# MADS: A Multi-modal Academic Document Segmentation Dataset for Smart Question Bank Management

Utathya Aich[1,*], Swarnendu Ghosh[2] and Tulika Saha[3]

[1]CNH Indutrial ITC, India
[2]Institute of Engineering & Management, (University of Engineering & Management), Kolkata, India
[3]University of Liverpool, United Kingdom

**Abstract**

In today's world, most major academic institutes and organizations conduct competitive exams to assess eligibility of students for admission or recruitment. Due to the rising craze among participants, traditional methods are not optimized enough to get ahead in the race. The inclusion of AI enabled tutoring is mandatory for such exams. One such area of implementation is smart question bank management system. Though we have large volumes of questions of competitive exams in physical mode, however, they are harder to process visually for systems as they consist of several types of text and non-text elements such as numbers, equations, images alongside textual paragraphs. For this purpose, we propose *MADS*, which is a multi-modal academic document segmentation dataset consisting of images of documents containing heterogeneous questions from the competitive exams like GMAT, GRE, GATE, SAT, UGC-NET. These documents consist of textual paragraphs along with numbers, images and equations. The dataset comes with bounding box annotation in two popular format YOLO and PASCAL-VOC formats to aid the development of efficient document segmentation algorithms. Additionally, benchmarks have been provided for state of the art deep learning based implementations such as Faster RCNN and YOLO-v8. From application point of view, the proposed dataset can identify different objects in an image so that later it can be used for semantic relationship and question answering applications enhancing comprehension and personalized learning experiences, thus, supporting the goal of providing quality education.

**Keywords**

Document Image Analysis, Multi-modal Document Processing, Text Classification, Deep Learning

## 1. Introduction

Competitive examinations are one of the most commonly used tools for academic performance assessment. These are generally conducted for selection of candidates suitable for a specific branch of study or work. There are multiple such exams which have become popular in both the national and international levels. Due to this increase in competition, students and teachers are finding it hard to optimize the preparation process using traditional methods which often leads to depression amongst them [1]. While e-documents are more suitable for automated systems, it is hard to find organized question banks or materials available in the electronic format. Hard copies of question banks are available but they are difficult to be directly processed as text,

*Liverpool'24: Symposium on NLP for Social Good, April 25–26, 2024, Liverpool, UK*
*Corresponding author.
utathya.aich@cnh.com (U. Aich); drghosh90@gmail.com (S. Ghosh); sahatulika15@gmail.com (T. Saha)

as they contain a mixture of texts, equations, images, numbers and so on. One of the major challenges with such documents containing a mixture of several mediums is localizing and segmenting the appropriate textual and non-textual elements. All these components have text like properties and they can mess up standard OCR techniques. The solutions are more scarce when it comes to solving queries containing multi-modal data. This becomes especially prominent for document images that does represent data as a sequence of Unicode characters, but as pixels. To implement a truly multi-modal question answering system, it is essential to segment this various components from complex documents before these advanced image processing tools can be used. For this purpose, we propose "MADS" which is a multimedia academic document segmentation dataset. For this specific work, we are primarily focusing on questions of competitive exams of national and international levels such as GMAT, UGC-NET, GRE, GATE and SAT. This covers a large variety of examinations catering to students of various fields. The images in these documents contain a mixture of equations, diagrams and numbers embedded within the body of the questions along with multiple options to choose from as well. The proposed dataset comes with bounding box annotation corresponding to 4 classes namely equations, diagrams, numbers and texts offering a transformative resource that aligns with Sustainable Development Goal of Quality Education. By meticulously annotating various elements such as text, images, equations, and numbers within question papers, this dataset lays the groundwork for advancing educational research and technology applications. Leveraging this dataset enables the development of innovative tools and algorithms aimed at enhancing teaching methodologies, personalized learning experiences, and educational accessibility. Through the identification of text, images, and equations, educational materials can be optimized for accessibility features such as text-to-speech conversion and alternative formats for students with disabilities. This ensures that all learners, including those with visual impairments or learning disabilities, can access educational content on an equal basis. The availability of the proposed dataset allows for the development of intelligent tutoring systems and question-answering algorithms that promote deeper understanding of educational concepts. Active participation and sustained engagement in the learning process can be obtained through the immediate feedback and adaptive learning pathways.

**Contributions :** The key contribution of this work are as follows : **(i)** To establish the problem statement for multi-modal academic document image segmentation and its future applications; **(ii)** Provide a challenging dataset of multi-modal document images consisting of questions from various types of competitive examinations; **(iii)** To provide with necessary annotation for document image segmentation into 4 classes, namely, equations, numbers, images, and text; and **(iv)** To provide benchmarks using state-of-the-art detection algorithms.

## 2. Related Work

There has been previous approaches to managing question banks and exam protocols through AI based technologies [2]. However, most of the approaches deal with already existing electronic question banks [3]. There has not been much work that can automatically process the already existing large volumes of question banks available in the printed medium in the form of

previous year question papers, study materials, educational magazines, and so on. However, there have been several applications of computer vision on multi-modal documents from other domains[4, 5, 6]. Some of these approaches primarily focus on text and non-text separation in various scenarios [7, 8, 9]. In terms of multi-modal text datasets we have applications in multiple areas that have similar set of challenges to our proposed domain. The Tobacco-3482 [10] dataset consists of document images belonging to 10 different classes such as forms, letters, resumes, memos, forms and so on. The RVL-CDIP dataset [11] consists of 400,000 grayscale images in 16 classes, with 25,000 images per class. Multi-label classification have been performed on academic papers to extract components such as titles and keywords [12]. Moreover, some multi-modal document image datasets that deal with mathematical equations [13] or geometry [14] problems have also been explored. In terms of exam related problems, there are some similar works done in specific subject groups such as social or natural sciences [15] or medical entrance exams [16]. In these methods there are implementations that address multilingual Q&A problems and also multiple choice based questions. However, after a through survey it is evident that there is a lack of datasets operating in unrestricted domains and provide fundamental annotations regarding the multi-modal contents. Furthermore in the proposed dataset, we are providing samples which do not have unicode representations thus, making it equivalent to digitally scanned print media.

## 3. Dataset

Due to the unavailability of multi-modal question bank dataset through which one can segregate different text and non-textual elements from a given question through document segmentation, we propose "MADS" and discuss its creation below. The sample dataset is publicly available under Creative Commons License (CC) by the authors[1].

### 3.1. Data Collection

As mentioned above, managing question banks can be a tedious process as the questions just does not contain free text but its different form of representation such as equations, labels, figures etc., mostly found in competitive examinations. As a result, we chose different competitive exams which are relevant to the global community as our base to collect questions of different form. Our dataset sources are GRE, GATE, SAT, UGC-NET, GMAT which are typically well-known competitive exams to pursue higher education and is very popular amongst students. We collected official and sample questions from these exams openly available in web by extracting pages from downloaded PDFs corresponding to mock questions and converted them into image format (.png). We utilised the official question bank for GATE and UGC-NET examinations and mock questions for the remaining ones. The collected data samples or questions are in the form of images consisting of equation, number, image and text. These questions are publicly available on the Main exam website as previous year questions. Some questions are taken from mock questions available for free on the different mock exam websites [17, 18, 19, 20, 21, 22, 23, 24, 25, 26]. A sample raw question is shown in Figure 1.

[1]https://github.com/MADS-dataset/MADS_Dataset_official

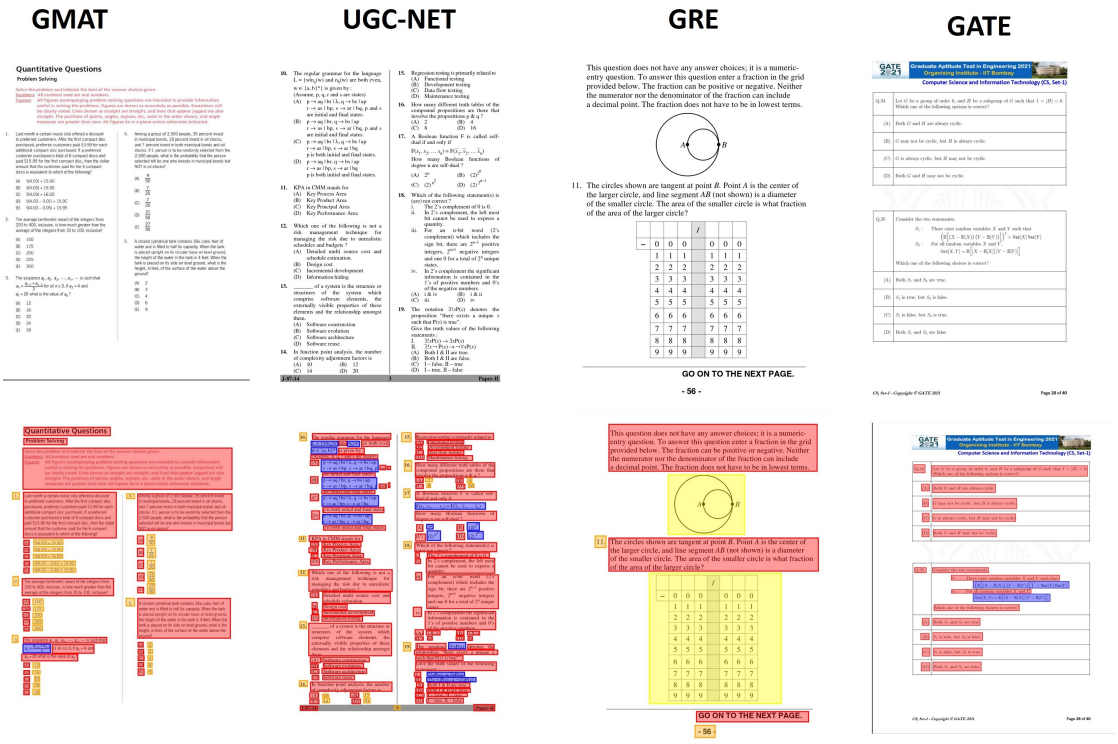
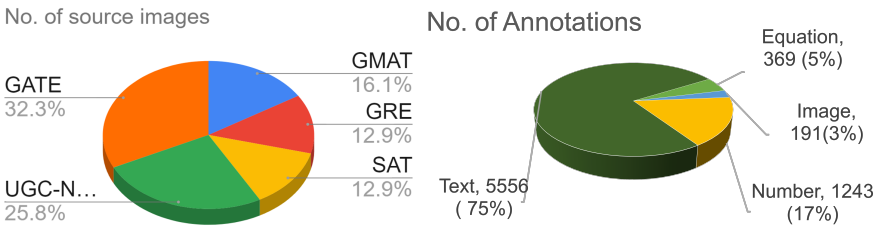



**Figure 1:** Samples of MADS from different sources of examination: top row - original question, bottom row - annotated sample question. Red ='Text', Orange ='Number', Yellow ='Image', Blue ='Equation'

## 3.2. Data Annotation

Next, the task was to annotate the images (typically questions) for extracting relevant information from these images which are question text, image, label and equation. All the sample questions were uploaded to an open-source annotation tool, Label- Studio[2] for creating bounding boxes. Three annotators from the authors' affiliation were asked to draw bounding boxes for these samples through this tool. The annotators were explained and demonstrated the task and then were initially asked to annotate 10 samples each for these four categories present in the image. These samples were then checked by the authors and the errors were resolved if any. The annotators were then finally provided with all the remaining samples equally divided amongst the three for annotation. On an average, there were at least one bounding box present each for image and text class in each sample of the dataset. To create the gold standard annotated dataset, we maintained the Intersection Over Union (IoU) score [27] between the annotated box to be atleast 80 and in addition to that Cohen kappa score to be greater than 90% for the acceptance of the bounding box with the class label. This Cohen kappa score is the agreement between the annotated labels by the annotators and the authors verifying the annotations.

[2]https://labelstud.io/




(a) (b)

**Figure 2:** Statistics from MADS: (a) Distribution of different source representation, (b) Distribution of different class labels

### 3.3. MADS

MADS now comprises of 230 question samples annotated for the presence of four categories of information, namely, question text, image, label and equation with the help of bounding box. An annotated sample from MADS is shown in Figure 1. As is visible, it contains a mixture of equation, text, image and number, and is challenging for machines to identify these said parts in an image easily. Some of these questions contain both numeric, text and equation on the same line. Some images include both image and equation at the same place. In some images questions are in two column format, which makes the dataset more challenging to segment the regions. It is indeed difficult to identify and differentiate amongst these and through *MADS* we aim to tackle such diverse situations. Largest contribution of the dataset comes from the GATE question which is 32.3% of the whole, followed by UGC-NET, GMAT, GRE and SAT. The distribution of the dataset is shown in Figure 2a. It has been observed that the dataset exhibits a predominance of text comprising of 5536 bounding boxes which is 75.5% of the annotations. The lowest number of bounding boxes are present with images which is 191 and is 3% of the dataset. The class based statistics is depicted in Figure 2b.

## 4. Methodology

The aim of *MADS* is to facilitate easy training of models in order to identify different categories in a given image of a question. The dataset can facilitate in identifying different objects in an image which can be later be used for semantic relationship and question answering. The trained model on *MADS* should then be able to identify and segregate different information present in the question for smart question bank management and facilitate future research directions in this area. In this section, we aim to benchmark *MADS* using different state of the art vision models for detecting the bounding box.

### 4.1. Benchmark Setup

We benchmark *MADS* using two state of the art vision models as follows :

• **YOLO-v8[3]:** YOLO-v8 is an advancement of the YOLO [28] model. The advanced model is developed by Ultralytics. It has a high rate of accuracy on the COCO dataset[4]. It is an anchor free model which means it predicts the center of an object rather than offset from a known anchor box. This model is more robust to noise and occlusions than other available models. The model uses a new backbone network called Panoptic Feature Extractor (PEE), a new loss function called CIoU loss, and a new training strategy called SimOTA.

• **Faster R-CNN:**[29] Ross Girshick developed Faster R-CNN. Compared to past models like R-CNN, a new layer called ROI pooling layer has been proposed in this model. The model is a single stage network in comparison with other previous models. Faster R-CNN does not need much disk storage compared to R-CNN as it does not cache the extracted features.

The pre-trained YOLO-v8 is fine-tuned and Faster R-CNN is trained on *MADS* to benchmark the dataset using state of the art vision models for the task of detecting useful information in the form of bounding box.

## 4.2. Implementation Details

*MADS* is divided into train and test set with a ratio of 85:15. We conducted the experiment five times and reported average of the results based on different models. Vanilla YOLO-v8[5] medium model is fine-tuned on *MADS*. This model has 25.9 million parameters. Vanilla Faster R-CNN model[6] with ResNet-50 is used at its backend to train on *MADS*. All the parameters are set to their default values. The learning rate has been set to 0.001, batch size is 64. Number of anchors have been set to 3. As there are 4 classes for detection in *MADS*, we have 4 output neurons. Confidence threshold has been set to 0.25 by default. YOLO-v8 uses LeakyReLU as its activation function. These parameters might be tuned in future for obtaining better performance. We used two evaluation metrics - Intersection Over Union (IoU) and Mean Average precision (mAP) score to benchmark the performance of the models.

**Evaluation Metrics.** The metrics IoU and mAP score are explained as follows:

• **IoU Score:** This metric is commonly used to evaluate the performance of object detection algorithms. It measures the overlap between the predicted bounding box and the ground truth. The IoU is calculated using the following formula:

$$IoU = Area\_of\_Overlap \,/\, Area\_of\_Union \tag{1}$$

where, Area_of_Overlap is the area common to both the predicted and ground truth regions and Area_of_Union is the total area covered by both the predicted and ground truth regions. Our experiments are evaluated same threhold of IoU used in COCO.The predicted annotations are evaluated using IoU threshold of 0.5 and 0.9 respectively .

• **Mean Average Precision (mAP):** Mean Average Precision is a commonly used metric to evaluate the performance of object detection or information retrieval systems. It provides a

[3]https://docs.ultralytics.com/

[4]https://cocodataset.org/#home

[5]https://github.com/ultralytics/ultralytics?tab=readme-ov-file

[6]5https://pypi.org/project/detecto/

single scalar value for two IoU threshold. We first find the average precision of each class then average of all the classes is done to find the mAP.

## 5. Results and Analysis

We take the average among all results for each model from the experiments to get the final result of the *MADS* dataset. Based on the predicted IoU score, we create a threshold of 50% and 90% to record the mAP score. Based on this threshold, we compute different metrics such as accuracy, precision and recall of the models in order to determine its performance. Table 1 depicts the accuracy of Yolo-v8 and Faster R-CNN models trained on *MADS*. As observed, YOLO-v8 performs better than Faster R-CNN by a significant improvement of about 15% in terms of accuracy when the IoU threshold is 50%. Similarly, when the IoU threshold is set to 90%, YOLO-v8 shows about 3% improvement with respect to Faster R-CNN. On an average, it is observed that the Yolo-v8 model showed a standard deviation of ±0.5 and ±1 on IoU threshold of 50% and 90% respectively for the overall accuracy. On the other hand, Faster R-CNN tends to have a standard deviation of ±0.7 and ±2.6 on IoU threshold of 50% and 90% respectively for the same.

**Table 1**
Average Class wise Precision and Recall of Faster RCNN and YOLO-v8 by set IoU for box overlap at 50% and 90%

| Model | Accuracy | Average Precision | | | | Average Recall | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Equation | Image | Number | Text | Equation | Image | Number | Text |
| Faster RCNN @ IoU50 | 79.1% | 47.6% | 35.9% | 64.7% | 90.2% | 32.1% | 48.3% | 71.7% | 91.2% |
| YOLO-v8 @ IoU50 | 93.7% | 73.4% | 80.5% | 86.6% | 96.5% | 69.2% | 77.1% | 92.02% | 96.5% |
| Faster RCNN @ IoU90 | 94.5% | 63.8% | 97.5 | 95.3% | 96.9 | 97.46 | 100% | 49.9% | 96.9% |
| YOLO-v8 @ IoU90 | 97.1% | 98.3% | 97.5% | 50.8% | 98.02% | 66.3% | 88.9% | 97.6% | 98.7% |

Table 1 also creates a benchmark on the precision and recall for each of the classes by the different models for 50% and 90% threshold. Experimental results noted that the class level precision tends to have a standard deviation of ±3.6 for equation, ±2.2 for image, ±1.9 for number, ±0.9 for text on IoU threshold of 50% for Faster R-CNN. For YOLO-v8 on IoU threshold 50%, precision for class level showed a standard deviation of ±1.5 for equation, ±1.1 for image, ±1.08 for number and ±1.2 for text. YOLO-v8 performs better than Faster R-CNN with a narrow performance improvement of about 3% when the IoU threshold is 90%. On IoU threshold of 90% it is observed that YOLO-v8 has a standard deviation of ±2.1 for equation, ±1.9 for image, ±2.9 for number and ±2.01 for text whereas Faster R-CNN showed a standard deviation of ±3.7 for equation, ±2.4 for image, ±4.7 for number and ±3.5 for text. The reason behind YOLO-v8 superior performance can be attributed to the fact that Faster R-CNN uses two stage detectors during training while YOLO-v8 uses a single shot detector. This gives a huge advantage to YOLO-v8 to look through the whole image at once whereas Faster R-CNN uses regions to localize the object within the image. We also report the precision and recall for individual class labels. The mAP score for the Faster RCNN for IoU50 is 59.6% whereas for IoU90 is 88.37%.YOLO-v8 has a mAP score of 84.25% for IoU50 and 86.15% for IoU90 respectively. It is observed that the text tag seems to be the easiest to identify based on the performance as the dataset has the highest

number of text tag annotations. Sample prediction of Figure 1 from the YOLO-v8 model is shown in Figure 3.

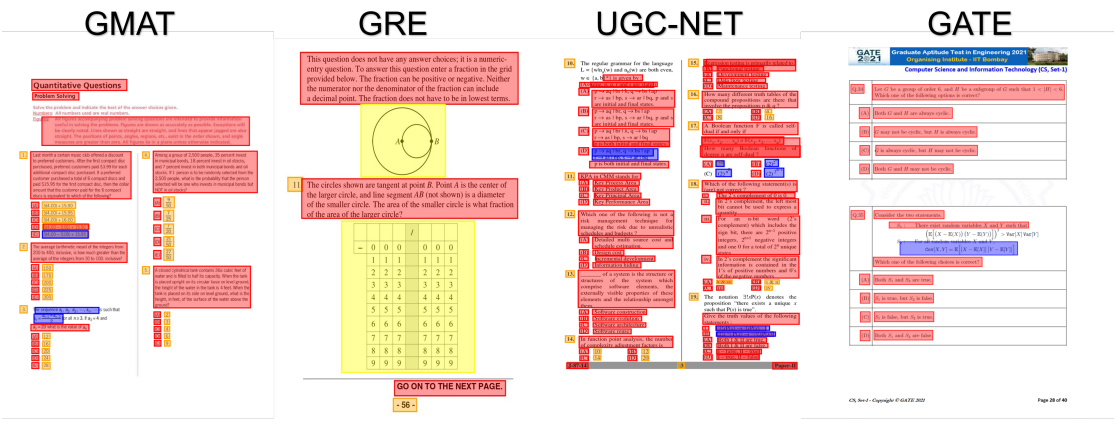



**Figure 3:** Predicted samples from YOLO-v8 for the images in Figure 1

With the increase in IoU threshold from 50% to 90%, it is observed that the models are able to correctly classify the different tags. When the threshold is tuned to be 50%, more bounding boxes are identified and there seems to be mis-classification for the same. YOLO-v8 model lacks to classify number tags despite increase in precision for other tags while increasing the threshold from 50% to 90%. Here, Faster R-CNN outperforms YOLO-v8 while identifying number tags on IoU threshold of 90%. Though the YOLO-v8 performs better than Faster R-CNN in almost every scenario, challenges do exists. Both the algorithms faces difficulty while identifying equation and image interchangeably when they are mixed. Isolating such instances while preserving their semantic relationships poses a considerable challenge. Some challenging image snippets are shown in Figure 4. The models tend to find difficulty in segregating equation and images. These issues can be further resolved by fine-tuning the hyper parameters. Size of the dataset needs to be scaled up (which is an ongoing effort) to achieve a better performance.

## 6. Conclusion and Future Work

In this paper, we established a novel problem statement for multi-modal academic document image segmentation and steer discussion focused on its future applications. Due to the unavailability of any such existing dataset relevant to the task, we propose a dataset, namely, *MADS* consisting of questions from various types of competitive examinations and gold-standard annotations to extract information from these questions through the task of bounding box detection. We benchmark *MADS* with the help of several state of the art vision models. The dataset exhibits a predominance of text documents compared to other object classes, revealing a bias in the performance of the base algorithms towards text detection. Challenges arise when labels are annotated within the bounding boxes of text. In case of text, characters are distributed in a horizontal and vertical format, meaningful segments can be enclosed in a rectangular bounding box. To address this bias, fine-tuning strategies can be implemented to improve

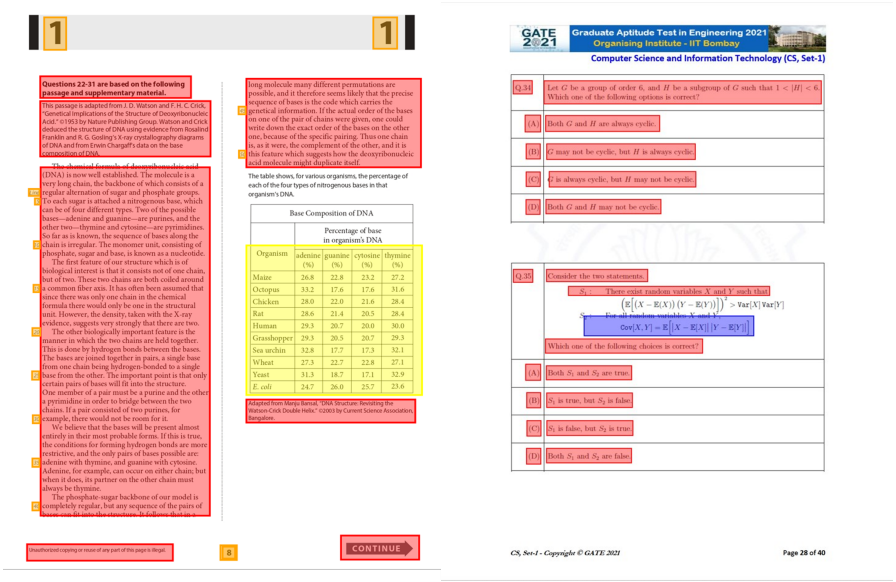



**Figure 4:** Challenging image snippets from MADS

the accuracy for other class labels. This presents an intriguing area for future research, as overcoming these complexities would contribute significantly to the advancement of the field. The primary goal for releasing this dataset is to spur a domain of automated teaching based learning method to aid students appearing for such competitive exams. At its first iteration, this dataset provides the opportunity to digitize existing question banks and annotating them during this process. At this point the dataset primarily focuses on segregation of text, equations, figures, and numbers. Finer segregation may be incorporated in the future versions of the dataset. Future iterations will focus on increasing the volume of the dataset and broadening the domain, embedding multi-modal questions for processing in large language models and vision language models, integrating GPT based services to retrieve solutions for questions, personalized mock test generations and so on. We summarise that this dataset will drive novel research contributions and applications in the field of smart question bank management and education in general.

## Acknowledgement

Dr. Swarnendu Ghosh is thankful for the infrastructure support from IEM Centre of Excellence for Data Science and the Innovation & Entrepreneurship Development Cell, IEM Kolkata.

## References

[1] A. Shrivastava, D. Rajan, Assessment of depression, anxiety and stress among students preparing for various competitive exams, International Journal of Healthcare Sciences 6 (2018) 50–72.

[2] G. Kurdi, J. Leo, B. Parsia, U. Sattler, S. Al-Emari, A systematic review of automatic question generation for educational purposes, International Journal of Artificial Intelligence in Education 30 (2020) 121–204.
[3] G. Nalawade, R. Ramesh, Automatic generation of question paper from user entered specifications using a semantically tagged question repository, in: 2016 IEEE Eighth International Conference on Technology for Education (T4E), IEEE, 2016, pp. 148–151.
[4] S. Bakkali, Z. Ming, M. Coustaty, M. Rusiñol, Visual and textual deep feature fusion for document image classification, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 562–563.
[5] R. K. Srihari, Z. Zhang, A. Rao, Intelligent indexing and semantic retrieval of multimodal documents, Information Retrieval 2 (2000) 245–275.
[6] J. Bateman, Multimodality and genre: A foundation for the systematic analysis of multimodal documents, Springer, 2008.
[7] X. Bai, B. Shi, C. Zhang, X. Cai, L. Qi, Text/non-text image classification in the wild with convolutional neural networks, Pattern Recognition 66 (2017) 437–446.
[8] K. Dutta, M. Bal, A. Basak, S. Ghosh, N. Das, M. Kundu, M. Nasipuri, Multi scale mirror connection based encoder decoder network for text localization, Pattern Recognition Letters 135 (2020) 64–71.
[9] L. Unsworth, Image/text relations and intersemiosis: Towards multimodal text description for multiliteracies education, in: Proceedings of the 33rd IFSC: International Systemic Functional Congress, Pontificia Universidade Catolica de Sao Paulo, 2007.
[10] J. Kumar, P. Ye, D. Doermann, Learning document structure for retrieval and classification, in: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), IEEE, 2012, pp. 1558–1561.
[11] A. W. Harley, A. Ufkes, K. G. Derpanis, Evaluation of deep convolutional nets for document image classification and retrieval, in: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2015, pp. 991–995.
[12] G. Mustafa, M. Usman, L. Yu, M. T. Afzal, M. Sulaiman, A. Shahid, Multi-label classification of research articles using word2vec and identification of similarity threshold, Scientific Reports 11 (2021) 21900.
[13] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, J. Steinhardt, Measuring mathematical problem solving with the math dataset, arXiv preprint arXiv:2103.03874 (2021).
[14] M. Seo, H. Hajishirzi, A. Farhadi, O. Etzioni, C. Malcolm, Solving geometry problems: Combining text and diagram interpretation, in: Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 1466–1476.
[15] M. Hardalov, T. Mihaylov, D. Zlatkova, Y. Dinkov, I. Koychev, P. Nakov, Exams: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering, arXiv preprint arXiv:2011.03080 (2020).
[16] A. Pal, L. K. Umapathi, M. Sankarasubbu, Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering, in: Conference on Health, Inference, and Learning, PMLR, 2022, pp. 248–260.
[17] GMAT, Gmat sample question paper 2023 with 100 q and a | eduaims, https://eduaims.in/gmat-sample-paper-pdf/, Sample Questions.

[18] Hank walker, https://people.engr.tamu.edu/d-walker/index.html, Practice Problems.
[19] GRE, Gre_practicebook_2004.pdf, https://www.prepscholar.com/gre/blog/wp-content/uploads/sites/3/2016/09/GRE_practicebook_2004.pdf, Practice Problems.
[20] GRE, Gr9768.pdf, https://wmich.edu/mathclub/files/GR9768.pdf, Practice Problems.
[21] M. PREP, 5 lb. book, https://dl.icdst.org/pdfs/files1/eceb4737c3836a94ef7ba0b88ae5510b.pdf, Practice Problems.
[22] S. S. Questions, Sat study guide 2020 - practice test 9, https://satsuite.collegeboard.org/media/pdf/sat-practice-test-9.pdf, Sample Questions.
[23] S. S. Questions, Sat study guide 2020 - practice test 10, https://satsuite.collegeboard.org/media/pdf/sat-practice-test-10.pdf, Sample Questions.
[24] S. S. Questions, Sat study guide 2020 - practice test 3, https://satsuite.collegeboard.org/media/pdf/sat-practice-test-3.pdf, Sample Questions.
[25] U. NET, University grants commission - net, https://www.ugcnetonline.in/previous_question_papers.php, Officail Question papers.
[26] GATE, Gate 2022 official site, https://gate.iitkgp.ac.in/old_question_papers.html, Official Question papers.
[27] Jaccard index - wikipedia, https://en.wikipedia.org/wiki/Jaccard_index, IOU Similarity.
[28] J. Redmon, S. K. Divvala, R. B. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, CoRR abs/1506.02640 (2015). URL: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Redmon_You_Only_Look_CVPR_2016_paper.pdf.
[29] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in neural information processing systems 28 (2015).