 Project Plan

The project timeline extends from 24th April 2024 to 15th August 2024, with a comprehensive visual plan illustrated in Figure 4. Preliminary research involved identifying applications of Topological Data Analysis (TDA) in the context of vendor fraud detection from invoices. Additionally, a practical introduction to TDA and a review of numerous research papers on topology and geometrical analysis were conducted. A thorough review of literature has been compiled ready for the final report. Data collection, preparation, and exploratory data analysis have been nearly completed. Research on datasets featuring bounding boxes (topology) including Kaggle invoice dataset and YOLO data analysis methods have been fully reviewed. The SROIE dataset has been collected, and exploratory data analysis on the SROIE invoice data is complete. Outliers have been identified and eliminated. The implementation phase, identified as the primary component of the project, is scheduled from May 15 to August 6. This phase involves various AI approaches, beginning with machine learning algorithms and advancing to topological data analysis methods, which form the core of this research. Results will be evaluated by comparing baseline approaches to topological methods to accurately identify vendors using bounding box coordinates. Although it is anticipated that the topological approach will yield high overall accuracy, a more robust model should be achieved. If time permits, additional datasets may be explored with the proposed architectures, including class imbalance correction and data augmentation. Metrics such as Mean Intersection over Union (MioU) could also be compared with the obtained results. Project deliverables like interim report, poster creation, thesis report and oral presentation have been defined and progress is examined.

Project Repository

- https://github.com/deepikachandru/TopologicalAnalysis

References

[1] Frédéric Chazal and Bertrand Michel: An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists

[2] Edelsbrunner, Letscher & Zomorodian Topological Persistence and Simplification. Discrete Comput Geom 28, 511–533 (2002).

[3] Zomorodian, A. and Carlsson, G. (2005). Computing persistent homology. Discrete Comput.

[4] Carlsson, G. (2009). Topology and data. AMS Bulletin, 46(2):255–308.

[5] Chazal, F., Cohen-Steiner, D., and Merigot, Q. (2010). Boundary measures for geometric inference. Found. Comp. Math., 10:221–240

[6] Monica Nicolau, Arnold J. Levine, and Gunnar Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proc. Nat. Acad. Sci., 108(17):7265–7270, 2011.

[7] Kramar, M., Goullet, A., Kondic, L., and Mischaikow, K. (2013). Persistence of force networks in compressed granular media. Physical Review E, 87(4):042207.

[8] Skraba, P., Ovsjanikov, M., Chazal, F., and Guibas, L. (2010). Persistence-based segmentation of deformable shapes. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, pages 45–52.

[9] V. Lempitsky, P. Kohli, C. Rother and T. Sharp, "Image segmentation with a bounding box prior," 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 2009, pp. 277-284, doi: 10.1109/ICCV.2009.5459262. keywords: {Image segmentation; Iterative algorithms; Power generation economics; Mice; Active contours; Computer vision; Linear programming; Image reconstruction},

[10] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shjian Lu, and C.V. Jawahar- ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction

[11] SROIE v2 dataset from Kaggle- https://www.kaggle.com/datasets/urbikn/sroie-datasetv2 - Introduced by Huang et al. in ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction

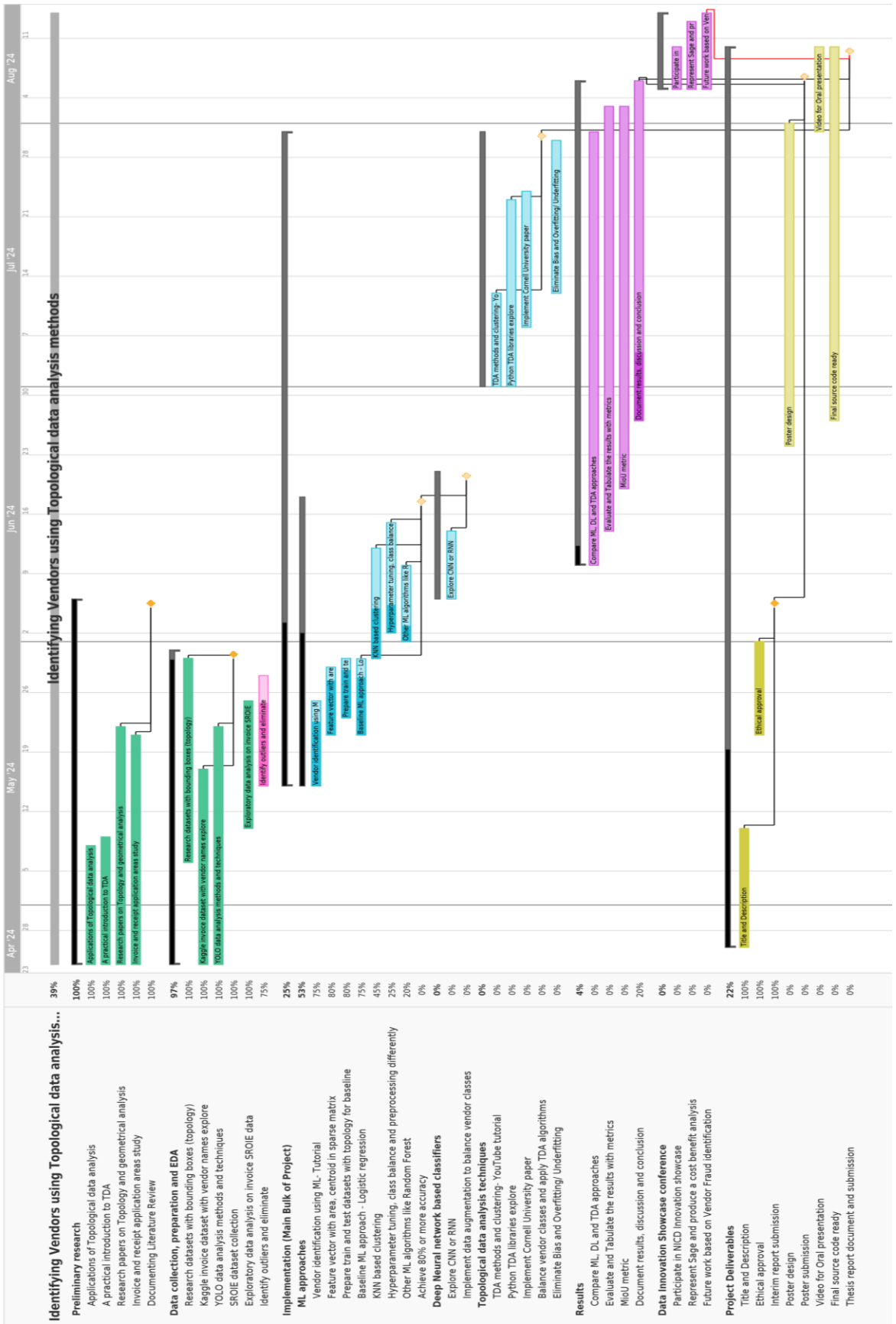


Figure 4: A Gantt Chart of the proposed project. Tasks are scaled according to the date range, with dependencies indicated by lines and milestones by ◆

DATA MANAGEMENT PLAN.

0. Project title, author, version and date

Project: *Identifying Vendors using Topological data analysis methods*

Author: *Deepika Chandramouli*	***Version:*** *1*	***Date:*** *5th June 2024*

1. Description of the data

1.1 Type of study

This study involves a comparative analysis of different AI methodologies, focusing on Topological Data Analysis (TDA) to enhance vendor identification processes. It includes extensive exploratory data analysis, implementing machine learning models, and integrating TDA techniques to evaluate their effectiveness. The study compares these approaches against existing methods, leading to a cost-benefit analysis and actionable recommendations.

1.2 Types of data

Only publicly available datasets are considered for this project. The Scanned Receipts OCR and key Information Extraction (SROIE) dataset v2 (which is the publicly available version) from Kaggle is used for the research (https://www.kaggle.com/datasets/urbikn/sroie-datasetv2/data). It has quantitative data and images with train and test folders containing scanned invoice and receipts. The dataset has been collected by several researchers as part of a competition and rearranged to Kaggle. The samples are in English along with bounding box annotations for text and vendor information. If time permits, Intelligent Data Capture and Analysis for Receipts (IDCAR) dataset will be explored which was the source for SROIE. (https://rrc.cvc.uab.es/?ch=13&com=introduction).

1.3 Format and scale of the data

SROIE dataset contains part of IDCAR data with focusing only on receipts and invoices. The dataset contains 973 scanned receipts. For each receipt you have an .jpg file of the scanned receipt, a .txt file holding OCR information and a .txt file holding the key information values like vendor, address, total and date [10][11]. The dataset is 875 MB in size. Each image has dimensions roughly around 463 x 1013 pixels. These files are processed using Python libraries such as OpenCV for image processing and TensorFlow/Keras for machine learning tasks. Since it is already a benchmark suite the FAIR principles (Findability, Accessibility, Interoperability and Reusability) are readily satisfied.

2. Data collection / generation

2.1 Methodologies for data collection / generation

The existing dataset will be utilized for research. The data is publicly available from the Kaggle SROIE dataset v2, following community data standards. The images and annotations provided in this dataset adhere to commonly accepted formats and conventions, ensuring compatibility and consistency with existing research practices. No new image data or bounding box annotations are aimed to be gathered or produced. The results of the research comprising data will however be tabulated clearly in the report. Thus, annotated images are stored either locally or in GitHub with utmost transparency. The FAIR principles of data will be adhered to whenever necessary.

2.2 Data quality and standards

Our focus lies in ensuring the high quality of the data we utilize and documenting our processing methods effectively. By utilizing established datasets, we eliminate the need for data entry or recording processes. Additionally, our research methodology does not involve peer review of data. Instead, we conduct experiments to assess the accuracy of bounding boxes in matching receipts, ensuring the reliability and validity of our findings. Our emphasis is on implementing good practices and standards to maintain the quality of the dataset.

3. Data management, documentation, and curation

3.1 Managing, storing and curating data.

SROIE dataset v2 has been downloaded from Kaggle to a local machine and uploaded to a Google Drive account to be used in conjunction with the Google Colab environment. Data will be stored and managed using Google Drive. Backup of training data is not a concern since it is available in cloud. Checkpoints are performed to store any trained weights as part of algorithms and those files are again placed in google drive for seamless access. Milestones like achieving a best accuracy from a model weight are stored in GitHub for further documentation and thesis report generation.

3.2 Metadata standards and data documentation

Any trained weights from the algorithms and hyperparameters will be detailed and appended to the main report. Further, the exact values of the hyperparameters to reproduce the run will be specified in the README file created in the project GitHub repository. Such a README file will also contain how the schema was for the training, which algorithm was used, which hardware specifications were used, and which environmental settings were used for that training and validation cycle. Clear documentation of instrument metadata, including hardware and software tool specifications for data collection and analysis are structured. In addition, history of data is captured from google Colab, google drive and GitHub to trace its origin and history, establishing transparency and accountability. It will also help record the time taken for each cycle to benchmark the efficiency of any models created, especially in scenarios where live image segmentation for applications in autonomous driving is required.

4. Data security and confidentiality of potentially disclosive information

4.1 Main risks to data security

Since training data is entirely public data with no personal information, security issues with third party storage solutions (raised by ncl.ac.uk https://www.ncl.ac.uk/library/academics-and-researchers/research/rdm/working/) are not a concern. The data is stored in google drive with no issue on the size and security.

5. Data sharing and access

5.1 Suitability for sharing

Yes, the data utilized is completely accessible to the public and has undergone extensive peer review, being cited in numerous published papers. It can easily be downloaded from Kaggle using the link shared in 1.2. This project aims to identify vendor companies from the dataset publicly available, thus sharing data, code and experiments will be encouraged.

5.2 Discovery by potential users of the research data

The research data as said in 5.1 is a publicly available data from Kaggle SROIE v2 dataset. Any researchers in the field of topology and geometrical analysis can make use of this data. Further in this research, all code libraries are kept in the public project GitHub repository (https://github.com/deepikachandru/TopologicalAnalysis). Should the project achieve its goal, publication will be pursued, given its broad applicability across domains for fair AI. This will include the generation of a DOI to aid in its discovery. This could potentially help Sage Plc company (with whom the research is conducted) identify its account payable vendors eliminating risks and frauds in invoice.

5.3 Governance of access

The research data could be used by any user interested to research on topology and receipt annotated data. It will be accessed through publicly available Kaggle dataset with no decisioning, restrictions or governance made as it is open source.

5.4 Restrictions or delays to sharing, with planned actions to limit such restrictions

The README file will outline clear procedures for data sharing. Since the dataset has complete public accessibility, any concerns regarding confidentiality are eliminated. Intellectual property such as algorithm or implementation procedure will be openly distributed to foster advancements in the field, with explicit terms detailed in the README file of GitHub repository. The author will disclaim any liability for the misuse of the developed software, code or dataset.

6. Responsibilities and Resources

Apart from the PI, who is responsible at your organisation/within your consortia for:

- *study-wide data management- Deepika*
- *metadata creation- Deepika*
- *data security- publicly available and secured data, further processing is held responsible by Deepika*
- *quality assurance of data- publicly available and quality assured dataset in accordance with FAIR principles*

Are there any resources (e.g. storage/ training) that you will require to fulfil the plan?

None

7. Relevant institutional, departmental or study policies on data sharing and data security

Policy	**URL or Reference**
Data Management Policy & Procedures	https://www.ncl.ac.uk/media/wwwnclacuk/research/files/ResearchDataManagementPolicy.pdf
Information Security	*https://services.ncl.ac.uk/itservice/policies/InformationSecurityPolic y-v2_1.pdf*
Other	*Data sharing and security are publicly available and accessible*

8. Author of this Data Management Plan (Name) and, if different to that of the Principal Investigator, their telephone & email contact details

Deepika Chandramouli