IDENTIFYING VENDORS USING TOPOLOGICAL DATA ANALYSIS METHODS

Deepika Chandramouli
MSc in Data Science
School of Computing, Newcastle University
d.chandramouli2@newcastle.ac.uk

Abstract

This study explores how Topological Data Analysis (TDA) can be used to improve the accuracy of vendor identification from scanned receipts. For businesses, it is critical that vendor identification is accurate to streamline accounts payable processes as well as prevent any fraudulent activities. This study leverages persistent homology in extracting and analysing topological features that come from bounding box coordinates which have been produced by Optical Character Recognition (OCR). The study compares two approaches upon SROIE dataset: A baseline model that makes use of traditional feature extraction methods like calculating areas and centroids from bounding boxes and a TDA-based method that employs persistent homology to produce feature vectors from persistence diagrams. Both these methods are applied to traditional machine learning algorithms to classify and predict vendor companies. The hypothesis is that TDA might show significant trends within unstructured invoice data which could improve vendor recognition ability. This study is conducted in collaboration with Sage Group and its goal is to extend TDA's applicability in invoice processing to make informed decisions in financial processes.

1 Introduction

In recent years, artificial intelligence (AI) and machine learning have brought forth important developments in various sectors. However, topological data analysis (TDA) is one of such areas that has not been adequately explored. TDA, having roots in applied algebraic topology and computational geometry, emerged as a common technique for analysing complex data structures [1]. This research intends to use TDA for improving the identification of vendors from scanned receipts which is an important process in financial transactions where precision and speed are essential. 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 is to utilise the topological bounding box information from invoices or receipts to identify and predict vendor companies. By leveraging TDA's unique capabilities, vendor recognition systems' accuracy and efficiency can be improved by addressing some limitations of current methodologies.

1.1 Background

Businesses from which goods or services are acquired are referred to as vendors (e.g., Walmart, Aldi), and precise identification of these vendors from receipts is fundamental for financial analysis and recording. The study aims at solving the challenges posed by conventional methods in exact identification of suppliers. Existing techniques have difficulties with heterogeneous nature of receipt data that consists of different text formats, fonts, layouts and qualities. This further complicates data extraction and analysis due to unstructured and noisy nature of cash register information [2]. A perfect example for this is seen in how traditional rule-based systems that process specific patterns

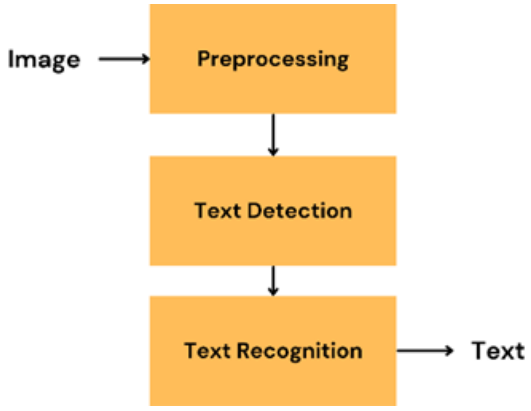

Figure 1: OCR- Image to text extraction process

and templates are unable to work with various types of receipts from vendors. These systems lack the flexibility to adapt to new or unusual receipt layouts, leading to high levels of mistakes in identifying suppliers [3]. On the other hand, Scanned Receipts Optical Character Recognition (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 utilised in the research's algorithms and analyses to effectively determine the vendor linked with the invoice. Figure 1 illustrates the process of OCR which converts image to text.

1.2 Motivation

Scanned Receipts are facing many breakthroughs in practical tasks including document-intensive processes such as entity, amount and name recognition and hence it has found a growing area in the financial and accounting sector [4]. 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. In order to enhance the efficiency as well as precision of financial processes it is crucial to tackle these problems. Moreover, converting extracted information into meaningful insights remains challenging. In the case of structured data, traditional machine learning models perform well, while they are not applicable on the intricate patterns present in receipt data. Consequently, it leads to lack of effectiveness in financial procedures and greater chances for mistakes and fraudulent activities [5]. Therefore, new techniques are needed that would be able to manipulate these complexities within receipt information better.

This research intends to tackle these challenges by applying TDA methods to examine topological features obtained from scanned receipts' bounding box information. It is motivated by the fact that topology and geometry offer a strong framework to reveal both qualitative and quantitative features of data [1]. Additionally, it provides stable geometrical, mathematical, and algorithmic frameworks for analysing sophisticated topological configurations [8][10]. This consists of using OCR extracted texts and bounding boxes as shown in Figure 2, and then applying TDA techniques on them to increase accuracy in vendor identification. The hypothesises is that TDA will outperform conventional machine learning approaches because it can identify complex patterns and structures in data better [6] [7]. This study proceeds as follows: First, a baseline model for bounding box receipt data is developed employing traditional machine learning algorithms which serve as a benchmark for performance comparison. Then TDA techniques are applied to extract and analyse topological features from the receipt data. Their effectiveness is compared with those obtained through the baseline model. Finally, actionable recommendations are provided for improving the vendor identification process.

This approach not only aims at enhancing the vendor recognition technology but also serves as an example which can gauge the general use of TDA methods in several AI based services [8] [9]. The research outcome shall provide a fresh perspective on financial transactions and real solutions that could shape future achievements in systems automating invoice processing and preventing frauds.

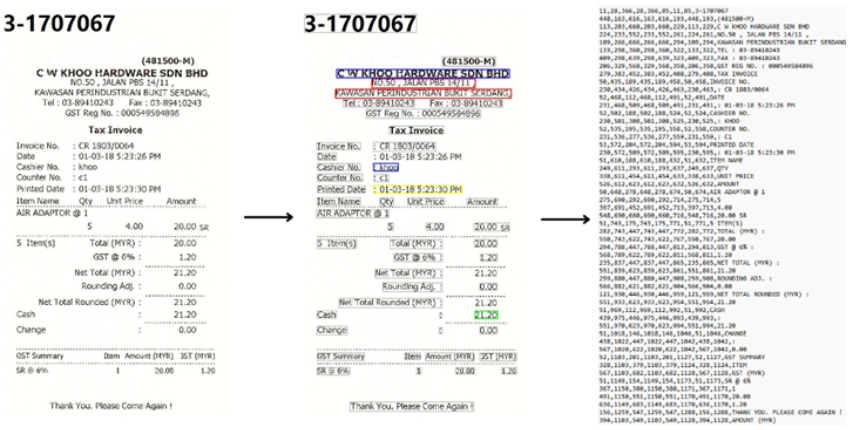

Figure 2: SROIE Dataset Sample- OCR Steps for Text Detection and Extraction

2 Related Work

TDA emerged as a distinct field through the groundbreaking contributions by Edelsbrunner et al. (2002) and Zomorodian and Carlsson (2005) in persistent homology, it became a separate area of study that is concerned with finding important topological characteristics from datasets for example, it involves examining complex shapes as well as their structures within various scales [6][8][9]. 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 [11]. Other areas like material science explored by Kramar et al (2013) investigate structural characteristics in persistence systems particularly in machine behaviours [12]. 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 [13]. Another reason 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 [8].

TDA has been revealed to be useful in identifying fraud-related behaviours and improving risk assessment models in case of financial transaction data. For instance, S. Mitra and K. Rao (2021) applied TDA to credit card transaction data in order to reach identification of any anomalies as well as patterns suggestive of fraudulent transaction behaviour. They found out that TDA could detect fraudulent purchases which are missed by conventional statistical methods [14]. This shows how TDA can help financial systems to be more secure and reliable. Furthermore, Gidea and Katz (2017) utilised persistent homology to analyse how stock prices move in time, exposing concealed frameworks and associations which cannot be seen through common methods for examining time series. This indicates that the use of TDA can go deeper into understanding the operation of these markets leading to more precise financial predictions and models [15].

A summary of research on OCR and document recognition techniques based on authors, methods used, datasets and performance metrics are compared in Table 1. Each paper uses different methods to handle OCR (Optical Character Recognition) and document analysis problems. For instance, Adams et al. (2017) used persistent homology to analyse 3D point clouds and synthetic data but mentioned problems with noise management and computational speed. Liu et al. (2020) used Transformer-based OCR techniques on SROIE dataset while identifying challenges regarding handling various document formats. Yao et al. (2021) proposed an end-to-end OCR pipeline with contextual embeddings that solved some issues concerning receipt layouts and noisy images. Li et al.,2022 has introduced a structured invoice recognition model which showed high accuracy but had limitations concerning range of recognised documents as well as processing speed. In this regard, the table describes how each study fills existing gaps in the field, such as improved noise handling techniques or adaptable to various document formats or real time processing abilities are preferred. Together with these findings, the table provides detailed information on document recognition research and its developments to

Table 1: Summary of Methods and Gaps in Research

Authors	Method Used	Datasets	Gaps in the Research	Performance Metrics
Adams et al. (2017) [16]	Persistent Homology, Sublevel Set Persistence	3D point clouds, Synthetic data	Necessity for better noise handling and computational efficacy in high-dimensional data.	Average Noise Handling Accuracy = 87.5%
Liu et al. (2020) [17]	OCR with Transformer models	SROIE	Lack of handling diverse formats, incorporation with text extraction systems, management of noisy data, adaptation to multiple languages as well as real-time processing and scalability are some of the gaps in research.	Precision = 82.58%, Recall = 82.03%
Yao et al. (2021) [18]	End-to-End OCR Pipeline with Contextual Embeddings	SROIE	The research did not fully address challenges related to various receipt layouts, handling noisy images, real-time performance.	Accuracy = 90.1%
Li et al. (2022) [19]	Structured invoice recognition based on StrucTexT model with knowledge distillation	SROIE, FUNSD, self-built dataset of VAT invoices	Limited range of documents recognised, possible problems in segregating complex parts, size of the model, and speed of execution are affected.	Accuracy = 94% on SROIE and FUNSD, Accuracy = 95% on self-built dataset

overcome existing challenges. This research seeks to tackle challenges related to processing time and varied dataset formats, while utilising the geometry and shape of features.

3 Methodology

3.1 Overview of Methodology

The problem that is tackled in this research is company classification depending on spatial properties from bounding box coordinates. To understand or distinguish firms, one must analyse their spatial arrangement, and sizes as represented in bounding box data. In order to solve this issue, we will use two different approaches that give different outlooks at data. This research uses two complementary methods to categorise vendor company names based on spatial characteristics obtained from bounding box coordinates, with each method chosen for its capacity to represent diverse dimensions of information. The first approach is concerned with the extraction of geometric features, where feature vectors are constructed from the area and centroid of each bounding box of receipt data. This approach is selected because of its efficiency in understanding important spatial properties like size and central tendency that are generally significant for classification tasks [20]. The second approach relies on Topological Data Analysis (TDA) in which bounding boxes coordinates are converted into point clouds to study their topological traits. This technique describes important topological attributes like connected components and loops, which remain constant through different scales providing more quality information about data's spatial structure [6] [21]. Later this topological information is used to create a high-dimensional classifier that can deal with non-linearities. Owing to this dual methodology we could explore easily the simple geometric features as well as the complicated ones hoping to realise an all-inclusive evaluation of invoice data. By mixing these methodologies we manage to attain a trade-off between simplicity of understanding and capturing in-depth all intricate patterns thus making our model both reliable and efficient at once.

3.2 Geometric Feature Extraction for Vendor Classification- Baseline

This is a developing problem in structured document data analysis, like receipts, where companies have to be classified based on their spatial properties that are derived from bounding box coordinates. This approach extracts geometric features from bounding boxes that capture vendor information from receipts in the form of text. It is in the spatial configuration of these bounding boxes that each company can be effectively represented, differentiating between them by the underlying geometric patterns in the data [22].

3.2.1 Bounding Box Area

Bounding box area is such a significant geometric property because it gives information about the size of spatial region that is enclosed by certain piece of text or image in a receipt or invoice. Since the size is determined by area, we can estimate its length and the number of characters it may contain. The area A of bounding box is calculated using the formula:

$$A = (x_2 - x_0) \times (y_2 - y_0) \tag{1}$$

where x_0, y_0 are the coordinates of the lower-left corner, and x_2, y_2 are the coordinates of the upper-right corner of the bounding box. This area is a critical feature because it directly correlates with the amount of information or prominence a company gives to specific elements on their receipts, which could be a distinguishing characteristic [23].

3.2.2 Bounding Box Centroid

One more noteworthy characteristic that defines the central tendency of space layout is the centroid of the bounding box. The centroid $C(x_c, y_c)$ is calculated as follows:

$$x_c = \frac{x_0 + x_2}{2}, \quad y_c = \frac{y_0 + y_2}{2} \tag{2}$$

The centroid acts as a representation of the overall area of the bounding box and thus is important in visualising how content is arranged around a given receipt. By showing alignment trends and spread which might be unique to some vendors, this property becomes particularly beneficial [24].

The decision about geometric characteristics such as area and centroid are based on an understanding that these qualities are essential in the spatial arrangement of text and images within a document. Area provides an uncomplicated gauge of size, often associated with how much significance or stress a particular vendor attaches to some record. Meanwhile, the centroid denotes spatial equilibrium which may depict the style of design that is done by the company, or it can show the common structure utilised for its layouts. These geometric features are highly suitable for classification tasks due to its simplicity and huge meaning associated with the bounding box. By reducing the bounding box data to a set of feature vectors, we significantly streamline the process of analysing and comparing companies, while still retaining the most critical aspects of the spatial information. This approach, therefore, provides a strong baseline for classification, capable of distinguishing between companies based on the spatial properties of their receipt layouts [25]. Furthermore, geometric features have been shown to be robust against small variations such as shifts in position, changes in size, layout and fonts, allowing for their use in cases where data consistency cannot be ensured for classification tasks.

3.3 Topological Data Analysis for Vendor Classification

This instance of topological data analysis is a technique for classifying vendors using bounding box coordinates, because it captures a complicated relationship between space and time that goes beyond the distinctions of space or form. While the methods based on geometry consider specific points like area and centroid, TDA by persistence diagrams captures more complex underlying structure using the birth and death of topology characteristics that are hidden in data [27]. The approach identifies where there are connected elements in a dataset or closed loops, which can uncover the fundamental framework and trends that characterise vendor receipts. In addition, TDA is an effective method of working with non-linear and high dimensional information; this positions it alongside more conventional geometrical strategies [6].

3.3.1 Homology Groups and Persistence Diagrams

Homology Groups: Homology is a mathematical concept that describes and measures what is expected of an object's shape in terms of counting its 'holes' up to a certain dimension. For instance in Figure 3, a ball has no holes and a doughnut has one hole [8]. This helps in the distinction of shapes, where the emphasis is on these stable characteristics which are less susceptible to noise [9]. Higher-dimensional analogs of this approach are given by homology where the sets of holes are

considered as "homology groups" [21]. The 0th homology group – H_0 describes how many connected components a shape has. For instance, if there is one connected ball then $H_0 = 1$, but if there are two balls that do not touch each other then $H_0 = 2$. The 1st homology group (H_1) coincides with the idea of holes or loops that are present in a given form. For example, one type of doughnut has a single loop and translates into the first homology group, $H_1 = 1$. The 2nd homology group or the H_2 goes a step beyond this idea to recognise holes or hollows in a shape. For instance, a solid spherical object is perfect and has no planes of any 2 dimensional voids and, therefore $H_2 = 0$ while a hallow spherical object has one such perfect plane of 2 dimensional void and thus is $H_2 = 1$ [6]. The bounding box coordinates are converted into the below homology groups and taken as a feature vector.

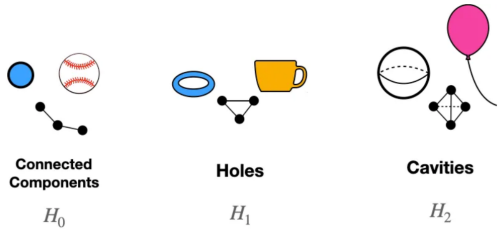

Figure 3: Homology Groups- H_0, H_1, H_2 [28]

Persistence Diagrams: The persistence diagram of TDA represents the birth and death of the features at a different scale of the data. The persistence diagram is obtained from point cloud with bounding box coordinates giving an overview about the type of topological elements, which exist in the dataset at different scales [26]. Homology groups offer crucial assistance in recognising as well as categorising topology data variables. To compute diagrams of persistence, we follow several sequentially executed steps: A point cloud is created, such that each pair of coordinates (x_i, y_i) on the point cloud represents a point. A Vietoris-Rips complex is generated for a given distance threshold $\varepsilon = 0.01$. This involves connecting every two points that lie within this distance to form simplices (triangles, tetrahedra and so on). It is an extension of graphs to higher dimensionality [26].

Using the bounding box array of data, we calculate the homology groups H_0 (which represent connected parts) and H_1 (which represent loops) and H_2 (which represent cavities) for each simplicial complex as ε increases [26]. The birth and death of such types of shapes relying on points are followed through the process [26]. Every characteristic in persistence diagrams has its start times (birth) and end times (death) defined by the values of ε at which it appears and disappears respectively:

$$\text{Persistence of feature } (f) = \text{death} - \text{birth} \tag{3}$$

One can visualise this persistence diagram (as in Figure 4) by plotting every one of such structures (bounding box coordinates in this case) as homology groups, where the first coordinate stands for time of origin and another one indicates termination moment.

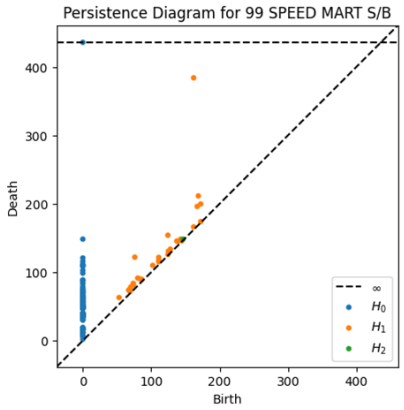

Figure 4: Persistence diagram for 99 Speed Mart S/B Vendor

Figure 4 represents a 2D space for 99 Speed Mart vendor, where many points signify different kinds of topology. The points in x-axis would be taken as birth of the feature, whilst for the points in y-axis it is taken as death of the feature. In this diagram, H_0 (blue points) are the connected components that occur first and are present almost up to the time most of them collide at a point on infinity line. H_1 (orange point) comprise objects that are created and disappeared at intermediate scale, thus, they are moderately stable. H_2 (green) points are connected with temporary and short lived voids, which appear at a later time and have a short duration only. This is demonstrated at the diagram where the features which are distinguishable at other scales are laid bare displaying the topological features of the data. TDA is aptly fitting for this activity due to its ability to encapsulate very own arrangement of data through various levels, a thing that usual geometric practices often overlook. TDA emphasises on sustainability of topological properties therefore making it durable against changes in the format of receipts like receipt layout, such as shifts, rotations, or slight changes in scale which are usually observed from actual data [26].

3.4 Data Collection and Pre-processing

3.4.1 Overview of the Datasets

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 [27]. 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: Intelligent Data Capture and Analysis for Receipts (IDCAR) dataset and rearranged to Kaggle. The samples are in English along with bounding box annotations for text and vendor information. SROIE dataset contains part of IDCAR data focusing only on receipts and invoices. The dataset contains 973 scanned receipts. For each receipt you have .jpg file of the scanned receipt, a .txt file holding OCR information and a .txt file holding the key entity values like vendor, address, total and date [4][27]. 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. Visualisations of the bounding box coordinates indicate the accuracy of the boxes on the receipt images. The SROIE dataset has its special value because of its composition and data diversity. It comprises several key folders for distinct purposes:

Box Folder: It contains text files that provide the bounding box coordinates of each recognised piece of text. Each line presents eight integers in the form $(x_0, y_0, x_1, y_1, x_2, y_2, x_3, y_3)$ specifying consequently the coordinates of the bounding box and then the corresponding text. Such information is crucial to understand how texts are organised spatially on receipts, which is very important for geometric feature extraction or topological data analysis.

Entities Folder: Text files which hold important annotated entities extracted from receipts like company name, date, total amount and address are kept here. These annotations are important for supervised learning tasks where the purpose is to predict or extract certain company entities from scanned documents.

Img Folder: It contains actual receipts images in different formats. These images are used in OCR tasks where image text conversion into machine-readable format is required and visual analysis on these receipts can be done too. This folder is utilised only for mapping the filename with entities and boxes. All these folders are kept constant throughout training as well as testing phases, which makes it easy to access and handle the information.

3.4.2 Data Collection

First, the original Kaggle SROIE dataset v2 shared for public is downloaded from the internet and saved in a personal cloud storage. The dataset is widely popular in the research community and adheres to the general data standards. Therefore, this strategy encourages heterogeneity and complements other scholarly activities. It includes a quantity of images captured at supermarkets, restaurants or any other places where there is stock to sell. To check each receipt along with its proportional axes for the bounding box, all folders were given different identities separately with the help of common file names. In the context of this study, no more image acquisition will be carried out as the existing data provides maximum information.

3.4.3 Data Preprocessing

Data preprocessing is one of the main phases in preparing SROIE dataset for analysis and model training. The ultimate aim at this stage is to clean, transform and organise raw data so that it can be further worked upon. The bounding box coordinates and entity annotations are parsed in this stage of preprocessing. In the Box Folder we have text files which have coordinate points meant for each receipt with associated text therein. Each file consists of these eight integers representing bounding box coordinates – $(x_0, y_0, x_1, y_1, x_2, y_2, x_3, y_3)$ followed by recognised text. The line items and corresponding boxes will vary for different invoice samples. This data is extracted and arranged itself into excel format that is easy to manipulate and access. Equally relevant for supervised learning tasks where models are trained to predict these entities from scanned documents, annotations found within the Entities Folder are used.

After extracting the data, cleaning and normalising the text information proceeds. The removal of superfluous characters like single quotes or extra spaces from all text in both bounding boxes and entity files are performed for this purpose. This step ensures a consistent format of the text, a crucial aspect in ensuring that it can be accurately recognised and analysed at a later stage. More so, normalisation entails converting all texts into lower case letters or eliminating any special symbols which could disrupt their handling during processing. For each bounding box the angles of orientation are computed to detect misalignment. This calculation applies an *arctangent* function to the upper left and right corners of the bounding boxes as per equation (4). The result indicates that all bounding boxes are normalised and uniform

$$\text{angle} = \tan^{-1}\left(\frac{y_2 - y_1}{x_2 - x_1}\right) \times \frac{180}{\pi} \tag{4}$$

Dealing with missing and null values is an additional important step. If there are missing values in bounding box coordinates or entity annotations, such coordinates are identified and removed. If the entries are incorrect, they are flagged or even eliminated. For this reason, the correction or deletion of mismatched coordinates and wrong text takes place to enhance the overall quality of the data set concerned. Six invoice samples are removed through this step which had missing entity values. Lastly, bounding boxes and entity annotations for the receipts are drawn to make some determinations on the correctness of the data. Receipt images have bounding boxes over-laid on them to be visually checked if they align to the text, as the annotations. Considering the training set, out of total 236 companies, only the most populated 20 companies were chosen while for test set the same companies were taken for better comparability. This methodology has been effective as targeting these 20 companies has enhanced the model's performance and precision when compared to using all the classes. Preparation of different categories using the colour codes of the boxes – company names (blue), addresses (red), dates (yellow) and totals (green) helps in easy identification of each set of data as per Figure 2 (middle image). These visual checks are also essential for cross-checking the correctness of the annotations and gaining insight into the spatial distribution of the textual content on receipts, which forms the basis of the data preparation and serves as a guarantee that the subsequent processes of analysis and model training are applicable to the data.

3.4.4 Model Implementation

Baseline Model: In the baseline model implementation phase, ample attention was paid to the process of building features from coordinates of the bounding boxes and using it for classification problems. The process was initiated with the geographical properties of the bounding box including area and centroid coordinate calculation. Particularly, the function called `calculate_area_and_centroid` carries out the computation of area and centroid of each bounding box. Area is the difference between the x and y coordinates of the rectangle in which the bounding box is contained. Centroid is the middle of the diagonal line of the rectangle. These features are necessary for extracting the spatial details of the text regions on the receipts. Then, the `create_sparse_matrix` function is called to create a sparse matrix for these features. Since each of the bounding boxes within the area contributes to this matrix, the columns of this matrix are area, `centroid_x`, and `centroid_y`. The obtained matrix is obviously very sparse, and it is converted to dense matrix for further processing. As for the baseline model, the `process_dataset` function prepares the training and test datasets by obtaining the bounding boxes, calculating geometric characteristics of the cubes defining the objects' position in 3D space, and combining the results into matrix format that can be used for machine learning algo-

rithms: `Logistic Regression, Random Forest, XGBoost, and K- Nearest Neighbors`. Features and Labels are derived from the given data set and matrix is reshaped according to the standard models used in classification process. Indeed, the baseline application implies the use of these matrices for training and testing, with the ultimate goal of setting a performance benchmark.

TDA Approach: In TDA, strategy was to apply summary statistics which included the sum and mean of the persistence diagrams obtained from various homology groups including H_0, H_1 and H_2. But, as it has been observed earlier, this strategy provided a comparatively lower accuracy measure on the test data. Additional experimentation also involved the use of all summary statistics including `max and standard deviation` generated from the persistence diagrams which resulted in some improvement although the competency levels were slightly below the expected levels. The second refinement strategy aimed to reduce the number of features used to the ones from the first and second homology groups only, H_1 and H_2. The zero-dimensional features (H_0) were removed because this kind of features led to many infinity and zero values, which negatively impacted the model's performance. Focusing on H_1 and H_2, which deals with higher-dimensional topological invariants of the data, the accuracy of the used model enhanced. Summary statistics were removed in this case and the homology groups were appended into a single feature vector for applying ML model. This approach showed better results than the baseline and tries to emphasise on the rationality of paying attention to specific topological characteristics for improved predictive results.

In TDA approach, coordinates of the bounding box are converted with the help of the function `bounding_boxes_to_points` into two-dimensional array of points. This step makes the data ready for the topological analysis aimed at calculating persistent homology by transforming the data in a way suitable for this kind of algebraic analysis. The `compute_persistence_features` function adopts the ripser package for the purpose of receiving topological features of these point clouds. The ripser library computes persistent homology with high efficiency with respect to the Rips complex method, which is aimed at revealing the object's multi-scale topology, like connected components, loops, and other higher-dimensional holes. Namely, H_1 and H_2 persistence diagrams are computed, points located close to the diagonal are omitted as per the given threshold ε. This helps to keep only the fundamental characteristics of the topological property being considered. These features are then flattened and concatenated to the feature vectors resulting in the final form of topology of bounding box.

After pre-processing of the topological features, the next stage that transforms the extracted feature vectors of both the training and the testing dataset is normalisation. This is achieved by finding the length of the largest of all the feature vectors and then summing up the vectors and then padding the resultant vectors with zeros to make the length equal to the maximum length. This padding also ensures that each vector has the same dimension as the other vectors, and this is desirable for the entry of the machine learning models that works with the fixed sizes. The matrices arising from the processes above, including the `X_train and X_test` contain transformed topological features, set up in a form that can be immediately used in learning activities.

Model Evaluation: The baseline model indicated better results compared to TDA. The `XGBoost` model performed the best compared to other models across different sets of feature vectors: This means comparison of results on the baseline features, the output obtained using topological data analysis and their combined TDA features. Hence, cross validation using GridSearch was employed to find the best set of hyperparamters. A grid was set up for tuning the hyperparameters used in our model, with regards to the number of trees `(n_estimators)`, learning rate, maximum tree depth `(max_depth)`, subsample, and number of features at each split `(colsample_bytree)`. To sum up, the results also proved that the classifier of `XGBoost` had higher average accuracy compared to Random Forest or Logistic Regression than other classifiers on the feature sets.

The TDA approach alone indicated lower accurate results compared to the baseline, hence a powerful model combining both the feature vectors were developed. These two set of features were then merged to create a complete feature set, the baseline feature which was the `X_train_baseline` was concatenated horizontally with the persistence features set `X_train_df` to form the final training set; `X_train_combined`. The same was done to the test data and a resultant matrix was formed as `X_test_combined`. The resulting model showed an accuracy close to the baseline model indicating that TDA along with geometric features (area, centroid) can be utilised for classification of vendors. `XGBoost` also performed the best when the baseline and topological features were summed up in the combined feature vectors where the model obtained high values for the enrichment of the combined

data. Basic geometric information from the basic shape features and the rich topological information allowed `XGBoost` to encompass all the features of the test set and deliver a high-test accuracy.

4 Results and Discussion

Based on the results from using the traditional geometric features in the baseline model, the baseline method has a higher test accuracy result than the TDA based model. This points to the fact that geometry is important in making a distinction between vendors or in classification problems. It is fair to say that the area and centroid alone had the core elements to identify the vendor. The TDA-based model gives fairly decent results but fails to beat the baseline, thus implying that merely relying on persistence features might not be enough to extract all the information from the dataset. Table 2 below shows the accuracy of different Machine learning algorithms on the Baseline approach with mean accuracy and standard deviation results at various `train_test_split` including the train test folders present in SROIE dataset (given as 'Dataset train/test').

Table 2: Baseline Approach Accuracy Results for Different Train-Test Splits

Model	Dataset train/test	60-40	70-30	80-20	90-10	Mean Accuracy	Standard Deviation
Logistic Regression	0.8385	0.8515	0.8212	0.8713	0.8431	0.8468	1.80%
K-Nearest Neighbors	0.7826	0.7129	0.7417	0.7822	0.7255	0.7406	2.61%
Random Forest	0.9441	0.9109	0.8874	0.9010	0.9020	**0.9003**	0.84%
XGBoost	**0.9503**	0.8812	0.8609	0.8812	0.8824	0.8764	0.90%

To avoid low reliability of results, all the experiments were conducted 4-5 times. All models were executed for five-fold cross validation, whereas the reported accuracy levels' reflect the average and variance of the models. This is because, it enables assessment of the performances of the model in a more holistic manner over the stochastic nature of training. The experiments used stratified `k-fold cross-validation` with `k = 5` to fine-tune hyper-parameters and, at the same time, to balance the number of classes in the folds. This way, the variance is reduced and an accurate generalised performance of the model is obtained, thus reducing the problem of overfitting. In the course of the experiments, various splits were applied such as 60/40, 70/30, 80/20 and 90/10 into training and testing sets which is depicted in Table 2. To enhance the reproducibility, the same random seed was used for all experiments. The models were built in Python and the traditional machine learning algorithms were from the `scikit-learn` package and the gradient boosting methods from the `xgboost` package, while the `ripser` package was used to compute the topological features.

Split 80-20 seems to achieve better results. Out of all the models assessed, `Random Forest` boasts of the highest mean accuracy of 90.03% with a very small standard deviation, which shows good results on splits. `XGBoost` comes second with a mean accuracy of 87.64% and a slightly higher standard deviation accompanied by other important parameters such as precision, recall, F1-score, and support. This is still lower when compared to the train test split that the dataset had which provides high accuracy of about 95% for XGboost and 94% for Random Forest. The classification report for the Baseline approach, utilising XGBoost as the classifier, reveals that "Gardenia Bakeries (KL) Sdn Bhd" vendor achieved the highest prediction accuracy followed by "Unihakka International Sdn Bhd".

Following the baseline approach, various TDA methods were evaluated with a focus on summary statistics such as `average`, `max`, `sum`, and `standard deviation` for the H_0, H_1, and H_2 persistence features. The number of bounding boxes per sample was treated as a hyperparameter during this assessment. The results indicated that increasing the number of bounding boxes per sample led to improvements in both accuracy and performance. However, the summary statistics were found to be less informative, resulting in less accurate models. Consequently, only H_1 and H_2 persistence features were utilised directly instead of relying on summary statistics and H_0. Therefore, full bounding boxes without statistics were used for the samples, and normalisation was applied across the samples, assigning zeros to boxes that were not present. Table 3 presents the model accuracy for different numbers of bounding boxes where summary statistics of persistence diagrams were used as feature vectors for 20-100 bounding boxes initially. It also shows that Full Bounding Boxes without summary statistics yielded the highest performance. This is still not better than the baseline indicating the importance of geometry.

Table 3: TDA approach- Model Accuracy for Different Numbers of Bounding Boxes

Number of Bounding Boxes	Random Forest	Logistic Regression	K-Nearest Neighbors	XGBoost
20 Bounding Boxes	71%	60.87%	61.49%	68.94%
50 Bounding Boxes	76%	70.81%	63.98%	75.16%
100 Bounding Boxes	80%	72.81%	60.87%	78.26%
Full Bounding Boxes (with summary statistics)	80%	72%	61%	78%
Full Bounding Boxes (without summary statistics)	81%	67%	76%	**85%**

The final results comparing the baseline and topology models are shown in Figure 5. The baseline model accurately predicts the vendor as "Gerbang Alaf Restaurants Sdn Bhd," whereas the topology model inaccurately predicts the vendor as "Popular Book Co. (M) Sdn Bhd." In the case of TDA, the persistence features for both receipts appear similar, leading to confusion in the model and resulting in incorrect predictions. Despite the refinements, the TDA model did not surpass the performance of the baseline model. However, incorporating baseline geometric features alongside TDA features improved the model's accuracy to approximately 95%. This result matches the baseline's performance but does not exceed it. This combination underlines the fact that geometric features are strongly supported and complement the topological ones, thus improving the overall model and achieving better results for the task. Out of all models tested in the combined study, the `XGBoost` classifier had the best results with good accuracies. This underlines the fact that `XGBoost` is capable of well combining the complex feature set, both geometric and topological, to provide the best classification results. The good result of the combined model with `XGBoost` indicates that the combination of multiple types of features can give a more comprehensive and therefore better description of the data for classification tasks, which is especially suitable for this kind of research.

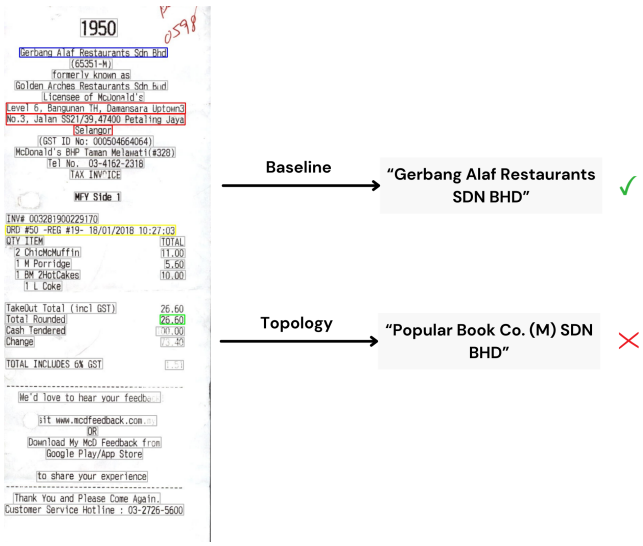

Figure 5: Comparison of Vendor Predictions by Baseline and Topology Models

To summarise, including both geometric and topological persistence features, the H_1 and H_2 diagrams resulted in a performance in par with the baseline model developed based on only geometric features including area and centroid. This model seem to be robut and the running time is just 30-40s for the ML models. This overcomes the time efficiency gap that was mentioned in the literature review. Finally, while integrating topological features induced new benefits it did not improve results any further than simple geometric features. Figure 6 shows the classification report for Baseline approach with `XGBoost` classifier for top 5 most populated vendor classes.

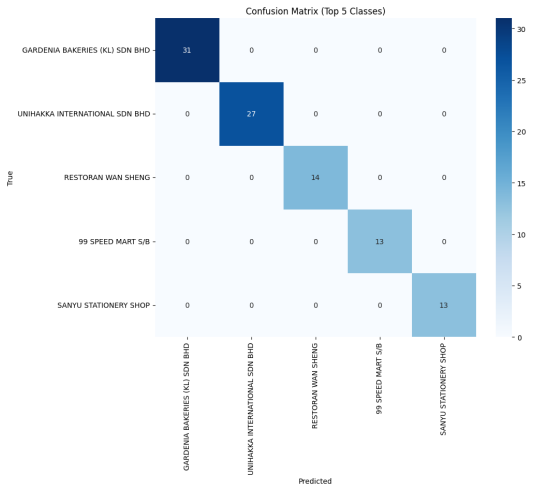

Figure 6: Baseline approach- Classification Report for XGBoost Classifier

5 Conclusion

This study evaluated the classification of vendors using SROIE dataset across three different feature vectors model such as Baseline, TDA and combined approach. The baseline model offered a very good test accuracy indicating that geometric features like area and centroid are inevitable in vendor classification tasks where image is the primary source of data. TDA approach offered a decent accurate result but did not outperform baseline showing that only persistence features are not enough for an accurate vendor classification. There were also few areas where TDA was able to identify the vendors correctly and baseline couldn't. This indicates that topology is also an important characteristic to consider while applying machine learning algorithm. From the persistence diagrams it is clear that the TDA model has the capacity to identify slight topological features that the baseline model was unable to pick. TDA integration made the classifications more accurate, and it helped know the structure of the data a must when identifying vendors.

Some of the shortcomings and limitations found in this study are as follows. A particular problematic aspect is that in case of H_0 features that tend to produce infinity and zero values, the optimal solution drastically dropped. Also, the sample size that was used in this study was reasonable for initial analysis, although increasing the sample size can always be a way to make the results even stronger. Increasing the size and the variety of the data might also shed more light on the aspects of the vendors' classification that could demonstrate the full potential of TDA. There were multiple classes, and the distribution of classes is unbalanced. Moreover, by having similarly named classes it created challenges that made the classification process even more difficult. These limitations point out some directions for further study, for example, the use of extended dataset, more balanced classes and the better ways to handle the issues arising with the similar class names. Further, the limitations can be overcome by performing data augmentation and class imbalance correction using SMOTE technique. This will reduce the overfitting and maintain the overall variance of the data which is essential in a multi- classification problem. Additionally, deep learning algorithms like CNN, RNN could also be applied to enhance the model's performance. Furthermore, the computational complexity of the TDA approach turned out to be low where it took 4 hours to compute the features, while the accuracy was decent indicating the possibility of improving the choice of topological characteristics or the development of new algorithms based on them.

The study result points to the possibility of TDA for the invoice and accounts payable field with a decent performance. This paves way of creating new techniques for higher and broader TDA application, and the addition of other analytical approaches that would improve the algorithm current output. Further work could be to extended to try TDA with deep learning methods to capture more complex details, to play with the persistence landscapes, and to expand the use of TDA to areas such as image, 3D object detection, and text analysis. Overall, the research offers valuable insights into the potential of TDA in data science, suggesting that its broader application could drive significant advancements across various sectors.

References

[1] Chazal, F., & Michel, B. (2017). *An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists*. Springer.

[2] Berke Oral, E., Emekligil, E., Arslan, S., & Eryiğit, G. (2023). Information Extraction from Text Intensive and Visually Rich Banking Documents. *Information Processing & Management*.

[3] Wang, X., Wu, Y., & Zhang, W. (2019). Robust OCR for Complex Document Layouts: A Survey. *IEEE Access*, 7, 171377-171388.

[4] Zheng Huang, K., Chen, J., He, J., Bai, X., Karatzas, D., Lu, S., & Jawahar, C.V. (2019). ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction.

[5] Zhuang, Y., Fang, X., & Wang, X. (2017). Challenges and Opportunities: From Big Data to Knowledge in Financial Services. *IEEE Transactions on Knowledge and Data Engineering*, 29(10), 2216-2231.

[6] Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2), 255–308.

[7] Chazal, F., Cohen-Steiner, D., & Merigot, Q. (2010). Boundary measures for geometric inference. *Foundations of Computational Mathematics*, 10(3), 221–240.

[8] Edelsbrunner, H., Letscher, D., & Zomorodian, A. (2002). Topological Persistence and Simplification. *Discrete & Computational Geometry*, 28(4), 511–533.

[9] Zomorodian, A., & Carlsson, G. (2005). Computing persistent homology. *Discrete & Computational Geometry*, 33(2), 249–274.

[10] Oudot, S. (2015). *Persistence Theory: From Quiver Representations to Data Analysis*. American Mathematical Society.

[11] Nicolau, M., Levine, A. J., & Carlsson, G. (2011). Topology-based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17), 7265-7270.

[12] Kramar, M., et al. (2013). Persistent homology and machine learning for the analysis of material structures. *Journal of Computational Physics*, 242, 531-550.

[13] Skraba, P., & Bujnak, M. (2010). Shape segmentation using persistent homology. *Computer Vision and Image Understanding*, 114(10), 1061-1072.

[14] Mitra, S., & Rao, K. V. (2021). Experiments on Fraud Detection use case with QML and TDA Mapper. In *2021 IEEE International Conference on Quantum Computing and Engineering (QCE)*, Broomfield, CO, USA (pp. 471-472). IEEE. doi: 10.1109/QCE52317.2021.00083

[15] Gidea, M., & Katz, Y. (2018). Topological data analysis of financial time series: Landscapes of crashes. *Physical Review E*, 98(5), 052221.

[16] Adams, H., et al. (2017). Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(1), 218-252.

[17] Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., & Wei, F. (2023). TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11), 13094-13102. https://doi.org/10.1609/aaai.v37i11.26538

[18] Yao, Z., Li, X., Zhang, H., & Wang, J. (2021). End-to-End OCR Pipeline with Contextual Embeddings for Receipt Analysis. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

[19] Li, Z., Tian, W., Li, C., Li, Y., & Shi, H. (2023). A Structured Recognition Method for Invoices Based on StrucTexT Model. *Applied Sciences*, 13(12), 6946. https://doi.org/10.3390/app13126946

[20] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.

[21] Edelsbrunner, H., & Harer, J. (2010). *Computational Topology: An Introduction*. American Mathematical Society.

[22] He, L., Li, J., Liu, C., & Li, S. (2018). Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines. *IEEE Transactions on Geoscience and Remote Sensing*, 56(3), 1579-1597. doi: 10.1109/TGRS.2017.2765364

[23] Farin, G. (2002). *Curves and Surfaces for CAGD: A Practical Guide* (5th ed.). Morgan Kaufmann.

[24] Dolezel, P., Skrabanek, P., Stursa, D., Baruque Zanon, B., Cogollos Adrian, H., & Kryda, P. (2022). Centroid based person detection using pixelwise prediction of the position. *Journal of Computational Science*, 63, 101760. ISSN 1877-7503. https://doi.org/10.1016/j.jocs.2022.101760

[25] Pavlidis, T. (1977). *Structural Pattern Recognition*. Springer-Verlag.

[26] Barnes, D., Polanco, L., & Perea, J. A. (2021). A comparative study of machine learning methods for persistence diagrams. *Frontiers in Artificial Intelligence*, 4 Article 681174. https://doi.org/10.3389/frai.2021.681174

[27] SROIE. (2023). *SROIE v2 dataset*. Retrieved from `https://www.kaggle.com/datasets/urbikn/sroie-datasetv2`. Introduced by Huang et al. in ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction.

[28] M. Mehta, "Persistent Homology," DataDrivenInvestor, Jan. 13, 2024. [Online]. Available: https://medium.datadriveninvestor.com/persistent-homology-f22789d753c4