DS Journal of Digital Science and Technology
Volume 4 Issue 2, 23-42, Apr - Jun 2025
ISSN: 2583-5416/ https://doi.org/10.59232/DST-V4I2P102

Original Article

Constructing a Multi-Modal Dataset for Digital Learning Feature Extraction

Bao Chau[1], Anh Le[1], Quang Cao[1], Thuy Nguyen[1], Sang Ho[1], Giang Ma[1], Hai Tran[1*]

[1]*Faculty of Information Technology, Ho Chi Minh University of Education (HCMUE), HCM, Vietnam.*

*haits@hcmue.edu.vn

Received: 09 February 2025; Revised: 08 March 2025; Accepted: 06 April 2025; Published: 30 April 2025

Abstract - The rapid advancement of technology has catalyzed the widespread adoption of online platforms, transforming communication, learning, and professional practices across diverse sectors. In education, this technological shift has spurred the integration of digital tools to enhance pedagogical methodologies. A critical component of this integration is the development of high-quality, realistic datasets to train educational models and tools. This study aims to address this need by constructing VNEC2018, a standardized image dataset derived from Vietnam's 2018 General Education Program. The dataset is designed to support and elevate student learning outcomes through accurate and representative digital resources. By adhering to this rigorous framework, VNEC2018 is positioned to become a benchmark resource for educational technology, facilitating the creation of robust training models and addressing the evolving demands of digital-age pedagogy.

Keywords - Digital learning materials, Graph Neural Network, General education programs, Multilabel image classification.

1. Introduction

Machine Learning is a fast-growing branch and the core of Artificial Intelligence (AI) and computer science, which focuses on using data and algorithms to give computer systems the ability to learn, process data and improve themselves instead of being explicitly programmed, gradually improving their accuracy in making predictions based on input data [1]. Machine Learning has also become one of the buzzwords of technology in our time as it plays a significant role in many real-world applications such as image and speech recognition, traffic warnings, self-driving cars, medical diagnosis, etc [2].

Generally, a Machine Learning project is always functioned by the combination of the dataset and the model. In a particular way, the dataset is fundamental to building the model since it provides the information it must learn from. The proposed model uses dataset patterns and makes predictions. Together, they form the core structure of any machine learning workflow. Hence, datasets play an important role in this process, so researchers must be self-conscious about choosing the database based on its suitability and the quality it can acquire.

In the context of the technological explosion, people learn and utilize new modern applications for life and work, and the education system is changing little by little with the integration of digital learning technology in

teaching applications. Digital Learning Resources (DLRs), also called electronic learning resources, refer to the structured and formalized learning materials stored in the computer in the form of a certain structure, paradigm and scenario and which, when used, are to impart knowledge and offer learning on a smart electronic device. In the digital space, learning resources may be documents, slides, graphs, audio-visual media, interactive modules, or hybrids of these formats [3]. After the Covid-19 pandemic, online methods have become more and more popular and widely used, so the Vietnamese Ministry of Education and Training has issued a decision on the orientation of building digital learning materials and online courses Massive Open Online Courses (MOOCs) and related decisions on digital transformation in education by 2022 [4], which affirms the position and importance of the quality of digital learning materials. Moreover, those potential platforms people passionate about data have opportunities to explore and dig into, generating a newly-built dataset with various educational features supporting the process of training a model in the future.

The 2018 general education program in Vietnam is being implemented in educational institutions in the country to meet students' capacity needs in terms of knowledge and life skills, helping students develop comprehensive capacity to meet the needs of humanity [5]. However, when combined with technological innovation, the application of modern technological techniques today can partly help develop learners' capacity for the teaching quality of domestic schools and help them access diverse and easier materials. In addition, the data set after the study can be referenced or used by other digital learning platform construction projects, making it easy to extract and search for lesson content and, at the same time, helping learners learn more effectively during the learning process. Data from the curriculum will be the basis for the research process and create resources for educational projects as a premise for training AI models to provide analysis or suggest content for users in the future. Therefore, building a dataset for the 2018 General Education Program is a practical and necessary task, not only meeting current needs but also opening many development opportunities for education in the future. This image dataset will be a foundation to support the e-learning system and research in education. At the same time, it will create favourable conditions for students to access knowledge comprehensively. In response to the practical demands mentioned, this research introduces the development of the VNEC2018 image dataset - a systematically organized and standardized resource designed to enhance the implementation of Vietnam's 2018 General Education Program in creating scores of applications for students in their academic studies or instructors for their teaching methods and materials as well.

2. Related works

To have an accurate machine learning model, the dataset will be one of the crucial factors in the research process. In recent years, academic research on Machine Learning on building and analysing image data has been growing strongly to upgrade knowledge on building raw data to improve the performance and interpretation of models. Based on knowledge from previous studies, the collected data can come from various sources (self-collected or from available and grouped datasets). The study of F.A. Mbiaya, C. Vrain and F. Ros et al. has presented three sub-datasets created from the large Microsoft COCO dataset [6, 7]. Another dataset is the SCB-dataset of the Yang & Wang research group on student behaviour in the classroom at a university in Chengdu, China, which manually collects data from real videos to reflect classroom behaviours and uses the YOLOv7 algorithm to evaluate the quality [8]. In addition, according to Loris et al., the team increased the number of images in the dataset using data augmentation methods from four sets of VIR, BARK, GRAV and POR and gave promising results [9]. A study from 2016 by Yao et al. collected image data using the main query word to expand the original semantically related query using Google Books Ngrams Corpora (GBNC) to gather a more diverse and richer query and then removed irrelevant images using the CNN algorithm [10]. In summary, instead of going into the collection techniques, the above article focuses on the dataset's structure, processing process and filtering of noisy data after collection to build a high-quality image dataset. In addition, there are many other outstanding collection methods that support effective data collection and construction.

Regarding the datasets related to the topic, there is currently no specific dataset available to support research on the topic of digital learning materials for the 2018 general education program in Vietnam, however, there are still some typical datasets that can help with image classification in digital learning materials data. The first is slide images, which are collected from many sources using construction techniques suitable for small datasets [11]. The research results show that Slide Images is a potential and useful dataset in multi-modal classification. Another dataset mentioned in the above research article is similar to Slide Images, DocFigure, created from three available datasets (Figure Seer, Revision and DeepChart) [12]. Although there are related datasets on the classification of digital resources, there is still a significant research gap in building a specific image dataset for the general education program in Vietnam. Specifically, there is currently no dataset specifically designed to reflect the content and structure of the 2018 general education program in a systematic and fully annotated manner, serving the purpose of developing machine learning models in the education sector. According to the national program, this leads to a major challenge in applying modern technologies such as machine learning and artificial intelligence to support teaching and learning.

In addition, current datasets such as SlideImages or DocFigure have potential applications in classifying digital learning materials, but they are not attached to a specific curriculum, lack systematicity and do not meet the specialized requirements of general education content in Vietnam. Therefore, the problem is building a structured image dataset that reflects the curriculum content while ensuring high applicability in developing intelligent learning systems.

From this gap, the study proposes the construction of the VNEC2018 dataset as a solution to solve the above problem. The dataset not only aims to collect and standardize image resources from subjects in the 2018 general education program but is also designed to train machine learning models, contributing to promoting the digital transformation process in Vietnamese education.

3. Theoretical Basis

Digital learning resources, as defined by Circular 21/2017/TT-BGDDT, consist of various electronic educational tools. These comprise digital textbooks, electronic reference materials, online assessment tools, digital presentations, electronic spreadsheets, audio recordings, digital images and videos, recorded lectures, educational software applications, virtual simulations, and other educational content in digital format. [13].

The concept of "digital learning materials" or "electronic learning materials" is a concept defined by many authors. According to Parrott and Kok (1997) [14], electronic learning materials are learning materials provided in electronic format, which integrate different digitized multimedia formats such as text, audio, and animation. Akker et al. (1992) [15] supplemented this definition by including within its other types of texts and supporting materials appropriate to specific teaching objectives.

Tran Thi Lan Thu and Bui Thi Nga (2020) [16], digital learning materials are documents containing digitalized knowledge information content to serve teaching via computers; these materials may appear in various formats such as textual content, presentations, tabular data, audio-visual media, and even combinations of these types.

Electronic learning materials today use technological achievements to create opportunities for students to develop their own abilities, practice thinking and create two-way interactions between people and electronic devices. Learners can study anywhere, depending on their needs and personal conditions. Teachers can also organize teaching and share with those digital learning materials through knowledge information using various media. In addition, electronic learning resources can last a long time and save costs, only 25 - 30% of printed textbooks have the same amount of content, creating opportunities for learners to learn for life [3].

The 2018 general education curriculum is a new educational program of the Vietnamese Ministry of Education and Training to comprehensively develop students' abilities. The program is divided into three levels: Primary, Middle and High School, with a general program (program framework), subject programs and educational activities [5]. The general education curriculum consists of three key parts: mandatory subjects and activities, career-oriented electives, and open electives [5].

This framework transforms the broader objectives of education into real-world learning experiences. It equips students with essential knowledge while enabling them to apply their skills in daily life and future endeavours. Students develop career awareness, build positive social connections, and foster personal ethics and self-development by engaging with this structure. These outcomes promote individual fulfilment and prepare students to contribute effectively to societal progress and the advancement of humanity [5].

Fig. 1 Book connecting knowledge with Life (Kết nối Tri thức với Cuộc sống)

The general education program includes two sets of books: Connecting Knowledge with Life (Kết nối Tri thức với Cuộc sống) [17] and Creative Horizons (Chân trời Sáng tạo) [18]. Choosing the Connecting Knowledge with Life [17] book set will help ensure consistency in the content and structure of learning materials when building a data set. This book set has a tight design, which makes it easy to organize data by level and subject. Regarding the types of digital learning resources, the proposed dataset will be divided into four main types (images, tables, diagrams, and charts), and the data will be stored in image format. The data will also have text paragraphs related to the lesson for the purpose of classification (classification of subjects, classification of grade levels, etc... based on text) depending on the purpose of future research topics.

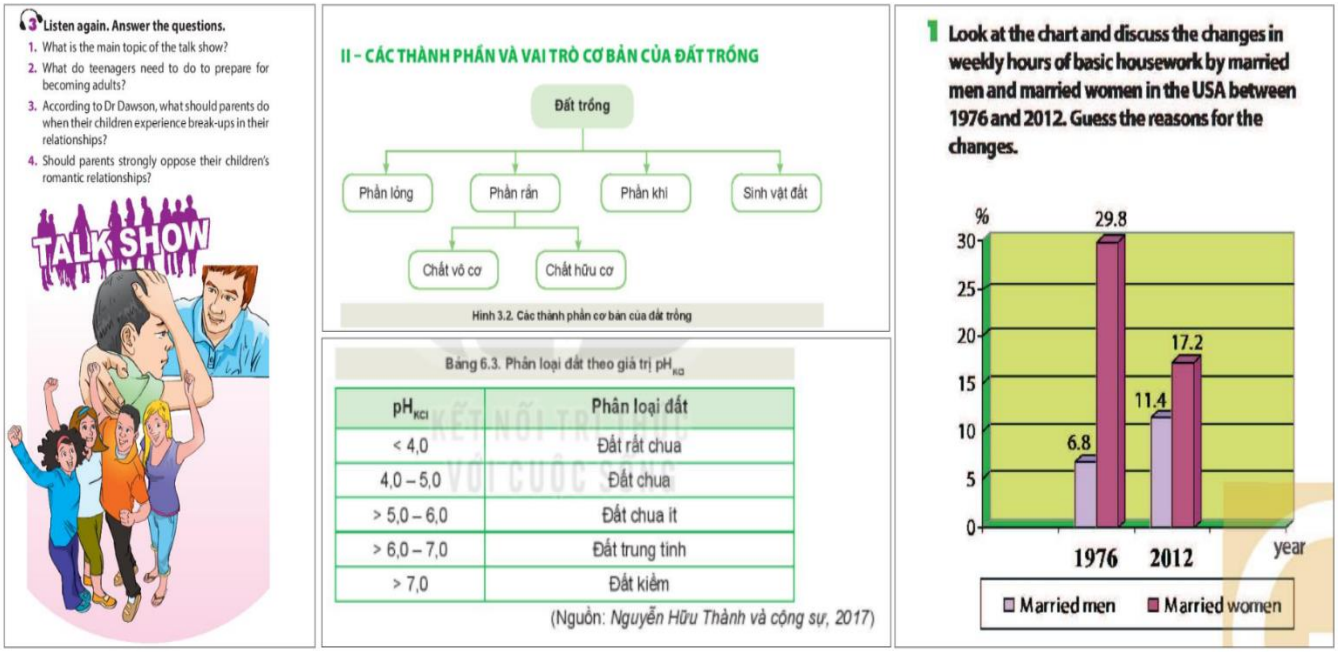


Fig. 2 Digital learning data visualization includes images, tables, diagrams and charts

Specifically, image data will be drawings or real-life photos of people or landscapes to convey lesson content. This data type is widely distributed in all subjects, making the content more vivid and understandable.

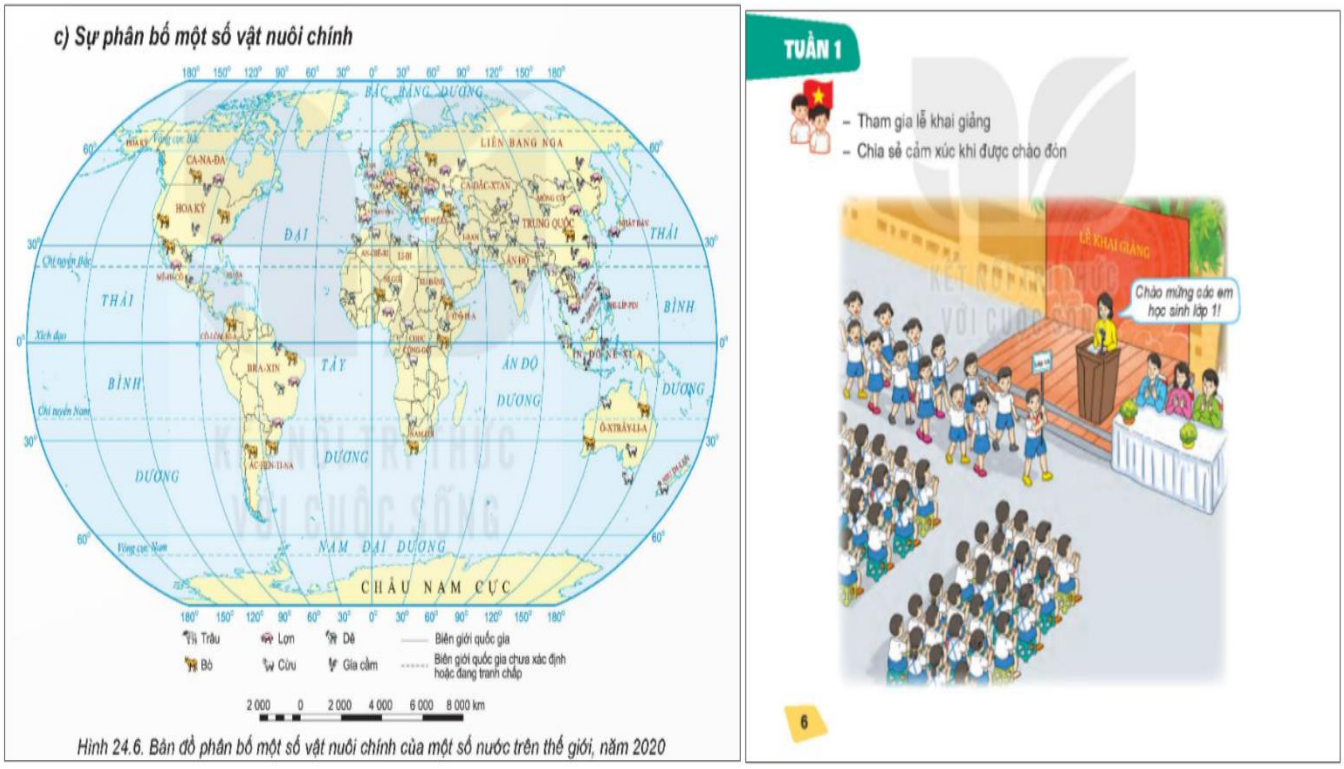
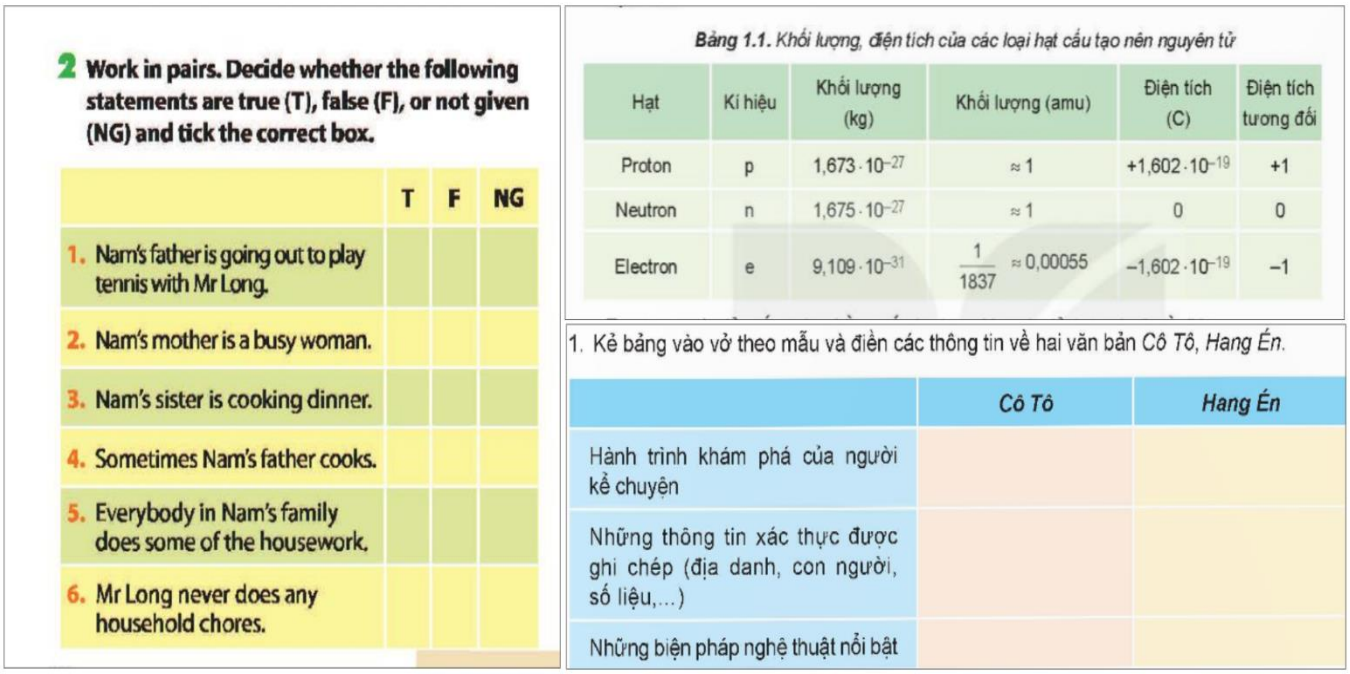


Fig. 3 Image data

A table is an arrangement of data in rows and columns. The table is widely used in research and data analysis. Based on the above criteria, tabular data will be collected into the proposed dataset.



2 **Work in pairs. Decide whether the following statements are true (T), false (F), or not given (NG) and tick the correct box.**

	T	F	NG
1. Nam's father is going out to play tennis with Mr Long.			
2. Nam's mother is a busy woman.			
3. Nam's sister is cooking dinner.			
4. Sometimes Nam's father cooks.			
5. Everybody in Nam's family does some of the housework.			
6. Mr Long never does any household chores.			

Bảng 1.1. Khối lượng, điện tích của các loại hạt cấu tạo nên nguyên tử

Hạt	Kí hiệu	Khối lượng (kg)	Khối lượng (amu)	Điện tích (C)	Điện tích tương đối
Proton	p	$1{,}673 \cdot 10^{-27}$	≈ 1	$+1{,}602 \cdot 10^{-19}$	+1
Neutron	n	$1{,}675 \cdot 10^{-27}$	≈ 1	0	0
Electron	e	$9{,}109 \cdot 10^{-31}$	$\frac{1}{1837} \approx 0{,}00055$	$-1{,}602 \cdot 10^{-19}$	−1

1. Kẻ bảng vào vở theo mẫu và điền các thông tin về hai văn bản *Cô Tô*, *Hang Én*.

	Cô Tô	*Hang Én*
Hành trình khám phá của người kể chuyện		
Những thông tin xác thực được ghi chép (địa danh, con người, số liệu,...)		
Những biện pháp nghệ thuật nổi bật		

Fig. 4 Table data

A diagram is defined as a specification of the functional requirements of an information system, including structure diagrams, behaviour diagrams or diagrams of processes, concepts, or relationships between things, capturing knowledge about the desired functionality of the system [19].

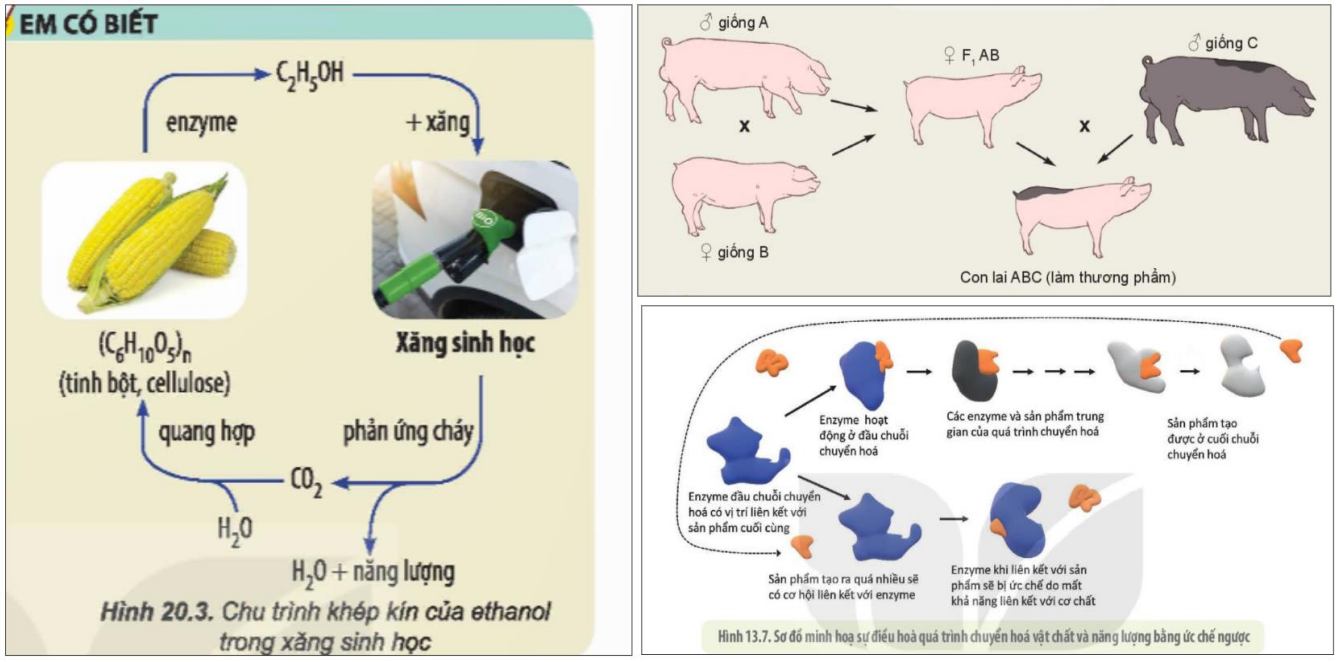


Fig. 5 Diagram data

A chart depicts numerical data or associated information through visual elements such as images, graphics, charts, graphs, or maps. Standard charts encompass bar charts, line charts, pie charts, histograms, and waterfall charts. Each type of chart is used to represent information separately [20].

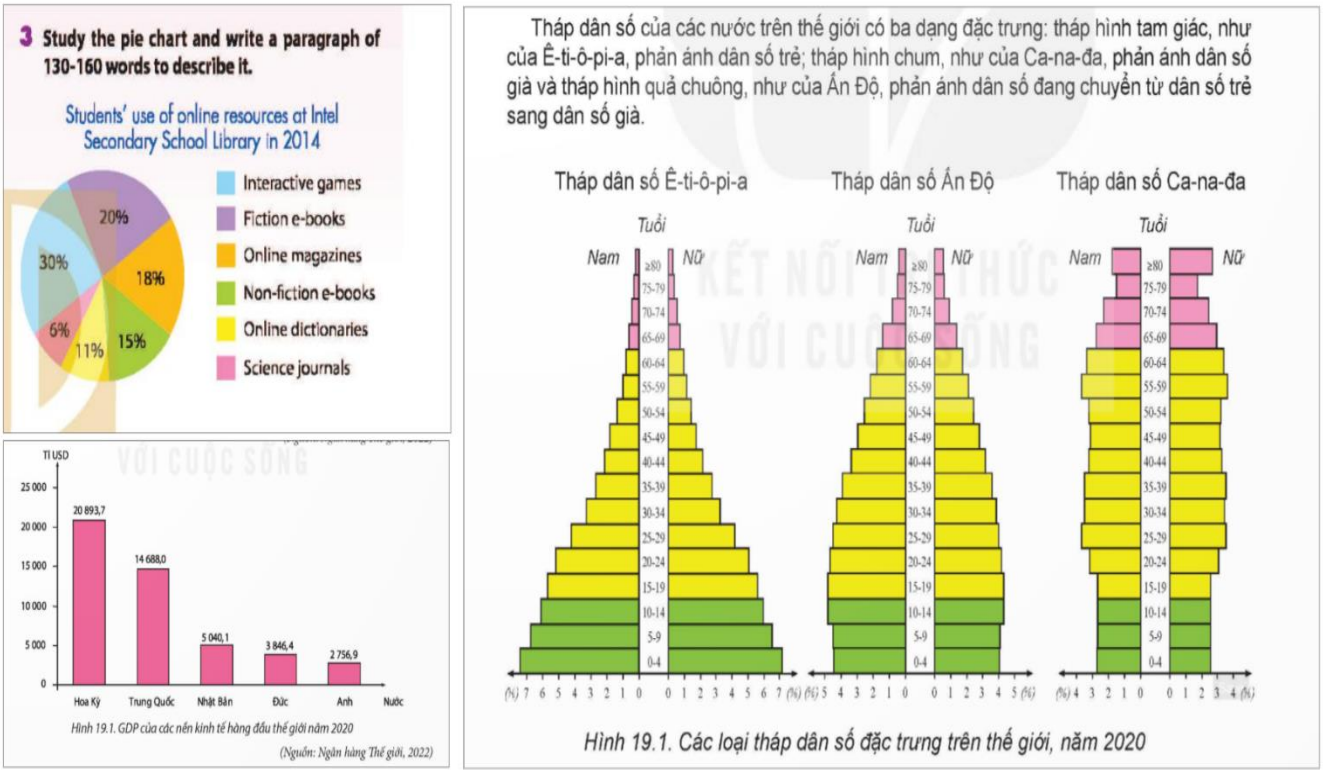


Fig. 6 Chart data

Current datasets on learning materials are also being researched and developed. SlideImages [11] and DocFigure [12] are two datasets that can meet current research needs. First, SlideImages is SlideImages collected from many sources with the task of classifying eight types of academic illustrations (charts, maps, tables, slides, ...) [11]. Although there are some errors in the classification process, the dataset still gives a good accuracy (80%) and can be fixed in the future. Besides, DocFigure has data similar to SlideImages; this dataset has twenty-eight classification classes and 33,000 extracted data images [12]. This dataset gives an accuracy higher than 90%, showing that it can serve well for future research articles. Morris et al. It is believed that the techniques applied well on DocFigure cannot work well on SlideImages because different purposes of use or development software will create different characteristics on the data, so in the research process, it is necessary to select and carefully study the techniques used in building a specific data set [11].

4. Data Description

The procedures section should provide sufficient details for other researchers to reproduce accurately. Organizing this information into clearly labelled subsections can enhance readability when describing several techniques. Based on the overall training program, the content of each subject is similar to the training content in the textbook. Most subjects are allocated with many different requirements but will be divided by topic, content (major, minor, main), and requirements to be achieved. In some non-compulsory subjects such as Music, there is only content with each fixed item, or it can also be technology with content oriented for grades 10 to 12. In addition, some subjects on career orientation at the high school level, such as Fine Arts, will be more in-depth about the major, helping students have a more certain perspective on that profession.

Some subject groups are renamed differently in each stage to suit the teaching level, such as Literature, with Vietnamese from grades 1 to 5 and Literature from grades 6 to 12, but the basic content remains unchanged.

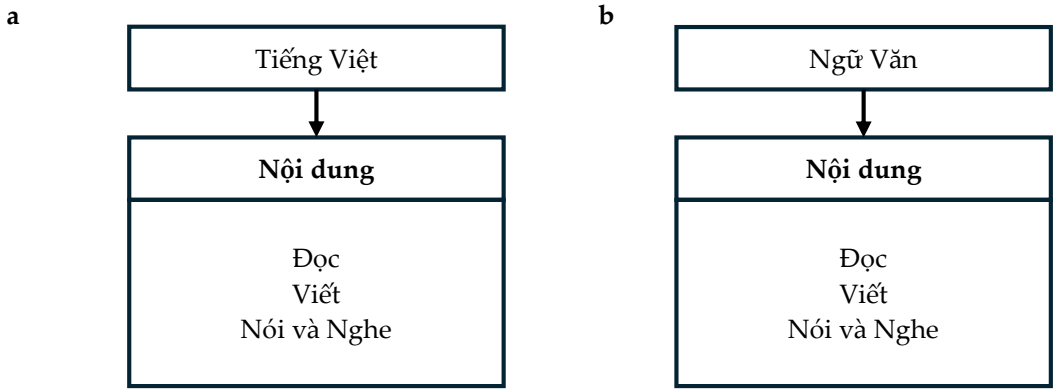


Fig. 7 Structure diagram of literature according to the curriculum (a) Grade 1..5, and (b) Grade 6..12.

Currently, with the need for international integration in Vietnam, English is gradually becoming extremely important because it is the main language used to communicate with foreign countries. This subject is one of the compulsory subjects for students at the present time. The subject structure is maintained according to diverse topics, helping learners to absorb more easily with four main skills and three language knowledge when learning a foreign language.

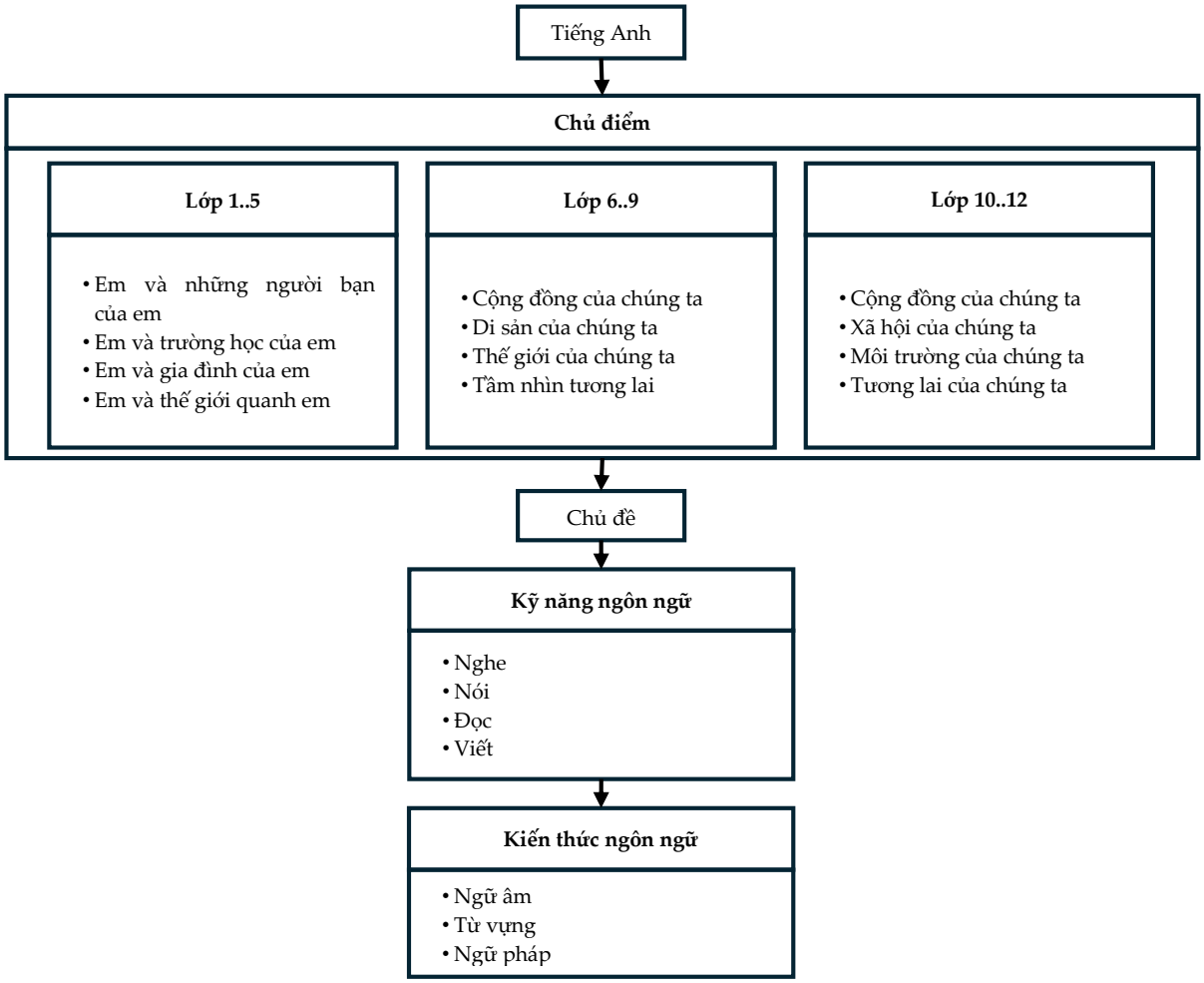


Fig. 8 Structure diagram of English according to the curriculum.

Regarding the specific structure distributed in textbooks, subjects have a similar structure but will differ in the number of lessons the subject provides. Based on the subject tree diagrams, most of these diagrams will usually include subjects, chapters, lessons, and activities, some subjects will have additional topics, sections, or other similar content. A subject may possess many designations based on the educational level within the curriculum; for instance, at the Primary level, it is referred to as Vietnamese, whereas at the Middle School and High School levels, it is termed Literature. In Vietnamese, the content is classified from practice to content type, mainly Reading, Writing, Speaking and Listening, but there will be differences in content type such as Practice, Extension, etc…

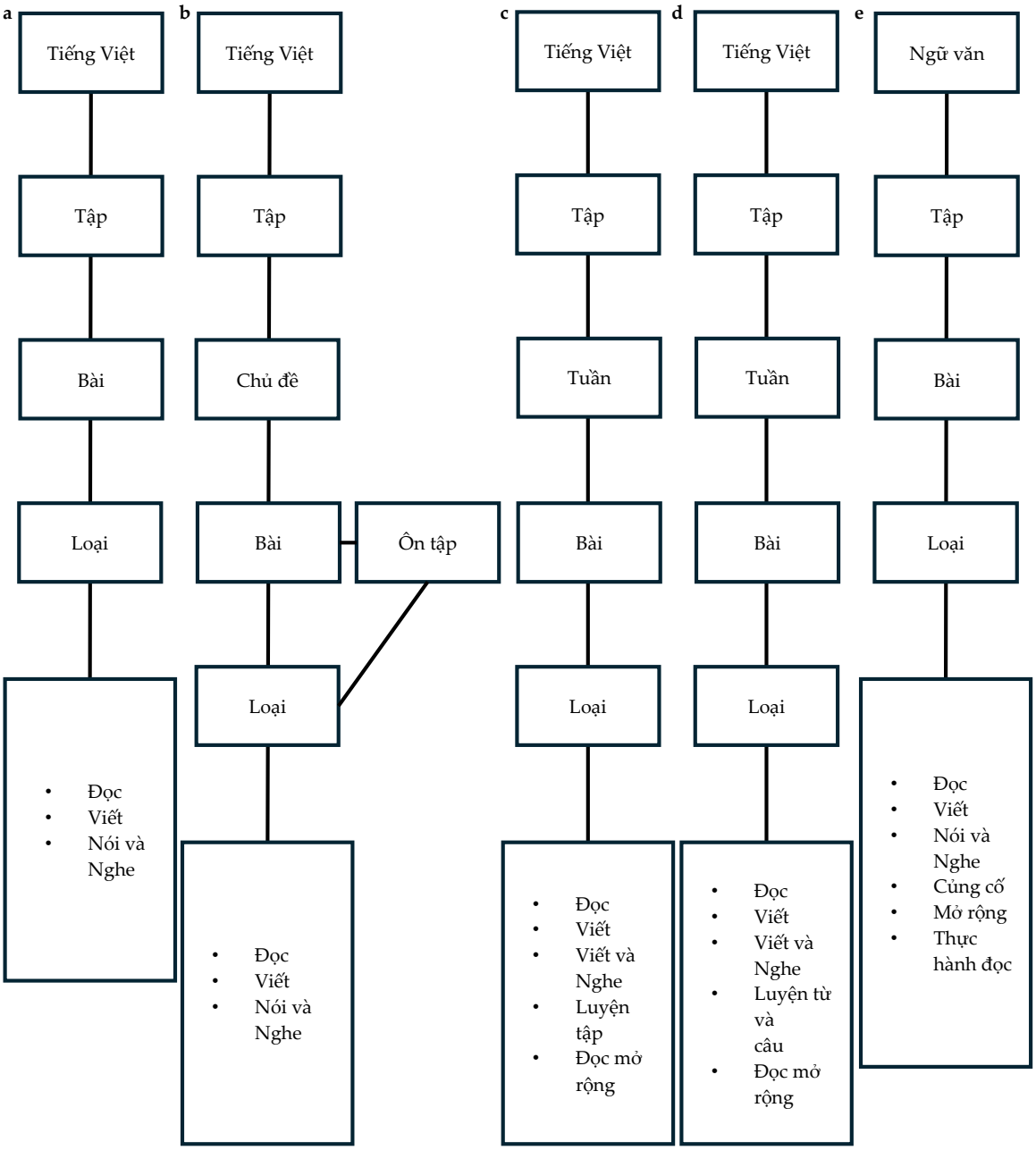


Fig. 9 Structure diagram of literature according to textbook (a) Grade 1 – Book 1, (b) Grade 1 – Book 2, (c) Grade 3..5, and (d) Grade 6..9.

Grade 10..12: For English, the subject name will not change, but the items in the lesson will be different. Grades 1 and 2 will be divided into units, including lessons, Reviews, and fun time. This subject will be divided into two sets in higher grades, and the lesson content will still be built, similar to grades 1 and 2. However, there will be no Fun time like the previous 2 grades.

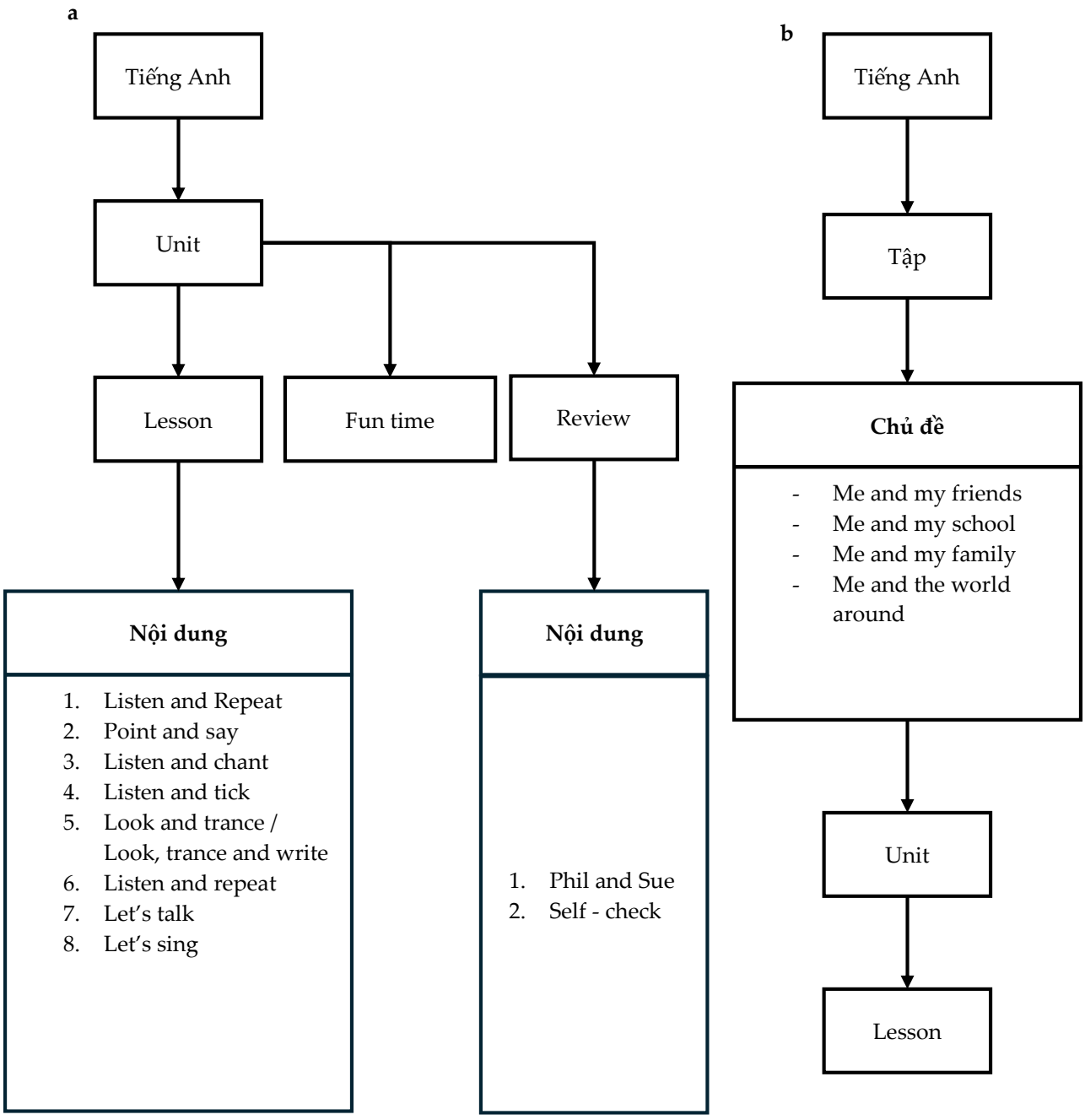


Fig. 10 Structure diagram of English from grade 1 to grade 5 according to textbook. (a) Grade 1..2. (b) Grade 3..5

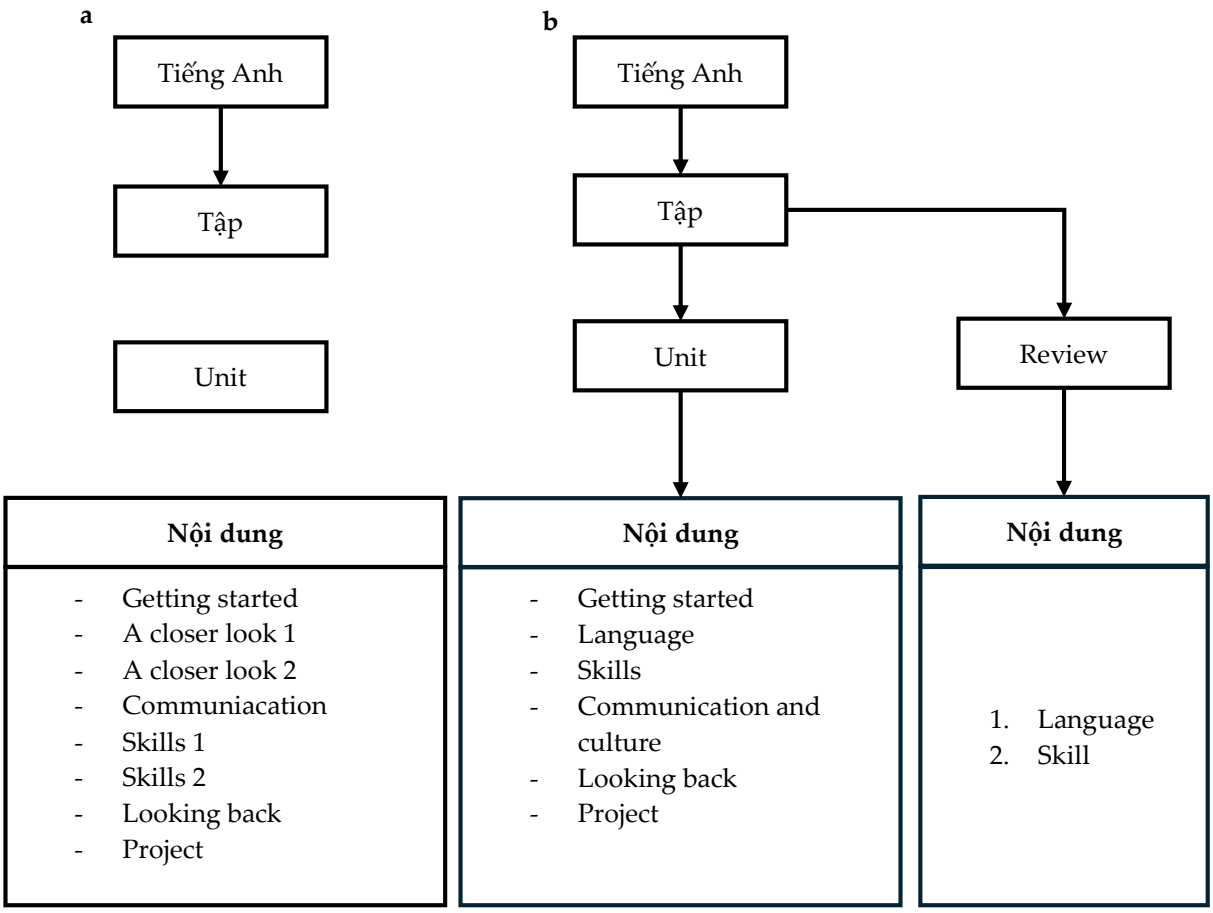


Fig. 11 Structure diagram of English from grade 6 to grade 12 according to textbook (a) Grade 6..9, and (b) Grade 10..12.

The dataset structure is built based on the official information about the program framework of [5] and is stored at https://drive.google.com/DataStructures.

5. Data Collection Method

The basic research dataset can be constructed using many methods depending on the needs and research objectives. For the study of F.A. Mbiaya, C. Vrain and F. Ros et al., 3 sub-datasets were created from the original Microsoft COCO dataset [6, 7]. This method cannot be applied to the proposed dataset because the 2018 General Education Program has only been put into practice in recent years, so previous research works up to now in Vietnam still do not have a dataset on this curriculum. However, the way F.A. Mbiaya, C. Vrain and F. Ros et al. distribute data by topics such as traffic, kitchen, etc., can be learned to apply to classifying digital learning resources.

Yang & Wang's SCB dataset on student behavior in a classroom at a university in Chengdu, China, involves manually collecting data from real videos and using the YOLOv7 algorithm for quality assessment [8]. This data can only be applied to problems with specific real-life contexts, such as classroom behavior classification. However, this manual collection method will also be more promising than F.A. Mbiaya, C. Vrain and F. Ros et al. when applied to the proposed problem.

In addition, according to Loris et al., the application of image augmentation in constructing datasets from each of VIR, BARK, GRAV and POR with accuracy above 90% shows that this is also a method that can be used for future works [9]. This collection method is an alternative for the proposed dataset in case of data shortage.

Yao et al.'s study used the main query word to expand the semantically related initial query using Google Books Ngrams Corpora (GBNC)[10]. By using words with similar meanings, the original dataset will have many images, including inaccurate images, causing data noise, so applying the CNN algorithm to filter and obtain a more optimal dataset is still necessary. This vocabulary-based data retrieval method can be used to expand the dataset with keywords related to the subject content in digital documents other than textbooks.

The research object of the topic is the 2018 General Education Program, the new education program of the Ministry of Education and Training of Vietnam, to comprehensively develop students' abilities. The program is divided into three levels: Primary, Middle and High School, with a general program (program framework), subject programs and educational activities.

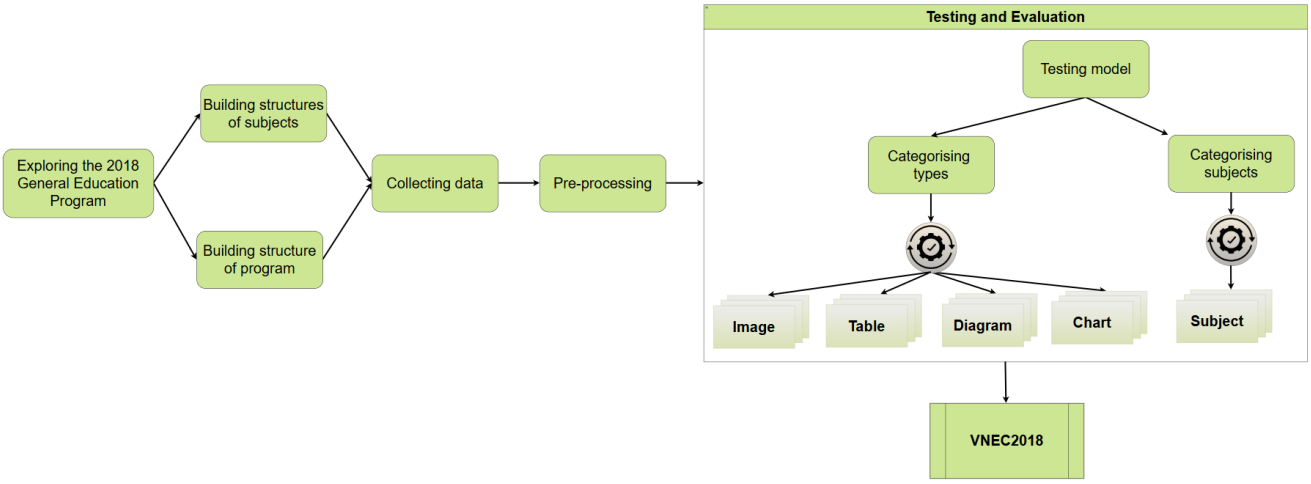


Fig. 12 Method of building and evaluating digital learning materials dataset of General Education Program 2018

Regarding the construction method, the dataset was manually collected and classified based on 4 main label objects: images, tables, diagrams, and charts. In addition, the classification is also based on the subject tree structure built from the overall education program, from which objects such as subject names, classes, subject classifications, etc., in the curriculum structure are also factors that help this dataset to develop and deploy creative research in many different directions. Moreover, the dataset is built in a multi-label direction for image objects, tables, diagrams, and charts (a data image can contain both images and tables, etc.). Therefore, this research was conducted to create a dataset that can represent the curriculum structure, contributing to serving multi-label classification problems and classification based on subject information in the training program, as well as some other purposes for future research.

In addition, to clarify the structure and increase the efficiency of the collection and identification of each criterion for the data set, the tree diagrams showing the overall curriculum and the tree diagrams of each subject by class cluster are also outlined in the most general way to serve the analysis and research of related topics later. Each data will be reviewed and recorded through screen capture, then stored as an image and sent to the pre-processing stage to remove substandard data such as duplicate data blurred images. The number of data sets is presented at: https://drive.google.com/TotalOfData.

6. Data Experimental Evaluation

This should clearly explain the main conclusions of the article. There are two common problems in evaluating the overall data set: image classification and text classification. Specifically, for the image processing object, the problem of classifying digital resources (images, tables, diagrams, charts) is a reasonable test to evaluate the quality of the data set. In addition, text data will also be evaluated through the problem of classifying subjects based on text extracted from images for processing. In both problems, a small part of the original data set. Depending on the problem's requirements and the nature of each data group, the division and use of data in both problems will be clearly considered. The reason is that for the constructed data set, the structure of the curriculum and the lessons presented in the document greatly influence the number of labels of digital resources (images, tables, diagrams, charts). Visual media often account for a large number to help convey and express content vividly and easily, while diagrams and charts account for a very low proportion because they are only suitable for some abstract subjects specific to data, calculations, etc. Therefore, if users try to use all the proposed data, the data imbalance between labels will become increasingly large, affecting the model's performance and accuracy.

In addition, subject labels are also a significant issue for text processing problems. In the curriculum, the subject name and lesson structure of each subject in different grade levels can be changed. However, the knowledge and lesson content between these grades may be related to each other. This is a major obstacle in the text-based classification problem when it can cause the model to confuse different subject labels with the same basic content of that subject. Therefore, this approach will help ensure assessment accuracy and is more suitable for each subject's characteristics.

6.1. Classifying Digital Resources

This problem is solved with the subjects: Geography 10-12, History 10-12, Technology 6-8, Physics, Science and Computer Science. After the combination of the above subjects is uploaded via CSV files, this subset of data has a total of 4077 data, including 2851 training data, 814 validation data and 412 test data. The number of image labels, tables, diagrams and charts are 3531, 771, 224 and 46, respectively. The evaluation model for this image classification problem will use the CNN [21], VGG16 [22], ResNet50 [22] and, MobileNetV2 [23], ResNet18 [24] algorithms, in which the CNN model will be built for testing. Specifically, the images will be changed to a uniform size of 380x380 (px) when processing data. The reason is that during the data collection process, it is inevitable that there will be many images, tables, etc., with different sizes, leading to the input data also having many sizes. Therefore, uniformity in size between the data helps the training model become more convenient, avoid errors and help increase the calculation speed [25]. In addition, the selected size is also in the range of 224x224 (px) and 512x512 (px) to help the processing model be more balanced in processing speed and quality of image data characteristics [26].

During training, it is considered that the amount of data between labels is not balanced (due to the nature of the learning material), which can affect the model's accuracy. Therefore, data augmentation is necessary, which helps to balance the data that accounts for a low proportion in the subset of data used [27] and can also help the model achieve higher performance [28]. After training, the output results of the models for the proposed subset of data have relatively high accuracy - all above 85%.

Table 1. Accuracy of models with the proposed dataset in the classification of digital learning resources

	Accuracy	Loss	Recall	Precision	F1-score
CNN	0.86	0.22	0.48	0.56	0.38
MobileNetV2	0.86	0.29	0.26	0.25	0.24
ResNet50	0.86	0.27	0.47	0.3	0.35
Resnet18	0.9	0.07	0.74	0.9	0.76
VGG16	0.92	0.14	0.57	0.92	0.61

6.2. Classifying Subjects

The data used for this text classification problem includes the following subjects: Geography 10-12, Computer Science, History 10-12, Physics, Chemistry, Vietnamese 4-5, English 7-9, Physical Education, Biology. After synthesis, this sub-dataset will have 9855 data, including 7095 training data, 1971 validation data and 789 test data.

The evaluation model for this text classification problem will use the algorithms BiLSTM [29], LSTM [29] and BiLSTM combined with CNN [30]. CNN is used to learn representations from characters, then combined with BiLSTM to capture the context of the two-dimensional word sequence [31].

Specifically, when uploading data to train the model, the images will be pre-processed before Optical Character Recognition (OCR), with the main steps being to convert the original image to grayscale to reduce noise, remove color information, and increase the contrast between text and background to help recognize text more effectively [32]. In addition, converting to grayscale also helps reduce the complexity of calculations. Because the number of color channels is 1 (instead of 3 RGB channels like color images), the input data size will be greatly reduced, thereby helping to increase the speed of training and processing the model [33].

Then, adjust the size of the images - zoom in if the image is too small (under 300px), and zoom out if the image is too large (larger than 1000px). This helps small images not to be too blurry or difficult to recognize and helps to increase the detail of the characters. Images with too large sizes will also reduce the calculation complexity and optimize memory without affecting the quality of extracted text [32]. Finally, after data preprocessing is completed, the image data will be extracted thanks to the optical character recognition tool EasyOCR [34], and the extracted text will be stored in a CSV file for long-term training and evaluation.

After being extracted, the raw data will be put into the training phase using the 3 models. The data will be transformed during pre-training to suit the processing model. The results are obtained after the training and evaluation process, as shown in Table 2.

Table 2. Accuracy of models with proposed dataset in subject classification.

	Accuracy	F1-score	Recall	Precision
BiLSTM	0.84	0.84	0.84	0.84
LSTM	0.83	0.82	0.83	0.83
BiLSTM - CNN	0.85	0.85	0.85	0.86

The results after training show that the accuracy of the models is 80% or higher. The BiLSTM algorithm combined with CNN gives higher results than the remaining algorithms, proving that BiLSTM initially handles text data well, but when combined with CNN, it gives better results (85%).

6.3. Discussion

Overview of the proposed dataset, the total amount of data is 30193 images, of which images account for 81.8% (25082 images), tables account for 11.9% (3654 images), diagrams account for 5.4% (1663 images), charts account for 0.8% (260 images), which shows that: for the problem of classifying digital resource images (including multi-label), the proposed dataset has a very large difference (image data accounts for 83% while charts account for less than 1%).

The reason is that objects such as diagrams and charts only appear in some subjects with specific knowledge, calculations based on data and requiring presentation in a form suitable for that subject for easy analysis, evaluation and application (for example, geography - a subject that requires data to be presented concisely, easy to see, easy to compare). At the same time, illustrations are used in most subjects, especially subjects at the primary level;

because this is an elementary level, students still do not have the ability to read, understand and analyze well. Hence, illustrations are a suitable solution for this level. Images help supplement, illustrate and explain information in a visual, vivid way combined with tables to record and arrange information and knowledge in an orderly and aesthetically pleasing way to convey knowledge effectively. That is the reason why the proportion of images and tables is larger than the other two types of data.

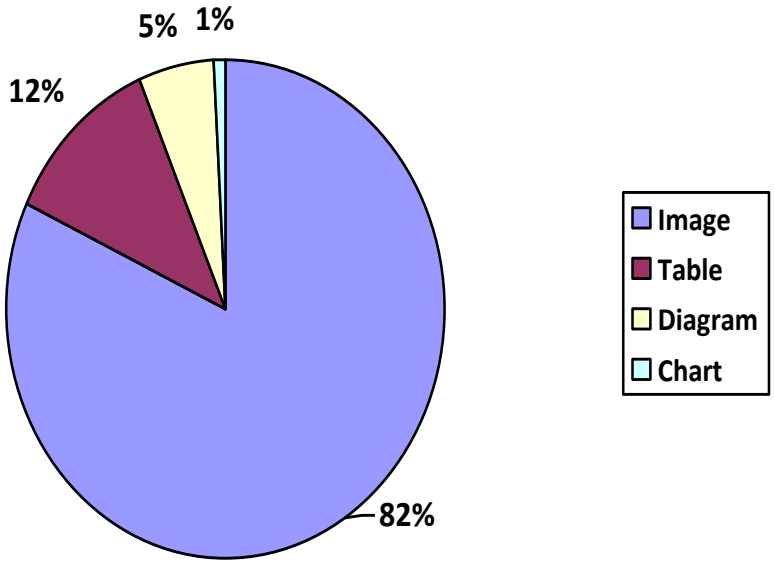


Fig. 13 Pie chart showing the distribution of data in the proposed dataset

The reason is that objects such as diagrams and charts only appear in some subjects with specific knowledge, calculations based on data and requiring presentation in a form suitable for that subject for easy analysis, evaluation and application (for example, geography - a subject that requires data to be presented concisely, easy to see, easy to compare). At the same time, illustrations are used in most subjects, especially subjects at the primary level; because this is an elementary level, students still do not have the ability to read, understand and analyze well. Hence, illustrations are a suitable solution for this level. Images help supplement, illustrate and explain information in a visual, vivid way combined with tables to record and arrange information and knowledge in an orderly and aesthetically pleasing way to convey knowledge effectively. That is the reason why the proportion of images and tables is larger than the other two types of data.

To address the problem of data imbalance, this study uses a subset method from the original dataset. Specifically, due to the difference between the classification labels, some subjects will be selected and grouped so that the ratio between the labels becomes more balanced. Since the original data cannot be completely randomly divided, this process focuses on prioritizing subjects containing diagrams and charts to correct the imbalance with image and table data types, which are dominant in the dataset. However, this only somewhat fixes it; data augmentation techniques can help produce more efficient results. To augment the data for the small class, the training dataset is processed using this technique to rotate, flip, change color, or add noise.

In terms of data quantity, although VNEC2018 is an image dataset with an average number like DocFigure [12] (33,000 images), it is extremely large when compared to the SlideImages dataset [11] (more than 3,000 images). Building a dataset with a large or small number depends on the research needs and the learning resources where the dataset is collected. For example, if the proposed dataset is collected only from the textbooks provided by the Ministry of Education and Training, the initial quantity of data will be approximately 30,000. In the future, if more

data from other sources besides this textbook is added, the data will certainly be more diverse, and the accuracy of the training process will also increase.

Table 3. Comparison table of accuracy of datasets in image classification problem table

	Model	Accuracy
SlideImages test	MobileNetV2	80%
DocFigure test	CNN	~90%
VNEC2018 test	MobileNetV2	86%
VNEC2018 test	CNN	86%

When analyzing the image classification problem on the SlideImages, DocFigure and VNEC2018 datasets, the results show a clear difference between the datasets and the models used. Specifically, the SlideImages dataset, when combined with the MobileNetV2 model, gives about 80% accuracy, while DocFigure, with the CNN model, achieves an accuracy of up to 90%. In particular, the VNEC2018 dataset developed by the research team achieves 86% accuracy when applying both the MobileNetV2 and CNN models, which shows a stable and encouraging performance. However, VNEC2018 only uses four classification labels, much less than SlideImages (9) and DocFigure (28).

This difference can be explained mainly by the way the classification labels of each dataset are designed. While VNEC2018 groups together chart types (bar, line, pie charts, etc...) and diagrams (processes, cycles, etc.) into a single label, SlideImages and DocFigure divide these types into many separate labels, allowing the model to learn more detailed and accurate features. In addition, VNEC2018 also encounters the problem of label imbalance, when image data accounts for a large proportion compared to the remaining labels, such as tables, diagrams, and charts. This makes it difficult for the model to learn the features of less frequent label types, leading to uneven classification results between groups.

However, one of the strengths of the VNEC2018 dataset is its stability and good generalization ability when applying different models, although the number of classification labels is somewhat smaller. This is thanks to the clear data structure and the close connection with the 2018 Vietnamese general education program, helping the deep learning model to identify features that are highly practical and close to the educational context of students. In addition, the fact that the problem is designed as a multi-label classification - allowing an image to belong to multiple groups at the same time - is suitable for current reality, when learning materials often include many different visual elements on the same page. This makes VNEC2018 highly applicable and opens up great opportunities for integrating this dataset into intelligent learning systems.

Beyond dataset comparisons, the study further evaluated the proposed dataset across multiple deep learning architectures to investigate each model's performance and generalizability in educational image classification. As shown in Table 2, the accuracy of the models used in this study varies significantly. Specifically, the three models, VGG16, ResNet18, and ResNet50, achieve 92%, 90%, and 86% accuracy, respectively. To explain this difference, we first compare VGG16 and ResNet50, two architectures with distinct differences in design and feature extraction. VGG16 has a sequential architecture in which convolutional layers are arranged in series in a simple but effective structure. The moderate depth of the network helps this model learn well on datasets that are not too large or have clear features. This explains why VGG16 achieves higher accuracy on the current dataset.

In contrast, ResNet50 has a significantly deeper design with residual blocks, an important component that helps the model learn more complex features [35]. However, in cases where the dataset is not large enough (only 4,077 images) and has a high label imbalance, ResNet50 may have difficulty generalizing, leading to overfitting and worse results than VGG16 [36]. Research by H. Kamal et al. also shows that ResNet50 does not perform as well as

VGG16 on small or imbalanced datasets when applied to medical images. Additionally, VGG16 may be more suitable in certain cases depending on the dataset's characteristics. Another study compared the performance of VGG16 and ResNet50 in lung disease classification and found that VGG16 achieved higher accuracy, suggesting that this architecture can better exploit important features of medical data [37].

A notable point in the experimental results is that ResNet18 achieves 90% accuracy, almost equivalent to VGG16 (92%) and significantly higher than ResNet50 (86%). This raises the question of why ResNet18 performs better than ResNet50 and even approaches VGG16, although they are both in the ResNet family. One of the important reasons lies in the depth of the model. ResNet18 has a shallower architecture than ResNet50, consisting of only 18 layers instead of 50 layers, which helps this model reduce complexity, avoid overfitting, and generalize better on small or imbalanced datasets. Deep models like ResNet50 can overlearn the training sample when the data is not rich enough, leading to poor generalization on the test set [38].

Furthermore, despite having fewer layers, ResNet18 still uses residual blocks, which preserve information and improve gradient propagation during training. Therefore, ResNet18 retains the important benefits of ResNet without the problems caused by too much depth like ResNet50. A comparative study on medical images also shows that when the data is small and imbalanced, ResNet18 often outperforms ResNet50 because the model is not too complex but still powerful enough to extract useful features [36]. In addition, VGG16 and ResNet18 have similar numbers of parameters, making the complexity of the two models not too different. This helps ResNet18 perform similarly to VGG16 on the current dataset. Meanwhile, ResNet50 has a significantly larger number of parameters, requiring more data to reach its full potential. This model may not achieve the expected performance when the data is insufficient or imbalanced.

Therefore, the image classification task using ResNet18 results close to VGG16 due to its moderate depth and better generalization ability than ResNet50 while still retaining the advantage of the residual block without being affected by overfitting. This shows that in problems with small and imbalanced data, a moderate model like ResNet18 may be a better choice than deeper networks like ResNet50.

In addition, the problem of classifying digital learning resources has a significant difference between the Precision, Recall, F1-score, and Accuracy indexes. In contrast, the problem of classifying subjects using text has a more even distribution of these indexes (ranging from 80-85%). The main reason comes from the difference between the two problems. Specifically, the problem of classifying subjects focuses on analyzing text and detailed content in image data to identify the corresponding subjects in the educational program. Meanwhile, the problem of classifying digital resources requires the model to analyze the characteristics of images, tables, diagrams, or charts more deeply to accurately identify each type of resource, leading to an uneven distribution between the evaluation indexes.

Secondly, compared with the image classification problem, the subject classification problem has a significantly better label balance. To ensure balanced representation during training, the subject data is meticulously prepared to represent each class equally. At the same time, content control helps minimize the similarity between subjects, limit confusion, and facilitate the machine learning model to identify each subject's unique features clearly.

In general, VNEC2018 has a major drawback: the data is unbalanced due to the uneven distribution of data in the labels. However, compared to the two datasets DocFigure and SlideImages, VNEC2018 is collected on textbooks used for the curriculum in a country, so it will focus more on electronic learning materials, helping to build smart educational tools such as lesson suggestions, competency assessments or supporting teachers to organize lectures based on knowledge in textbooks, while the two previously released datasets have advantages in image classification, data types, and research support.

In summary, the subject classification problem is mainly based on academic content with a higher level of data balance. In contrast, the digital resource classification problem requires a detailed analysis of the image feature structure and is imbalanced between labels. These are the important factors leading to the difference in performance between the two problems. From this, for the text classification problem, the proposed dataset is more suitable than the digital learning resource image classification problem based on the label balance criterion in the selected subsets.

In the future, we will overcome this shortcoming by searching for some data from other sources within the framework of the 2018 general education program, such as the Creative Horizon book series [17] combined with reference materials, online tests and MOOC courses to balance the image objects and the remaining 3 types of data (tables, diagrams and charts). Integrating more data on multimedia platforms (lecture videos, learning applications, etc.) also helps the dataset have more characteristics, diversifying the dataset and improving the accuracy in classification and content recognition, increasing the practical applicability of systems in the digital environment. In addition, it is also planned to add some specific types of labels instead of just the current general labels, such as human images, landscape images, life cycle diagrams, pie charts, bar charts, etc., to increase the specificity and diversity of information for the proposed dataset.

7. Conclusion

Based on the research, we have presented the 2018 Vietnamese general education curriculum and associated ideas for digital learning resources. Simultaneously, the VNEC2018 digital learning material picture dataset was also proposed, with 30,193 photos: the construction structure's characteristics and the quantity of each dataset component. The research team simultaneously extracted two sub-datasets and applied them to two distinct classification problems to assess the dataset's quality. The CNN, VGG16, ResNet50, ResNet18, and MobileNetV2 models were employed to solve the image classification problem. The models' accuracy ranged from 86 to 92%, with the VGG16 model achieving the best result (92%). Furthermore, three models categorize text retrieved from images: BiLSTM, LSTM, and BiLSTM-CNN. The accuracy ranges from 83% to 85%, with the BiLSTM-CNN model achieving the best result (85%).

Acknowledgments

We would like to thank Ho Chi Minh City University of Education for supporting this research. We truly appreciate the resources and academic environment provided, which allowed us to focus on our work without unnecessary difficulties.

References

[1] Neha Gupta, “Chapter One - Introduction to Hardware Accelerator Systems for Artificial Intelligence and Machine Learning,” *Advances in Computers*, vol. 122, pp. 1-21, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] Huu-Tai Thai, “Machine Learning for Structural Engineering: A State-of-the-Art Review,” *Structures*, vol. 38, pp. 448-491, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Trinh Le Hong Phuong, ” Building digitized lear's resourcees to enhance the teachat and learning of some contents of chemistry in high school, *Ho Chi Minh City University of Education Journal of Science, (8)*, 178-187. [Online]. Available: https://journal.hcmue.edu.vn/index.php/hcmuejos/article/view/1803/1792.

[4] Ministry of Education and Training, *Decision No. 3784/QD-BGDDT* on guidelines for developing digital learning materials on MOOC platform application, Law Library, 2022. [Online]. Available: https://thuvienphapluat.vn/van-ban/Cong-nghe-thong-tin/Quyet-dinh-3784-QD-BGDDT-2022-Huong-dan-xay-dung-hoc-lieu-so-tren-ung-dung-MOOCs-609186.aspx

[5] Ministry of Education and Training, General Curriculum of General Education Program (Circular No. 32/2018/TT-BGDĐT), Socialist Republic of Vietnam, 2018. [Online]. Available: https://moet.gov.vn/content/tintuc/Lists/News/Attachments/8421/chuong-trinh-tong-the-ctgdpt-2018.pdf.

[6] Franck Anaël Mbiaya et al., "Dataset for Image Classification with Knowledge," *Data in Brief*, vol. 57, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Tsung-Yi Lin et al., "Microsoft COCO: Common Objects in Context," Computer vision - ECCV 2014, 13th European Conference, Zurich, Switzerland, pp. 740-755, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[8] Yazhou Yao et al., "A New Web-Supervised Method for Image Dataset Constructions," *Neurocomputing,*, vol. 236, pp. 23-31, *2017*. [CrossRef] [Google Scholar] [Publisher Link]

[9] Loris Nanni et al., "Comparison of Different Image Data Augmentation Approaches," *Journal of Imaging*, vol. 7, no. 12, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Yazhou Yao et al., " Exploiting Web Images for Dataset Construction: A Domain Robust Approach," *IEEE Transactions on Multimedia*, vol. 19, no. 8, pp. 1771-1784, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[11] David Morris, Eric Müller-Budack, and Ralph Ewerth, "Slideimages: A Dataset for Educational Image Classification," *Advances in Information Retrieval*, *42nd European Conference on IR Research*, Lisbon, Portugal, pp. 289-296, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] K.V. Jobin, Ajoy Mondal, and C.V. Jawahar, "DocFigure: A Dataset for Scientific Document Figure Classification," *IEEE International Conference on Document Analysis and Recognition Workshops*, Sydney, NSW, Australia, pp. 74-79, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[13] Ministry of Education and Training, *Circular No. 15/2017/TT-BGDDT Amending Regulations on Working Regime for High School Teachers*, Law Library, 2017. [Online]. Available: https://thuvienphapluat.vn/van-ban/Bo-may-hanh-chinh/Thong-tu-15-2017-TT-BGDDT-sua-doi-Quy-dinh-che-do-lam-viec-doi-voi-giao-vien-pho-thong-341252.aspx

[14] L. Parrott and R. Kok, "Design and Development of Multimedia Courseware: An Overview," *Canadian Society for Bioengineering*, vol. 39, no. 2, pp. 131-137, 1997. [Google Scholar] [Publisher Link]

[15] Jan van der Akker, Paul Keursten, and Tjeerd Plomp, "The Integration of Computer Use in Education," *International Journal of Educational Research,* vol. 17, no. 1, pp. 65-76, 1992. [CrossRef] [Google Scholar] [Publisher Link]

[16] Rahul Shrivastava, Yogendra Kumar Jain, and Ajay Kumar Sachan, "Designing and Developing e-Learning Solution: Study on Moodle 2.0," *International Journal of Machine Learning*, vol. 3, no. 3, pp. 305-308, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[17] Connecting knowledge with life [Digital textbooks], Vietnam Education Publishing House, Hanoi, 2020. [Online]. Available: https://hanhtrangso.nxbgd.vn/sach-dien-tu?book_active=0&classes=1

[18] *Creative horizons* [Digital textbooks], Vietnam Education Publishing House, Hanoi, 2020. [Online]. Available: https://hanhtrangso.nxbgd.vn/sach-dien-tu?book_active=1&classes=1

[19] Janis Osis, and Uldis Donins, "Software Designing With Unified Modeling Language Driven Approaches," *Topological UML Modeling,* pp. 53-82, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[20] Stephen M. Kosslyn, "Understanding Charts and Graphs," *Applied Cognitive Psychology*, vol. 3, no. 3, pp. 185-225, 1989. [CrossRef] [Google Scholar] [Publisher Link]

[21] Leila Mohammadpour et al., "A Survey of CNN-based Network Intrusion Detection," *Applied Sciences*, vol. 12, no. 16, pp. 1-34, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[22] Chanthini Baskar et al., "Computer Graphic and Photographic Image Classification using Transfer Learning Approach," *Signal Processing,* vol. 39, no. 4, pp. 1267-1273, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[23] Ke Dong et al., "MobileNetV2 Model for Image Classification," *IEEE 2nd International Conference on Information Technology and Computer Application*, Guangzhou, China, pp. 476-480, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[24] PyTorch, ResNet18 - Torchvision main documentation. 2025, [Online]. Available: https://pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html#torchvision.models.resnet18

[25] Carl F. Sabottke, and Bradley M. Spieler, "The Effect of Image Resolution on Deep Learning in Radiography," *Radiology: Artificial Intelligence*, vol. 2, no. 1, pp. 1-7, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[26] Mingxing Tan, and Quoc Le, "Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks," *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, pp. 6105-6114, 2019. [Google Scholar] [Publisher Link]

[27] Guanghan Ning et al., "Data Augmentation for Object Detection via Differentiable Neural Rendering," *arXiv*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[28] Connor Shorten, and Taghi M. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1-48, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[29] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin, "The performance of LSTM and BILSTM in Forecasting Time Series," *IEEE International Conference on Big Data*, Los Angeles, CA, USA, pp. 3285-3292, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[30] Zhiheng Huang, Wei Xu, and Kai Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging," *arXiv*, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[31] Yoon Kim et al., "Character-Aware Neural Language Models," In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, pp. 2741- 2749, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[32] Abdeslam El Harraj, and Naoufal Raissouni, "OCR Accuracy Improvement on Document Images through a Novel Pre-Processing Approach," *arXiv*, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[33] OpenCV, *Color Conversions*, 2025. [Online]. Available: https://docs.opencv.org/3.4/de/d25/imgproc_color_conversions.html

[34] JaidedAI/EasyOCR, 2023. [Online]. Available: https://github.com/JaidedAI/EasyOCR

[35] Zhang A. et al., Residual Networks (ResNet) and Model Design, Dive into Deep Learning, 2013. [Online]. Available: https://vi.d2l.ai/chapter_convolutional-modern/resnet.html

[36] Kamal Kamal, and Hamid Ez-Zahraouy, "A Comparison between the VGG16, VGG19 and RESNET50 Architecture Frameworks for Classification of Normal and CLAHE Processed Medical Images," 2023. [CrossRef] [Google Scholar] [Publisher Link]

[37] Ngo Huu Huy et al., "Comparative Performance of Resnet50 and Vgg16 in Lung Infection Detection," *Advances in Information and Communication Technology, Proceedings of the 3rd International Conference*, pp. 733-744, 2025. *[CrossRef] [Google Scholar] [Publisher Link]*

[38] Kaiming He et al., "Deep Residual Learning for Image Recognition," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016. [Google Scholar] [Publisher Link]