Towards a Novel Classification of Table Types in Scholarly Publications

Jilin He[1], Ekaterina Borisova*[2], and Georg Rehm[2]

[1] Technische Universität Berlin (TU), Berlin, Germany
[2] Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin, Germany
jilin.he@campus.tu-berlin.de, ekaterina.borisova@dfki.de,
georg.rehm@dfki.de

Abstract. Tables are one of the prevalent means of organising and representing structured data. They contain a wealth of valuable information that is challenging to extract automatically, yet can be leveraged for downstream tasks such as question answering and knowledge base construction. Table Type Classification (TTC) is one of the tasks which contributes to better semantic understanding and extraction of knowledge in tabular data. While multiple classification schemas exist, almost all of them are focused on web tables. Therefore, these classifications might overlook certain types which are common in other areas such as scientific research. This paper addresses this gap by introducing ten novel TTC taxonomies tailored towards tables used in scholarly publications. We also evaluate the applicability of taxonomies derived from web tables to scientific tables. Additionally, we propose a new dataset containing 13,000 annotated table images, called TD4CLTabs. Our results indicate that both existing and newly proposed taxonomies are suitable and effective for classifying scientific tables.

Keywords: Table type classification · taxonomy construction · table understanding.

1 Introduction

Tables are used to summarise and present information in a structured manner across various areas such as business, finance, science, education, and healthcare [40]. With a growing interest in the field of Table Understanding (TU), several studies have focused on the automatic extraction of knowledge from tables [16,3,45,36] and applying it to various tasks, e. g., question answering [43,50,33,20,48,5,7,22,29,9], knowledge base construction [27,25], table-to-text generation [28], tabular data augmentation [45,44,12], content extension and completion [27,21], fact-checking [6,1], and natural language inference [17].

Table Type Classification (TTC) is the TU sub-task aimed to categorise tables according to a predefined schema based on their layout structure, content or purpose of use [45]. Classifying tables into specific types helps to uncover

* Corresponding author

the semantics of the data they contain, facilitating tasks such as detecting and filtering layout tables (which do not contain any meaningful data), recognising table structures, and information extraction [15,14,25,23]. Even though various TTC schemas exist [11,26,41,4,27,25,8], most were designed focusing on tabular structures that exist in web pages, commonly referred to as *web tables* [26]. As a consequence, these classifications might overlook certain table features and types, especially domain specific ones. In particular, they might not be fully applicable to tables found in scholarly papers. We refer to such tables as *scientific tables*, defining them as tabular structures found in (digital) scholarly publications and labelled as a table by the authors. To the best of our knowledge, there is only one study by Kruit et al. [25] that proposed a table type taxonomy derived from scientific tables. No taxonomies based on structural or layout features exist for the field of scientific publications. The present paper addresses this gap by developing ten novel taxonomies based on scientific tables. To this end, we collect a corpus of tables extracted from Computational Linguistics (CL) articles. We develop various taxonomies based on two well-established classification schemas and by considering table features identified in previous studies and our own corpus analysis. We train and evaluate classifiers on the dataset of scientific tables that we annotated according to the two pre-existing schemas and our newly proposed taxonomies.

Our contributions can be summarised as follows:

- We construct and release the TD4CLTabs dataset with 13,000 annotated images of scientific tables extracted from CL articles.
- We propose and evaluate ten novel TTC taxonomies defined based on scientific tables.
- We assess the applicability of taxonomies derived from web tables to scientific tables.
- We offer a list of table features which are potentially important for TTC. The list includes attributes considered by previous taxonomies, alongside those overlooked by these schemas but identified in the literature and in our TD4CLTabs dataset.

This article is structured as follows: Section 2 discusses related work. Section 3 describes our approach to the dataset and taxonomies construction. Sections 4 and 5 present the evaluation results and main findings, respectively. Section 6 outlines limitations. Concluding remarks are provided in Section 7.

2 Related Work

Tables are ubiquitous data structures, often stored in relational databases (e. g., MySQL, PostgreSQL), spreadsheets (e. g., Microsoft Excel, Google Sheets), web pages (e. g., Wikipedia), and scientific articles. Tables vary greatly in terms of their layout structures and content, posing challenges for automatic TU [2,46]. In order to effectively process and extract knowledge from tables, several TTC schemas have been proposed.

The existing schemas vary in their complexity, ranging from simple binary classifications to multi-layer taxonomies. Additionally, most TTC schemas have been designed based on tables found in web pages. For instance, in the pioneering work by Wang and Hu [42], web tables were classified into two categories: *genuine*, i. e., leaf tables (not containing other tables, lists, images, etc.) and *non-genuine*. Later Cafarella et al. [4] distinguished between *extremely small* tables, *HTML forms*, *calendars*, *non-relational* (contain low-quality data), and *relational* (contain high-quality data) tables. Subsequent studies proposed more fine-grained classifications by organising table types into hierarchical taxonomies. Crestan et al. [11] introduced the categories of *relational knowledge* tables, which contain relational data, and *layout* tables, which do not contain any meaningful data at all. The former class included sub-types defined based on the positioning of table headers: *vertical listing*, *horizontal listing*, *matrix*, *attribute/value*, *enumeration*, and *calendar*. The layout category contained *formatting* and *navigational* tables. Lautert et al. [26] refined this taxonomy by revisiting the relational knowledge tables class and incorporating types derived from cell features. On the first layer, relational knowledge tables were categorised as *horizontal*, *vertical*, and *matrix*. These were subsequently divided into *concise* (contain merged cells), *nested* (contain a table in a cell), *splitted* (contain repeated labels in headers), *simple* and *composed multivalued* (contain multiple values in a single cell) categories. Chen and Cafarella [8] devised an alternative TTC taxonomy focusing on the use-case of web spreadsheets. In contrast to previous studies, this taxonomy incorporates major classes such as *data frame* spreadsheets and *non-data frame* (flat) spreadsheets, along with their respective sub-categories. More recent studies have shifted back to single-level classification schemas. Eberius et al. [14] distinguished between three main table types, namely *matrix*, *horizontal listing*, and *vertical listing* (see Appendix A). Similarly, Lehmberg et al. [27] also classified tables into three major categories: *relational*, *entity*, and *matrix*.

In contrast to web tables, there is currently only one TTC taxonomy defined based on scientific tables extracted from Computer Science papers. It was proposed by Kruit et al. [25] for the development of Tab2Know, i. e., a novel end-to-end system for building a knowledge base from scientific tables. This taxonomy consists of four root classes (*observation*, *example*, *input*, *other*) with their respective sub-classes and primarily focuses on the narrative role tables play in scholarly articles rather than their structural characteristics.

As emphasised by Zhang and Balog [45], the established approaches to TTC were designed for different use-cases. Therefore, it is not surprising that existing schemas might overlook certain table features. For instance, Shigarov et al. [39,38] highlighted that current classifications fail to address header and cell-related characteristics such as header hierarchies, the presence of non-textual content and diagonally split cells. Additionally, the schemas do not consider the concepts of complicated tables (i. e., containing spanning cells) and void cells introduced by Chi et al. [10] and Rolan et al. [35], respectively (see Appendix B).

In earlier studies, TTC relied on traditional machine learning algorithms such as decision trees, support vector machines, and logistic regression [42,4,11,26,14,25].

Recent research has shifted towards the adoption of deep learning techniques such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms for automatic feature extraction from tables [31,18]. Previous approaches primarily utilised plain-text and HTML representations of tables. However, not all tables are readily accessible in a machine-readable format. For instance, scientific tables are commonly embedded in unstructured PDF documents. Such tables have to be extracted and transformed into a format suitable for training and testing models. One of the widely used approaches involves obtaining the image-like representations of tables from a PDF file [25,24,49] which can either be directly used as model input or first converted into structured formats like CSV or JSON.

3 Methodology

3.1 Data

To assess the applicability of web tables-based taxonomies to the area of science and to construct novel TTC taxonomies, we created a corpus of table images from scholarly articles in the ACL Anthology.[3] We fetched a total of 3,219 papers from the year 2022, chosen as the latest collection of publications in the readily available ACL Anthology corpus.[4] As ACL papers are available only in PDF, Tab2Know was used to obtain table images. Out of the 3,219 PDF files, Tab2Know successfully processed 2,687, resulting in a total of 15,292 table images. Since Tab2Know is designed to locate and extract tables without their respective captions and titles, these are not present in our corpus.

3.2 Taxonomies Construction

We applied two established schemas based on web tables to the corpus of scientific tables, i. e., the classifications proposed by Eberius et al. [14] and Crestan et al. [11]. We picked these two taxonomies based on their usage in recent applications and tasks. We did not consider the taxonomy proposed by Kruit et al. [25] since it classifies tables based on their narrative role in scientific articles rather than their layout structure.

In order to determine whether any adjustments are needed in the two taxonomies, such as excluding under-represented classes, we examined their presence and distribution in a sample of 1200 table images from our corpus. The results are presented in Figure 1. Eberius et al.'s schema, featuring the classes *listing* and *matrix*, was directly adopted to the TTC task due to their high frequency in the corpus. The taxonomy by Crestan et al. was adjusted by keeping *horizontal listing*, *vertical listing*, *matrix*, and *enumeration*, while disregarding other classes (e. g., calendar, form, layout tables, etc.) since these could not be observed in

[3] https://aclanthology.org
[4] https://github.com/shauryr/ACL-anthology-corpus

the sample data. Additionally, all tables of the attribute/value class were classified as either vertical listing or horizontal listing since they represent specific instances of these classes [11]. Together with the class *other tables*, which was introduced for tables that do not fit any of the pre-defined classes, we refer to the final two taxonomies as Baseline_I and Baseline_II, respectively. The graphical illustration of the baseline taxonomies is provided in Figure 2 (a).

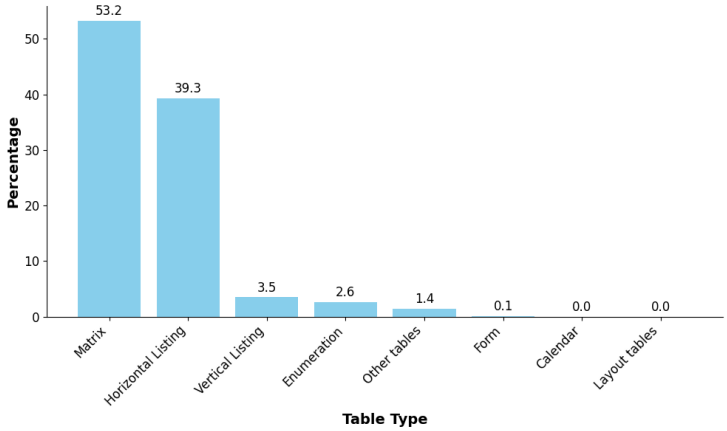


Fig. 1. The distribution of table types defined by Crestan et al. [11] and Eberius et al. [14] in a sample of 1200 table images extracted from the ACL Anthology Corpus.

In addition, ten novel taxonomies were defined by incorporating the table types from the baseline taxonomies as well as header and cell features. As a fist step, we determined which classes should be preserved from Baseline_I and Baseline_II by analysing the results of their preliminary frequency of occurrence (Figure 1). Hence, only the matrix and horizontal listing classes were considered while designing the taxonomies. Vertical listing and enumeration were disregarded due to their low frequencies in the dataset. Then, we compiled a list of table layout features which are neglected by the existing taxonomies but distinguished by previous studies (see Section 2). We further extended the list with additional features observed during the examination of the 1200 sample tables. The collected features fall into header and other table attributes and are outlined in Table 1.

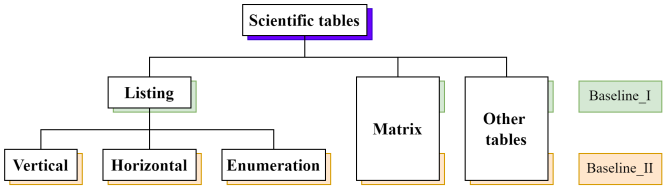
Initially, we constructed the TTC taxonomies by combining the selected table types and additional header features. We refer to these as Header-Feature Table Taxonomies (HFTTs) and present them in Figures 2 (b) and (c). Thus, taking into account the absence or presence of a header hierarchy, we extended Baseline_I with the classes *flat listing*, *flat matrix*, *hierarchical listing*, and *hierarchical matrix* classes, and called it HFTT_Novel_I. Then, we incorporated the positioning of hierarchical headers (HHs) within the classes matrix and hori-

Table 1. Header and other features potentially significant for Table Type Classification. Attributes identified based on a sample of 1200 tables extracted from ACL papers are highlighted in italics.

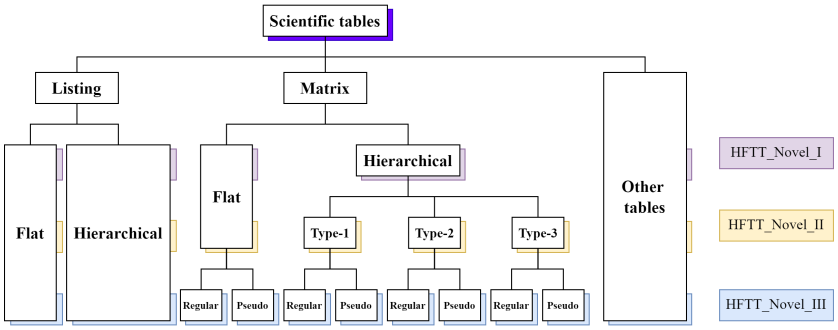
Header Features
– Positioning of headers [11,26]
– Hierarchy of headers [39]
– *Varied positioning of hierarchical headers in Matrix*
– Presence of diagonally split cells in Matrix [38]
Other Features
– Presence of missing and void cells [35]
– Presence of non-textual content [38]
– *Presence of hierarchical rows*
– Presence of spanning cells [26,10]
– *Presence of other complex cells*
– Table splitting [26]

zontal listing into HFTT_Novel_I. For the former, HH might exclusively appear in a column header (CH), row header (RH), or in both. We refer to these three additional classes as *type-1*, *type-2*, *type-3 hierarchical matrix*. In the case of horizontal listing, HH may be positioned on the left, right or middle of a table, potentially with repetitions. We name the resulting taxonomy HFTT_Novel_II. As can be seen from Figure 2 (b), for HFTT_Novel_III, we further distinguished between matrix with diagonally split cells at the top-left cell (*pseudo matrix*) and without those (*regular matrix*). Note that pseudo matrices often bear a resemblance to listing. For the final HFTT_Novel_IV, we excluded HH and the three respective HH positioning types related to matrix and pseudo matrix. Eventually, the ten different taxonomies developed vary in terms of their number of classes, from 3 to 17. Baseline_I contains the fewest number of categories, while FFTT_Novel_V includes the highest number.

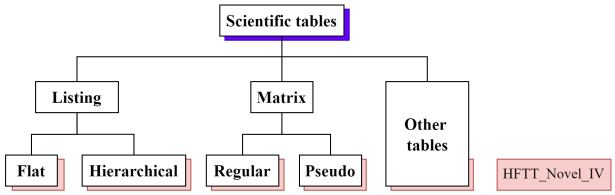
As outlined in Table 1, HFTT can be extended with other table features related to cell types and table splitting. Thus, each feature introduces a new category within each table type across HFTTs. When focusing solely on header features, the resulting table types are mutually exclusive. For instance, if a table is categorized as matrix, it cannot simultaneously belong to the listing class. Similarly, once it falls into the type-1 hierarchical matrix, it cannot be classified as type-2, type-3 or pseudo matrix. However, when considering both header and other table features, the resulting table types become inclusive. Thus, matrix can exhibit features such as spanning cells and being split at the same time, leading to a new category called *split complex matrix*. We refer to the refined HFTTs, containing header features, cell-related attributes, and table splitting, as Full-Feature Table Taxonomies (FFTTs). Figure 3 shows two examples.



(a)



(b)



(c)

Fig. 2. The table type taxonomies proposed in this study: Figure (a) depicts two baseline taxonomies, while (b) and (c) illustrate four newly defined taxonomies. The colours highlight each taxonomy and its respective classes.

Method	BC5CDR (Big Dict)	BC5CDR (Small Dict)
Fully Supervised	85.00 (77.68/93.83)	
Binary PU	72.38 (61.89/87.17)	79.14 (81.85/76.60)
Method	**CoNLL2003 (KB)**	**CoNLL2003 (Dict)**
Fully Supervised	96.09 (92.77/99.65)	
Binary PU	88.88 (81.04/98.39)	88.08 (79.09/99.38)

(a) Split Complex Matrix

Dataset	Source(s)	Target	Context	Evidence	#Instances	Task
English Datasets						
Rumour Has It (Qazvinian et al., 2011)		Topic	Tweet		10K	Rumours
PHEME (Zubiaga et al., 2016b)		Claim	Tweet		4.5K	Rumours
Emergent (Ferreira and Vlachos, 2016)		Headline	Article*		2.6K	Rumours
FNC-1 (Pomerleau and Rao, 2017)		Headline	Article		75K	Fake news
RumourEval '17 (Derczynski et al., 2017)		Implicit[1]	Tweet		7.1K	Rumours
FEVER (Thorne et al., 2018)		Claim	Facts		185K	Fact-checking
Snopes (Hanselowski et al., 2019)	Snopes	Claim	Snippets		19.5K	Fact-checking
RumourEval '19 (Gorrell et al., 2019)		Implicit[1]	Post		8.5K	Rumours
COVIDLies (Hossain et al., 2020)		Claim	Tweet		6.8K	Misconceptions
TabFact (Chen et al., 2020)		Statement	WikiTable		118K	Fact-checking
Non-English Datasets						
Arabic FC (Baly et al., 2018b)		Claim	Document		3K	Fact-checking
DAST (Danish) (Lillie et al., 2019)		Submission	Comment		3K	Rumour
Croatian (Bošnjak and Karan, 2019)		Title	Comment		0.9K	Claim verifiability
ANS (Arabic) (Khouja, 2020)		Claim	Title		3.8K	Claim verification
Arat(bic)Stance (Alhindi et al., 2021)		Claim	Title		4K	Claim verification

(b) Listing with the presence of hierarchical rows and non-textual contents

Fig. 3. Examples of scientific tables belonging to the Full-Feature Table Taxonomies.

3.3 Annotation

To label the corpus of 15,292 table images according to the defined taxonomies, we run an annotation project. LabelStudio[5] was used as the annotation tool and since there was only one annotator involved, a Master student of Data Science, no inter-annotator agreement (IAA) score was calculated. To ensure that the final corpus contains well-structured images, displaying only the complete and clear layout of tables, we filtered out inappropriate samples while annotating. To this end, we introduced the class *non-table* and used the following rules during the annotation:

- If a table is partially extracted, as if incorrectly cropped, it is not considered to be a complete table and should be annotated as non-table.
- If a table is fully extracted but labelled as Figure in a paper, it should be annotated as non-table.
- If a table is fully extracted but there is other information in the image, such as segments of text, it should be annotated as non-table.
- If a table is fully extracted but an image contains multiple scattered tables, it is considered as incorrect input and should be annotated as non-table.

As a result, 280 table images belong to the non-table category and were excluded from the corpus. We also checked the labelled data with respect to annotation errors. Consequently, 54 images were removed from the corpus.

The final dataset comprises 13,301 annotated scientific table images along with their respective metadata (image name, image label, image path, and dataset split). We refer to the final corpus as TD4CLTabs (Type Detection for Computational Linguistics Tables) dataset.[6] As a post-processing step, we encoded the categorical features with numerical values. Then we divided the dataset into a training set containing 10,347 table images and a test set comprising 2,954 samples.

3.4 Models

Considering recent advances of deep learning in computer vision (CV), alongside the proven successful application of table images for TU tasks such as table detection and table structure recognition [34,49,30,32,37], we approach TTC as an image classification task. In particular, TTC based on HFTTs was tackeled as a multi-class problem, while classification based on FFTTs was addressed as a multi-label task.

Two models, ResNet50 [19] and Vision Transformer (Vit) [13], were trained.[7] ResNet50 is a deep CNN model widely utilised in CV tasks, exhibiting efficient performance in image classification problems. ViT presents a newer approach to CV, utilising the Transformer architecture's unique ability to capture global

[5] https://labelstud.io
[6] https://zenodo.org/records/10972922
[7] The code is available on Software Heritage: https://shorturl.at/mCGHT

image information, outperforming traditional CNN models. We combined pre-encoded labels from all hierarchy levels into one flat list and fed them as input into the models along with table images.

ResNet50 was implemented using the Fastai framework.[8] For the Vit model, we utilised the Hugging Face implementation.[9] To enhance the robustness and reliability of the image classification models, cross-validation was applied with k set to 4. For both models, the batch size was set to 16. The resize dimensions of (500, 900) and (224, 224) were chosen for ResNet50 and Vit, respectively. FocalLoss was employed as the loss function for ResNet50, while the default CrossEntropy was used for Vit. The training process for ResNet50 extended to 30 epochs with early stopping enabled and a patience of 5 epochs. Vit was trained for 15 epochs with the option to save the best model. Both models utilised pretrained weights, with ResNet50 set to True and Vit using the 'google/vit-base-patch16-224-in21k' pretrained configuration.

3.5 Evaluation Metrics

To evaluate the performance of the two models on the multi-class classification task, error rate, precision (weighted), recall (weighted), and F1 score (weighted) were used. In the case of multi-label classification, hamming loss, macro and micro F1 scores were utilised.

4 Results

4.1 Dataset Analysis

The table images in our dataset have a wide range of resolutions, spanning from a minimum of 100×100 pixels to a maximum of either 1200×200 or 1000×1400 pixels. In terms of dimensions, tables average 7.60 rows and 6.68 columns.

The distribution of tables per class within each HFTT is presented in Figure 4. As can be seen, with the increase in the number of classes, the degree of data imbalance also rises. The analysis shows that matrix tables are approximately 15% more common than listings in the dataset. Interestingly, other tables comprise less than 5%. Among the matrix tables, those with HHs constitute approximately half of all (49%). Furthermore, the majority of such tables (about 64%) fall under type-1 hierarchical matrix, i. e., have HHs located in a CH. Matrix tables with diagonally split cells are quite frequent (about 71%). The least common across the matrix sub-categories are type-2 hierarchical and type-3 hierarchical. In terms of the listing class, horizontal tables are more frequent (about 84% of the total) than vertical and enumeration types. In contrast to hierarchical matrix tables, the number of hierarchical listings in the dataset is considerably lower (approx. 8% of all listings).

[8] https://www.fast.ai
[9] https://huggingface.co

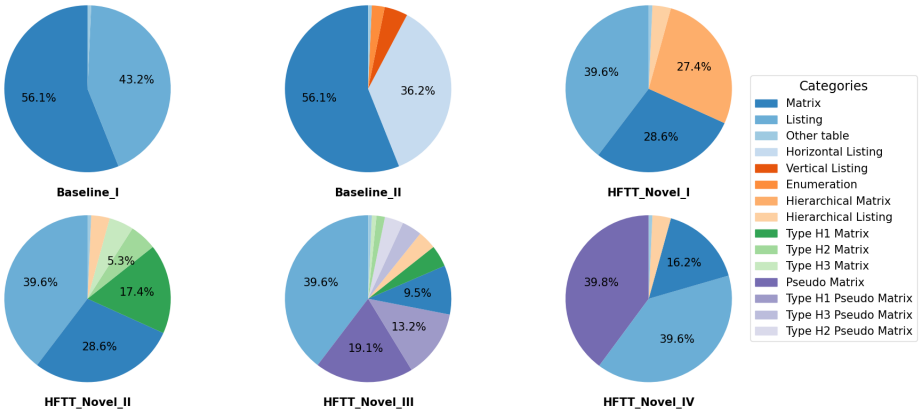


Fig. 4. The distribution of table types in the baseline and Header-Feature Table Taxonomies within the TD4CLTabs dataset. Note that only proportions exceeding 5% are explicitly labelled with numerical values.

Figure 5 illustrates the distribution of table splitting and cell-related features incorporated into FFTTs within the TD4CLTabs dataset. The results indicate the infrequent occurrence of those across the given corpus of scientific tables. The highest value of about 13% was achieved for the missing and void cells type, followed by the presence of hierarchical rows (approximately 10%). A limited number of tables contain cells with non-textual content (about 3%) and other complex cells (about 2%).

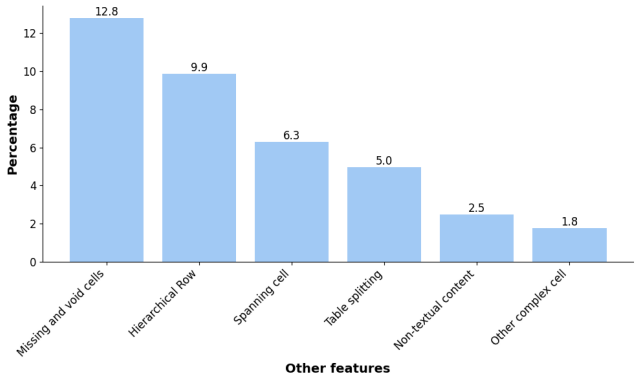


Fig. 5. The distribution of cell types and table splitting across the TD4CLTabs dataset

4.2 Table Type Classification

Table 2 presents the TTC results across HFTTs. The Vit model outperforms ResNet50 in all but one case, namely HFTT_Novel_II. We can also see a general trend of decreasing performance among the models as the number of classes in the taxonomy increases. The class imbalance indicated in Section 4.1 might have also influenced the predictions. The best F1 value (0.82) was obtained for Vit based on Baseline_I. This is not surprising since it is a 1-level schema with the least number of classes and the most balanced data. The second highest F1 scores (0.78) were achieved by Baseline_II and HFTT_Novel_IV, both of which contain two additional categories when compared to Baseline_I. Even though HFTT_Novel_III contains four more categories than HFTT_Novel_II, the models based on these taxonomies result in very similar results (approx. 1% difference). The study also shows that HFTT_Novel_IV achieved the highest scores among the novel taxonomies.

Table 2. Multi-class classification results based on baseline and Header-Feature Table Taxonomies

Taxonomy	**ResNet50**				**Vit**			
	Error Rate	Precision	Recall	F1	Error Rate	Precision	Recall	F1
Baseline_I	0.22	0.78	0.78	**0.77**	0.17	0.82	0.83	**0.82**
Baseline_II	0.23	0.76	0.77	**0.76**	0.21	0.77	0.79	**0.78**
HFTT_Novel_I	0.26	0.74	0.74	**0.73**	0.23	0.74	0.77	**0.75**
HFTT_Novel_II	0.27	0.73	0.73	**0.72**	0.26	0.72	0.74	**0.71**
HFTT_Novel_III	0.28	0.73	0.72	**0.71**	0.27	0.72	0.73	**0.72**
HFTT_Novel_IV	0.25	0.76	0.75	**0.75**	0.21	0.78	0.79	**0.78**

The results for multi-label classification based on FFTTs are provided in Table 3. In terms of micro F1, the Vit model demonstrates overall better performance compared to ResNet50 across all taxonomies, except FFTT_Novel_IV and FFTT_Novel_V. However, all models exhibit low macro F1 scores, indicating the dataset imbalance. The hamming loss values are also consistently low across the models (0.05-0.07), suggesting an overall good performance of the classifiers. Similar to the classification based on HFTTs, we note a trend where models tend to perform worse on FFTTs with a larger number of classes. Furthermore, the highest score (0.75) for FFTTs is about 7% and 2% lower compared to those obtained for the baselines and HFTTs, respectively.

To address the problem of class imbalance, we applied the random oversampling technique [47] on novel HFTTs.[10] This involved duplicating instances of the minority classes to align with the majority classes. As shown in Table 4, oversampling consistently improved F1 scores by 1-5% across the models. The

[10] Note that we have not addressed the data imbalance for FFTTs.

Table 3. Multi-label classification results based on Full-Feature Table Taxonomies. The threshold is set to 0.5. If the probability of the prediction is greater than 0.5, it as a positive prediction. Otherwise, it is a negative prediction.

Taxonomy	**ResNet50**			**Vit**		
	$F1_{Micro}$	$F1_{Macro}$	Hamming Loss	$F1_{Micro}$	$F1_{Macro}$	Hamming Loss
FFTT_Novel_I	**0.73**	0.54	0.07	**0.75**	0.38	0.07
FFTT_Novel_II	**0.69**	0.49	0.06	**0.72**	0.32	0.06
FFTT_Novel_III	**0.70**	0.55	0.07	**0.71**	0.37	0.07
FFTT_Novel_IV	**0.69**	0.58	0.06	**0.68**	0.36	0.06
FFTT_Novel_V	**0.66**	0.53	0.05	**0.61**	0.25	0.05
FFTT_Novel_VI	**0.70**	0.54	0.07	**0.72**	0.47	0.06

Vit model based on HFTT_Novel_IV is the only instance where a slight decrease in score (by about 2%) is observed. All other evaluation scores also increased in the majority of HFTT classifiers. Furthermore, comparable results to ResNet50 with Baseline_I were achieved on ResNet50 with HFTT_Novel_I and HFTT_Novel_IV. However, despite the overall improvement in model performance, the prediction accuracy for novel taxonomies still remains lower (by approximately 5%) than that of Baseline_I based on Vit.

Table 4. Multi-class classification results based Header-Feature Table Taxonomies after applying oversampling

Taxonomy	**ResNet50**				**Vit**			
	Error Rate	Precision	Recall	F1	Error Rate	Precision	Recall	F1
HFTT_Novel_I	0.22	0.78	0.78	**0.77**	0.24	0.77	0.76	**0.76**
HFTT_Novel_II	0.25	0.75	0.75	**0.74**	0.24	0.76	0.76	**0.76**
HFTT_Novel_III	0.24	0.77	0.76	**0.75**	0.24	0.77	0.77	**0.75**
HFTT_Novel_IV	0.22	0.78	0.78	**0.77**	0.24	0.78	0.76	**0.76**

5 Discussion

The study indicates that matrix and listing tables are the most commonly used across CL papers. In particular, matrix with hierarchical headers, frequently found in CHs, matrix with diagonally split cells, and horizontal listings are prevalent. Hence, these types are worth considering when classifying scientific tables. In contrast, the findings suggest that incorporating table splitting and cell features may not be advantageous, as they seem to be relatively uncommon in scientific tables.

The study further showcased the applicability of the TTC schema by Eberius et al. to scientific tables. In this sense, Crestan's et al. taxonomy also proved to

be adaptable after smaller adjustments. The models based on these baseline schemas demonstrate greater efficiency on TTC than those trained on the newly proposed taxonomies. Hence, although the two established classification schemas were designed for web tables, they are still suitable for scientific tables.

While the experimental results do not demonstrate a clear advantage of the novel domain-specific taxonomies, they do show the promising outcomes. Among the newly developed taxonomies, HFTT_Novel_I and HFTT_Novel_VI have proven to be the most successful. This could potentially be attributed to the smaller number of categories within those, indicating a lower level of complexity, compared to other schemas. These taxonomies also achieved efficiency comparable to the results obtained for ResNet50 with the baseline schemas.

6 Limitations

While this study sheds light on devising TTC taxonomies for scientific tables, it is not without limitations. First, the annotations may be subjective and contain errors due to the involvement of only one annotator. Having at least one additional annotator and curator, and subsequently validating the results by calculating the IAA score, would be beneficial. Second, the novel taxonomies were constructed and tested based on scientific tables from CL papers. Thus, the applicability of those to other domains remains an open research question, which we leave for future work. Third, the study considered only two existing web table based taxonomies, limiting the analysis to types within them and potentially neglecting other categories relevant to scientific tables. Finally, the hierarchy of the taxonomies' labels was not taken into account in this study. Additionally, to tackle class imbalance, we considered only oversampling and applied it only to taxonomies with header features. Future endeavours could incorporate the label hierarchy in the model training process and focus on annotating more samples for the minority classes or on utilising other automatic methods for solving class imbalance (e. g., resampling).

7 Conclusion

In this paper, we developed and evaluated the effectiveness of ten novel TTC taxonomies tailored for tables found in scholarly publications. Additionally, we examined the applicability of well-established schemas designed for and based on web tables to the use-case of scientific tables. The findings reveal that existing taxonomies are indeed suitable for classifying scientific tables. However, while established taxonomies demonstrate their efficiency, comparable performance can also be achieved with two novel domain-specific taxonomies. Finally, our study indicates that header features are essential for classifying scientific tables, whereas cell features and table splitting have not shown to provide significant advantages. The proposed taxonomies can be beneficial for downstream tasks such as information retrieval from scholarly papers by helping to reduce the

search space, data integration allowing mapping of scientific tables with similar structures across different datasets, and scientific table structure recognition.

Acknowledgments. The work presented in this paper was partially supported by the consortium NFDI for Data Science and Artificial Intelligence (NFDI4DS, no. 460234259)[11] as part of the non-profit association National Research Data Infrastructure (NFDI e. V.). The NFDI is funded by the Federal Republic of Germany and its states.

A Examples of Matrix, Horizontal Listing, and Vertical Listing Tables

	Lake	Area
1	Windermere	5.69 sq mi (14.7 km²)
2	Kielder Reservoir	3.86 sq mi (10.0 km²)
3	Ullswater	3.44 sq mi (8.9 km²)
4	Bassenthwaite Lake	2.06 sq mi (5.3 km²)
5	Derwent Water	2.06 sq mi (5.3 km²)

(a) Horizontal listing

Government[3]	
• Type	Mayor–Council
• Body	New York City Council
• Mayor	Bill de Blasio (D)
Area[2]	
• Total	468.9 sq mi (1,214 km²)
• Land	304.8 sq mi (789 km²)
• Water	164.1 sq mi (425 km²)
• Metro	13,318 sq mi (34,490 km²)
Elevation[4]	33 ft (10 m)

(b) Vertical listing

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Totals	87	13	100

(c) Matrix

Fig. 6. Examples of web tables falling under the categories within the table type classification schema by Eberius et al. [14]. The samples are taken from WDC Web Table Corpus 2015.

B Illustrations of Table Features

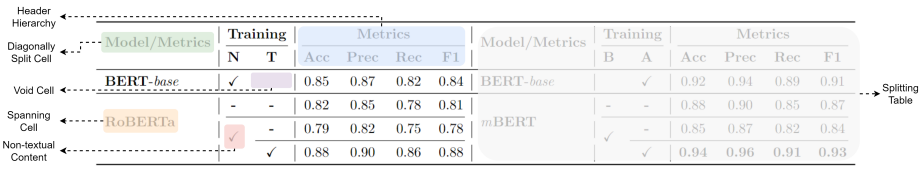


Fig. 7. Illustration of a splitting table with spanning cells, diagonally split cells, void cells, hierarchical headers, and cells with non-textual content.

[11] https://www.nfdi4datascience.de

References

1. Aly, R., Guo, Z., Schlichtkrull, M.S., Thorne, J., Vlachos, A., Christodoulopoulos, C., Cocarascu, O., Mittal, A.: The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In: Aly, R., Christodoulopoulos, C., Cocarascu, O., Guo, Z., Mittal, A., Schlichtkrull, M., Thorne, J., Vlachos, A. (eds.) Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER). pp. 1–13. Association for Computational Linguistics, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.fever-1.1
2. Bonfitto, S., Casiraghi, E., Mesiti, M.: Table understanding approaches for extracting knowledge from heterogeneous tables. WIREs Data Mining and Knowledge Discovery **11**(4), e1407 (2021). https://doi.org/https://doi.org/10.1002/widm.1407
3. Borisov, V., Leemann, T., Sessler, K., Haug, J., Pawelczyk, M., Kasneci, G.: Deep neural networks and tabular data: A survey. IEEE Transactions on Neural Networks and Learning Systems pp. 1–21 (2022). https://doi.org/10.1109/tnnls.2022.3229161
4. Cafarella, M.J., Halevy, A.Y., Zhang, Y., Wang, D.Z., Wu, E.: Uncovering the relational web. In: WebDB. pp. 1–6. Citeseer (2008)
5. Chen, W., Chang, M.W., Schlinger, E., Wang, W., Cohen, W.W.: Open question answering over tables and text. arXiv (2021)
6. Chen, W., Wang, H., Chen, J., Yunkai Zhang, H.W., Li, S., Zhou, X., Wang, W.Y.: TabFact: A large-scale dataset for table-based fact verification. In: International Conference on Learning Representations (ICLR). Addis Ababa, Ethiopia (2020)
7. Chen, W., Zha, H., Chen, Z., Xiong, W., Wang, H., Wang, W.Y.: HybridQA: A dataset of multi-hop question answering over tabular and textual data. In: Cohn, T., He, Y., Liu, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 1026–1036. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.91
8. Chen, Z., Cafarella, M.: Automatic web spreadsheet data extraction. In: Proceedings of the 3rd International Workshop on Semantic Search Over the Web. SSW '13, Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2509908.2509909
9. Cheng, Z., Dong, H., Wang, Z., Jia, R., Guo, J., Gao, Y., Han, S., Lou, J.G., Zhang, D.: HiTab: A hierarchical table dataset for question answering and natural language generation. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1094–1110. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.78
10. Chi, Z., Huang, H., Xu, H.D., Yu, H., Yin, W., Mao, X.L.: Complicated table structure recognition. arXiv preprint arXiv:1908.04729 (2019)
11. Crestan, E., Pantel, P.: Web-scale table census and classification. In: Proceedings of the fourth ACM international conference on Web search and data mining. pp. 545–554 (2011)
12. Del Bimbo, D., Gemelli, A., Marinai, S.: Data augmentation on graphs for table type classification. In: Krzyzak, A., Suen, C.Y., Torsello, A., Nobile, N. (eds.) Structural, Syntactic, and Statistical Pattern Recognition. pp. 242–252. Springer International Publishing, Cham (2022)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

14. Eberius, J., Braunschweig, K., Hentsch, M., Thiele, M., Ahmadov, A., Lehner, W.: Building the dresden web table corpus: A classification approach. In: 2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC). pp. 41–50. IEEE (2015)
15. Ghasemi-Gol, M., Szekely, P.: Tabvec: Table vectors for classification of web tables. arXiv preprint arXiv:1802.06290 (2018)
16. Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A.: Revisiting deep learning models for tabular data. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 18932–18943. Curran Associates, Inc. (2021)
17. Gupta, V., Mehta, M., Nokhiz, P., Srikumar, V.: INFOTABS: Inference on tables as semi-structured data. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 2309–2324. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.210
18. Habibi, M., Starlinger, J., Leser, U.: Deeptable: a permutation invariant neural network for table orientation classification. Data Mining and Knowledge Discovery **34**(6), 1963–1983 (2020)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), http://arxiv.org/abs/1512.03385
20. Herzig, J., Müller, T., Krichene, S., Eisenschlos, J.M.: Open domain question answering over tables via dense retrieval. arXiv (2021)
21. Hu, K., Gaikwad, N., Bakker, M., Hulsebos, M., Zgraggen, E., Hidalgo, C., Kraska, T., Li, G., Satyanarayan, A., Çağatay Demiralp: VizNet: Towards a large-scale visualization learning and benchmarking repository. arXiv (2019)
22. Iyyer, M., Yih, W.t., Chang, M.W.: Search-based neural structured learning for sequential question answering. In: Barzilay, R., Kan, M.Y. (eds.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1821–1831. Association for Computational Linguistics, Vancouver, Canada (Jul 2017). https://doi.org/10.18653/v1/P17-1167
23. Kardas, M., Czapla, P., Stenetorp, P., Ruder, S., Riedel, S., Taylor, R., Stojnic, R.: AxCell: Automatic extraction of results from machine learning papers. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 8580–8594. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.692, https://aclanthology.org/2020.emnlp-main.692
24. Karishma, Z., Rohatgi, S., Puranik, K.S., Wu, J., Giles, C.L.: ACL-Fig: A dataset for scientific figure classification. arXiv (2023)
25. Kruit, B., He, H., Urbani, J.: Tab2know: Building a knowledge base from tables in scientific papers. In: The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part I 19. pp. 349–365. Springer (2020)
26. Lautert, L.R., Scheidt, M.M., Dorneles, C.F.: Web table taxonomy and formalization. ACM SIGMOD Record **42**(3), 28–33 (2013)
27. Lehmberg, O., Ritze, D., Meusel, R., Bizer, C.: A large public corpus of web tables containing time and context metadata. In: Proceedings of the 25th International Conference Companion on World Wide Web. p. 75–76. WWW '16 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2016). https://doi.org/10.1145/2872518.2889386
28. Moosavi, N.S., Rücklé, A., Roth, D., Gurevych, I.: Learning to reason for text generation from scientific tables. arXiv preprint arXiv:2104.08296 (2021)

29. Nan, L., Hsieh, C., Mao, Z., Lin, X.V., Verma, N., Zhang, R., Kryściński, W., Schoelkopf, H., Kong, R., Tang, X., Mutuma, M., Rosand, B., Trindade, I., Bandaru, R., Cunningham, J., Xiong, C., Radev, D., Radev, D.: FeTaQA: Free-form table question answering. Transactions of the Association for Computational Linguistics **10**, 35–49 (2022). https://doi.org/10.1162/tacl_a_00446
30. Nassar, A., Livathinos, N., Lysak, M., Staar, P.: Tableformer: Table structure understanding with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4614–4623 (2022)
31. Nishida, K., Sadamitsu, K., Higashinaka, R., Matsuo, Y.: Understanding the semantic structures of tables with a hybrid deep neural network architecture. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017)
32. Paliwal, S., D, V., Rahul, R., Sharma, M., Vig, L.: Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images. arXiv (2020)
33. Pasupat, P., Liang, P.: Compositional semantic parsing on semi-structured tables. In: Zong, C., Strube, M. (eds.) Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1470–1480. Association for Computational Linguistics, Beijing, China (Jul 2015). https://doi.org/10.3115/v1/P15-1142
34. Prasad, D., Gadpal, A., Kapadni, K., Visave, M., Sultanpure, K.: Cascadetabnet: An approach for end to end table detection and structure recognition from image-based documents. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 572–573 (2020)
35. Roldán, J.C., Jiménez, P., Corchuelo, R.: On extracting data from tables that are encoded using html. Knowledge-Based Systems **190**, 105157 (2020)
36. Sahakyan, M., Aung, Z., Rahwan, T.: Explainable artificial intelligence for tabular data: A survey. IEEE Access **9**, 135392–135422 (2021). https://doi.org/10.1109/ACCESS.2021.3116481
37. Schreiber, S., Agne, S., Wolf, I., Dengel, A., Ahmed, S.: DeepDeSRT: Deep learning for detection and structure recognition of tables in document images. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 1162–1167 (2017). https://doi.org/10.1109/ICDAR.2017.192
38. Shigarov, A.: Table understanding: Problem overview. WIREs Data Mining and Knowledge Discovery **13**(1), e1482 (2023). https://doi.org/https://doi.org/10.1002/widm.1482
39. Shigarov, A.O., Mikhailov, A.A.: Rule-based spreadsheet data transformation from arbitrary to relational tables. Information Systems **71**, 123–136 (2017). https://doi.org/https://doi.org/10.1016/j.is.2017.08.004
40. Shwartz-Ziv, R., Armon, A.: Tabular data: Deep learning is not all you need. Information Fusion **81**, 84–90 (2022). https://doi.org/https://doi.org/10.1016/j.inffus.2021.11.011
41. Wang, Y., Hu, J.: Detecting tables in html documents. In: Lopresti, D., Hu, J., Kashi, R. (eds.) Document Analysis Systems V. pp. 249–260. Springer Berlin Heidelberg, Berlin, Heidelberg (2002)
42. Wang, Y., Hu, J.: A machine learning based approach for table detection on the web. In: Proceedings of the 11th international conference on World Wide Web. pp. 242–250 (2002)
43. Zayats, V., Toutanova, K., Ostendorf, M.: Representations for question answering from documents with tables and text. arXiv preprint arXiv:2101.10573 (2021)

44. Zhang, L., Zhang, S., Balog, K.: Table2vec: Neural word and entity embeddings for table population and retrieval. In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. pp. 1029–1032 (2019)
45. Zhang, S., Balog, K.: Web table extraction, retrieval, and augmentation: A survey. ACM Transactions on Intelligent Systems and Technology (TIST) **11**(2), 1–35 (2020)
46. Zheng, X., Burdick, D., Popa, L., Zhong, X., Wang, N.X.R.: Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 697–706 (2021)
47. Zheng, Z., Cai, Y., Li, Y.: Oversampling method for imbalanced classification. Computing and Informatics **34**(5), 1017–1037 (2015)
48. Zhong, V., Xiong, C., Socher, R.: Seq2SQL: Generating structured queries from natural language using reinforcement learning. arXiv (2017)
49. Zhong, X., ShafieiBavani, E., Yepes, A.J.: Image-based table recognition: data, model, and evaluation. arXiv (2020)
50. Zhu, F., Lei, W., Huang, Y., Wang, C., Zhang, S., Lv, J., Feng, F., Chua, T.S.: TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 3277–3287. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.254