Scientific Research Publishing

World Journal of Engineering and Technology, 2026, 14(1), 103-123
https://www.scirp.org/journal/wjet
ISSN Online: 2331-4249
ISSN Print: 2331-4222

Multimodal AI Application for Vietnamese Digital Learning Material Classification

Giang Ma[1,2], Quoc Nguyen[2], Hai Tran[1]

[1]Department of Information Technology, Ho Chi Minh University of Education, Ho Chi Minh City, Vietnam
[2]Department of Information Technology, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam
Email: giangmn@hcmue.edu.vn, giangmninfo@gmail.com

How to cite this paper: Ma, G., Nguyen, Q. and Tran, H. (2026) Multimodal AI Application for Vietnamese Digital Learning Material Classification. *World Journal of Engineering and Technology*, **14**, 103-123.
https://doi.org/10.4236/wjet.2026.141006

Received: October 13, 2025
Accepted: December 28, 2025
Published: December 31, 2025

Abstract

This study proposes a multimodal AI model for classifying Vietnamese digital learning materials by integrating three key information sources: text content, image and graphic features, and document layout structures. The model is designed with a dual-branch architecture in which Vietnamese transformer models (BERT, PhoBERT) process textual information, while convolutional neural networks extract visual features from document images. A hybrid fusion mechanism combines multimodal representations at both intermediate and prediction levels to enhance the robustness of the classification process. Based on theoretical foundations and evidence from international multimodal research, this model is expected to outperform single-modal approaches, particularly when applied to visually complex learning materials such as slides, diagrams, and documents with diverse layouts. The proposed framework contributes conceptually to the development of multimodal learning material classification tailored to Vietnamese characteristics and offers potential practical value for automating classification, improving search and recommendation functions, and supporting digital transformation in Vietnam's higher education context.

Keywords

Multimodal Artificial Intelligence, Digital Learning Material Classification, Deep Learning, Transformer, Digital Transformation of Education

1. Introduction

In recent years, digital education in Vietnam has developed strongly under the impact of the national digital transformation policy. The Ministry of Education and Training has issued the Digital Transformation Strategy for the Education sector for the period 2023-2025, with a vision to 2030, to promote the application

of information technology and artificial intelligence in teaching, administration and provision of learning materials. This policy has created conditions for the emergence of a series of learning management systems (LMS), digital libraries, and open learning repositories, helping students and lecturers access rich and diverse learning resources. Digital learning materials are no longer just pure text (e-books, PDF documents) but also include lecture slides, teaching videos, illustrations, diagrams, charts and multi-modal documents with complex layouts such as scientific articles, textbooks with tables and specialized symbols. Managing, organizing and retrieving these resources is becoming an urgent requirement to improve the effectiveness of teaching and learning in the digital age.

However, current digital learning material classification systems still have many limitations. Most systems only process a single method—usually text—due to the availability of quite developed Vietnamese analysis tools, for example, news topic classification or sentiment extraction from student feedback. Studies such as Phuc *et al*. (2013) used LDA to classify Vietnamese web documents but only stopped at pure text [1].

When learning materials combine multiple methods (multimodal), they have complex layouts, and information from images, charts, and page layouts is often omitted or processed asynchronously, leading to low classification accuracy and limited ability to suggest learning materials according to content/topic/level.

Several international works have shown the potential of multimodal approaches. Canhui *et al*. (2023) proposed a hierarchical multimodal neural network for document layout analysis, showing that combining image information and page layout significantly improves classification performance compared to text alone [2]. Liu *et al*. (2024) also presented the HPMT model—a hierarchical multimodal transformer architecture for long document classification [3]. Abdallah *et al*. (2024) showed that combining OCR and layout analysis using deep learning can improve the accuracy of content recognition of scanned documents [4]. These results indicate a trend of shifting from single-modal to multimodal analysis internationally.

In the context of Vietnam, the research of Nguyen Khang *et al*. (2022) has built a dataset and benchmark for text recognition and layout detection of Vietnamese documents, providing an important foundation for layout analysis systems [5]. Scius-Bertrand *et al*. (2019) presented a method for text layout analysis and columnarization for ancient Vietnamese stele inscriptions, demonstrating the feasibility of Vietnamese-specific layout processing techniques [6]. Longer studies such as Mai and Toan (2012) applying Hausdorff distance for page layout analysis also laid the foundation for this approach, although it has not yet handled modern multimodal learning materials [7].

In addition, the classification of heterogeneous educational data has been mentioned by Phuc and Chau (2021), emphasizing that different data sources (forum posts, student feedback, electronic documents) need to be integrated for course-level classification [8]. At the same time, a new generation of multimodal large

language models (MLLMs) is emerging in the world. Yin *et al*. (2024) conducted a comprehensive review of MLLMs, showing that integrating information from multiple modalities (text, image, layout) effectively improves complex document comprehension tasks [9].

From the above studies, there is a research gap in the context of Vietnam: there is no specialized model that fully integrates the three components—text, images/graphics, and document layout—for the task of classifying digital learning materials for education. There is no Vietnamese dataset that is fully annotated for illustrations, tables, page layout positions, and visual quality, nor is there an empirical evaluation comparing single-modal and multi-modal methods in the Vietnamese environment with real data from schools and digital libraries. The problem is also complicated by technical challenges such as OCR of Vietnamese with accents, word segmentation, processing of charts and diagrams, and incorporating contextual information from images into determining learning material topics.

Therefore, the objective of this study is to propose a specialized multimodal artificial intelligence model for digital learning materials classification in the context of digital education transformation in Vietnam. The model is designed to combine and effectively exploit three components: text content (preprocessed appropriately for Vietnamese), illustrative images/graphics (charts, diagrams) and page layout features (page partitioning, column division, content area location, captions, titles). The specific objectives include: (i) building or customizing a Vietnamese dataset with full annotations for both text, images and layout; (ii) developing a multimodal digital learning materials classification model with higher accuracy than single-modal methods; (iii) experimental evaluation in real-world environments such as digital libraries, online learning systems, universities; and (iv) propose implementation solutions to support searching, suggesting learning materials, organizing learning material warehouses by topic/usage level, suitable for digital transformation of education in Vietnam. The expected results will have clear scientific significance in multi-modal machine learning, document image analysis and layout analysis; at the same time, have high practical value for managing digital learning materials, improving the ability to retrieve and use learning materials, supporting lecturers and students to quickly access appropriate documents, contributing to promoting the goal of digital transformation in the country's education sector.

2. Related Work

Digital learning materials, also known as electronic learning materials (Digital Learning Resources), are a collection of digitized learning resources that can be accessed, stored, shared, or edited through electronic devices. According to the definition of IEEE Learning Object Metadata (LOM), digital learning materials contain not only content but also metadata that describes the characteristics, purpose, and usage [10]. They serve as essential electronic media for teaching and

learning. One of the outstanding features of digital learning materials is their high flexibility and adaptability. Unlike traditional printed materials that require costly and time-consuming reprinting when changes are made, digital learning materials can be easily edited and updated instantly. This allows educators to respond quickly to changes in the curriculum, new information or student feedback. Furthermore, digital learning materials support multimedia, provide automatic feedback and allow for personalized learning, which is difficult to achieve with static materials. The ability to easily integrate with advanced technologies such as AI and Learning Management Systems (LMS) is also a key advantage, opening the door to deeper learning data analytics and optimized user experiences.

Over the past decade, along with the strong development wave of multimodal artificial intelligence in the world, the domestic research community has witnessed a significant increase in the quantity and quality of multimodal deep learning applications in processing, analyzing and classifying Vietnamese digital documents, especially digital learning materials that simultaneously contain text, image and layout information. One of the outstanding contributions is the research with the introduction of the OpenViVQA dataset—the first Visual Question Answering (VQA) dataset dedicated to the Vietnamese language, including tens of thousands of hand-crafted question-answer pairs associated with images reflecting the context and context of Vietnam [11]. To exploit this dataset, the authors propose three multimodal feature fusion models: FST (Feature Summation Transformer), QuMLAG (Question-guided Multi-Level Attention Graph) and MLPAG (Multi-Layer Perceptron Attention Graph), each model represents a different strategy for integrating visual and textual information to optimize question answering performance. Experimental results show that these multimodal models significantly outperform single-modal methods, confirming the advantage of simultaneously exploiting visual and linguistic data in tasks requiring complex inference from digital learning materials [11].

Following this research direction, the ViOCRVQA dataset was built, focusing on the problem of understanding and exploiting Vietnamese text information in images through OCR (Optical Character Recognition) technique [12]. The dataset consists of 28,282 images and 123,781 question-answer pairs, where each image contains text and each question is associated with the text content in the image. The images are mainly book covers, advertising/information images containing metadata such as title, author, publisher, genre, translator. The highlight of the research is the VisionReader model, an architecture that combines OCR components to extract text in images, visual features from the original image and layout/feature information related to objects in the image. On the test set, VisionReader achieved Exact Match (EM) = 0.4116 and F1-score = 0.6990. The results demonstrate that fusing three feature sources—OCR text, visual features, and object-layout features—improves contextual understanding in Vietnamese documents with diverse formats and contents.

Another important step forward in the Vietnamese VQA field is the develop-

ment of the ViTextVQA dataset. Unlike OpenViVQA, which focuses on general images, ViTextVQA emphasizes the model's ability to understand text appearing in images—such as titles, captions, charts, or information printed on objects—in the Vietnamese context. This dataset includes more than 16,000 images and 50,000 questions and answers, meticulously annotated to reflect diverse text types in life and education. This research also clarifies the role of OCR token order and input token selection strategy in the model's performance, thereby providing optimization directions for multimodal AI systems processing the Vietnamese language.

In addition to VQA datasets, document layout analysis and mining have also received great attention. The VNDoc dataset—a collection of administrative, educational, and commercial documents popular in Vietnam, is labeled not only by text content but also by layout and presentation structure [13]. This is especially useful for OCR and document classification systems, as layout plays an important role in identifying document types, understanding relationships between elements, and supporting accurate information extraction. By providing this dataset, the authors have enabled AI models to learn both the visual and structural aspects of documents—a key factor in classifying complex digital learning materials [13].

Not only stopping at educational and administrative documents, but Vietnamese researchers also expanded the application to the medical field, which has many characteristics and challenges in handwriting recognition, non-standard formats and high accuracy requirements. Ly *et al.* (2025) proposed a method for extracting information from Vietnamese medical prescriptions based on combining OCR with deep learning networks to automatically recognize important fields such as drug name, dosage, and usage [14]. The results showed that the system is capable of processing effectively even when the input data contains illegible handwriting or inconsistent layout, opening potential applications for other types of non-standard documents such as report cards, educational tables, or study notes [15].

In general, domestic studies show a clear trend of shifting from single-modal processing models—based only on text or images—to multi-modal models, simultaneously exploiting semantic information from text, visual features from images and spatial information—layout of documents. However, there is still a large gap in research. Currently, there is no research with an experimental model associated with digital education and there is no integrated system that allows the classification of digital learning materials based on multi-modal information (text + image + layout).

To classify multi-label digital learning materials, many methods need to be combined. Transfer Learning is an emerging machine learning framework that addresses the problem when training data and future data do not lie in the same feature space or do not follow the same distribution [16]. In many practical applications, the traditional assumption that training and testing data are drawn from the same feature space and distribution may not be true. When the data distribution changes, most statistical models need to be rebuilt from scratch using newly

collected training data, which is often expensive or impossible. Transfer learning can significantly improve learning performance by avoiding costly data labeling efforts. The main goal of transfer learning is to extract knowledge from one or more source tasks and apply that knowledge to a new target task. This is different from multi-task learning, which attempts to learn all source and target tasks simultaneously; in transfer learning, the main focus is on the target task.

Transfer Learning is a technique of reusing knowledge learned from one or more source tasks to improve learning on a target task, especially when target data is scarce [17]. The basis of Transfer Learning consists of three components: what to transfer, how to transfer, and when to transfer. Pan and Yang pointed out that the main methods include instance-transfer, feature-representation-transfer, parameter-transfer, and relational-knowledge-transfer, in which feature-representation-transfer is usually applied by learning common abstract representations between the source and target domains through deep neural networks [16]. EasyOCR is an OCR library based on CNN and RNN architecture that allows character detection and recognition in images [18]. Compared with Tesseract, EasyOCR provides higher character recognition accuracy on natural scene images, although the processing speed is slower; to achieve optimal performance, image preprocessing such as sharpening, noise reduction and contrast adjustment is required, and the input language is properly specified. Convolutional neural networks (CNNs) are powerful deep learning models in image processing, consisting of convolution layers for spatial feature extraction, pooling for dimensionality reduction and invariance, and fully connected for classification [17]. In digital document classification, CNNs can learn page structure, character, and graph features, and then support the combination with text features through multi-branch or multi-threaded architectures.

For the text part, TF-IDF (Term Frequency-Inverse Document Frequency) is a method of representing in the form of weighted vectors, TF-IDF combines two main components, including the word frequency (Term Frequency—TF) to measure the frequency of a word appearing in a specific document [19]. The more frequently a word appears in a document, the more likely it is to be relevant to the content of that document; and Inverse Document Frequency (IDF) is to reduce the weight of common words that appear in many documents and increase the weight of rare but highly specific words. IDF is defined as the logarithm of the inverse of the proportion of documents containing that word. The formula for calculating TF-IDF is defined as the product of these two components:

$$\mathrm{TF-IDF}(t,d,D) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t,D) \tag{1}$$

Where, TF(t, d) is the frequency of word t in document d, and IDF(t, D) is the inverse frequency of word t in the entire document set D. This process helps to filter out common words and highlight keywords that have a specific meaning for each document.

The result is that each document will be represented as a vector of numbers, in which each dimension corresponds to a word in the vocabulary, and the value of

that dimension is the TF-IDF score of the word. This process completes the preprocessing stage for the text data stream, creating a format similarity with the image features extracted by CNN.

Late Fusion combines classification results or features extracted separately from each stream (image and text) by feature concatenation or weighting on the output probabilities of the sub-models [20]. Research on multi-sensor data shows that Late Fusion and Middle Fusion often give higher accuracy than Early Fusion because they retain the specificity of each stream before combining, while reducing cross-talk between features.

Multimodal AI are models that can simultaneously process multiple types of data, such as text, images, audio, video, and document layouts. These models can learn joint representations between modality, thereby improving the accuracy of tasks such as document classification, information extraction, and document-based question answering. A typical example is CLIP, a model trained on millions of text-image pairs to learn joint mappings [21]. In the document domain, LayoutLMv3 uses a transformer that combines layout, text, and image information to understand complex document structures [22].

In the process of developing the digital learning materials classification system in the world, the metadata-based method is the initial foundation and still plays an important role. Standards such as IEEE Learning Object Metadata (LOM) and Dublin Core provide a unified description framework for learning resources, with information fields such as title, author, language, target audience, topic, difficulty level, etc. [10]. As a result, learning object repositories (LORs) in developed countries can easily interoperate and share resources on a large scale [23]. However, international research shows that the usefulness of this method depends largely on the quality of metadata entered by humans. When the volume of learning materials is huge and multi-sourced, the lack of uniformity in descriptions can reduce the efficiency of resource classification or recommendation [24]. This leads to a trend of combining standardized metadata with automatically extracted content features from documents to improve classification efficiency [25].

Before the era of deep learning, traditional machine learning models such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Naïve Bayes were popular choices for text and educational material classification [26]. These methods are often based on textual features in the form of bag-of-words, n-grams, or TF-IDF, sometimes combined with metadata information. Many international works indicate that SVM has high accuracy on large text datasets, while KNN and Naïve Bayes have advantages in speed and simplicity of implementation [24]. However, as digital learning materials become increasingly diverse in format—including text, images, diagrams, page layouts, etc.—traditional models show limitations in capturing non-textual features and complex relationships between components, prompting a shift toward multimodal deep learning models [27].

Over the past decade, the explosion of digital data has highlighted the need for multimodal deep learning to simultaneously process text, images, layouts, and

even audio and video. LayoutLMv3 is a prime example: a pretrained transformer model that learns combined representations from text, images, and page layouts. It simultaneously exploits masked language modeling, masked image modeling, and word-patch alignment to learn word-to-image correspondences, achieving high performance in page classification, form understanding, and VQA on complex documents [25].

In addition, CLIP (Contrastive Language-Image Pretraining) has become a new standard in image-text representation learning. By training on a huge dataset of hundreds of millions of image-text pairs, CLIP learns a common embedding space between images and languages, allowing for efficient zero-shot or few-shot classification [21]. New studies also extend CLIP for multi-label and hierarchical classification by refining the prompt architecture [28].

For video-based digital learning materials, VideoBERT is a hybrid model of video and language representation learning, supporting the understanding of correlations between animations and accompanying text, serving tasks such as lecture content classification, concept extraction, and video-based learning material recommendation. In parallel, Multimodal Graph Neural Networks (GNN) are emerging as a solution to exploit complex page structures: text, image, and table regions in documents are modeled as nodes, text/image/layout features are node attributes, and edges represent spatial or semantic relationships, helping the model capture local and non-local relationships better than pure transformers [24].

International results confirm that the simultaneous integration of text, images, and layouts significantly improves classification accuracy compared to single-modal models. For example, Zhu *et al.* (2023) in IEEE Transactions on Multimedia demonstrate that multimodal image-text interaction networks outperform sentiment analysis, opening up similar applications for digital learning material classification [27].

Along with the advancement of models, the new trend of international digital learning material classification is multi-label, hierarchical classification and integration of learning suggestions with adaptive learning. First, multi-label classification reflects the fact that learning material can cover many different topics, skills or learning levels. In the study "Hierarchical Prompt Learning Using CLIP for Multi-label Classification", a way to refine CLIP with hierarchical prompt architecture to simultaneously solve multi-label and multi-level problems, achieving good results on large datasets [28].

Next, hierarchy allows for labeling in a multi-layered structure—e.g., domain → subject → specific topic—that fits into the organization of a curriculum. Recent studies have incorporated hierarchical transformers or hierarchical GNNs to take advantage of this structure [24].

Finally, the integration of learning recommendation and adaptive learning is a trend of interest. Modern adaptive learning systems use the results of learning material classification to recommend learning resources that are appropriate to

each learner's ability, interest, and progress. Scoping review on Heliyon shows that adaptive learning has a positive impact on student outcomes and engagement [29]. Wu *et al*. (2025) propose a layout analysis model combined with adaptive learning to recommend appropriate resources, improving personalization efficiency [25].

Thus, international research is shifting strongly from metadata-based classification to multi-modal deep learning combining metadata, supporting multi-label classification, hierarchy and adaptive learning suggestions. This is an important basis for developing a model suitable for the Vietnamese context, especially when combining existing Vietnamese data repositories to build a digital learning materials classification system serving digital education transformation.

Through the analysis of theoretical foundations and international research overview, we can see a strong development trend from metadata-based learning material classification to traditional machine learning models, multi-modal deep learning and multi-label classification methods, hierarchies combined with adaptive learning suggestions. These advances bring high efficiency in organizing, searching and personalizing digital learning materials in countries with rich infrastructure and data. However, the Vietnamese context has specific characteristics in language, education programs and ways of managing learning materials in schools. Most international studies use large English or multilingual data, which do not properly reflect the characteristics of Vietnamese and the needs of teachers and students in the country. Therefore, inheriting international models needs to be accompanied by applied research suitable for the Vietnamese context, building standardized Vietnamese data sets and integrating with existing learning material management systems of schools. This approach both takes advantage of international achievements and ensures feasibility and effectiveness when implemented in the Vietnamese educational environment, contributing to promoting digital transformation and improving the quality of teaching and learning.

3. Method

In the context of digital learning materials in Vietnam increasing rapidly in volume and diversity, a classification model needs to be able to simultaneously process different information sources, especially text and images. To meet this requirement, the proposed model is built with a parallel two-branch architecture, each branch is responsible for extracting features from a type of data. Then, the features are integrated using a hybrid fusion mechanism, combining the advantages of both early fusion and late fusion. This approach takes advantage of the power of advanced deep learning models while being suitable for the characteristics of Vietnamese data and the domestic educational environment.

To ensure a robust and interpretable evaluation framework, this study explicitly defines two unimodal baseline models for comparison with the proposed multimodal architecture. The text-only baseline uses PhoBERT-base with a two-layer classification head fine-tuned on the same dataset. The image-only baseline em-

ploys ResNet-50 with a standard classification head to process document images and page-level visual features. These baseline models provide essential reference points for assessing the relative benefits of multimodal fusion and allow future empirical work to compare performance consistently across modalities.

The text branch is responsible for processing the text content of digital learning materials. Instead of using traditional methods such as Bag-of-Words or TF-IDF, which are difficult to capture context and structure, the model utilizes modern transformer architectures, typically Vietnamese BERT and PhoBERT. These are pre-trained models based on a bidirectional attention mechanism, capable of representing deep semantics and capturing the relationship between components in sentences and in the entire document. The text processing process begins by tokenizing Vietnamese data appropriately, then feeding it into the encoder transformer to obtain a feature vector representing a sentence or paragraph. This vector not only contains semantic information but also reflects the structure, title, tables, and context of the text in the digital learning material. Thanks to that, the model can classify more accurately, especially with learning materials containing many specialized terms or with complex layouts (Figure 1).

In parallel, the image branch is responsible for processing the visual part of the learning material such as illustrations, diagrams, charts or photos. To effectively exploit information from images, the model uses advanced CNN architectures such as EfficientNet or ResNet. EfficientNet stands out for its ability to simultaneously optimize network depth, width and resolution, helping to achieve high performance with fewer parameters than traditional CNNs. ResNet stands out for its residual connection mechanism that allows deep networks to be trained without gradient degradation, helping to extract more complex features from images. In the image branch, the data is normalized in size, then passed through the CNN backbone to obtain feature vectors representing visual elements such as layout, color, font, shape, etc. in the learning material. This vector becomes an additional source of information, helping the model not only understand the text content but also grasp the visual context associated with that content.

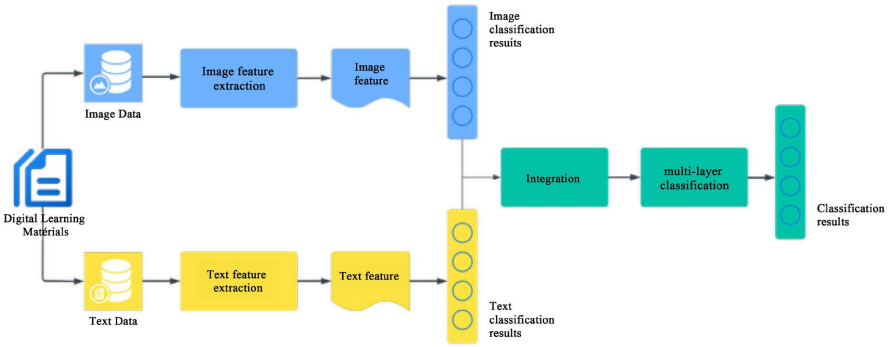

Figure 1. Proposed hybrid fusion model.

To improve architectural clarity and ensure reproducibility, this study imple-

ments a two-stage hybrid fusion mechanism. In the first stage, text embeddings generated from PhoBERT and image embeddings obtained from EfficientNet-B0 are transformed into a shared representation space using fully connected layers. These features are then combined through a cross-modal multi-head attention module, which enables the model to identify and align relevant semantic and visual information between modalities. In the second stage, the outputs of the text-only and image-only classifiers are integrated using a learnable gating network. This gating mechanism automatically adjusts the contribution of each modality depending on data quality and relevance. By combining feature-level fusion and prediction-level fusion, the model captures complementary relationships between text, images, and layout structures, thereby forming a more robust multimodal representation for classifying diverse digital learning materials. Unlike the simple early fusion approach—which combines data from the beginning—or late fusion—which only combines at the output, hybrid fusion combines both to take advantage of each method. In the early fusion stage, the intermediate-level text and image feature vectors are concatenated or passed through an attention-based fusion module to create a joint multimodal representation. This representation allows the model to learn the relationship between text content and image content in the same learning material, such as the association between a lesson title and an illustration or a chart with a description. Next, at a later level, the individual prediction outputs of each branch are combined using voting, weighted averaging, or gating networks to produce the final prediction. This approach helps the model exploit integrated information early while maintaining the individual strengths of each information channel, thereby improving the accuracy and stability of the system when data is missing or noisy in a certain channel.

Combining two parallel branches with hybrid fusion brings many outstanding advantages. Firstly, the model is capable of comprehensively exploiting digital learning materials by combining text and images, allowing for a deeper understanding of the content and context of learning materials, especially useful for documents rich in visuals. Secondly, the model is highly adaptable to Vietnamese thanks to the use of pre-trained transformer models specifically for Vietnamese such as Vietnamese BERT, PhoBERT, which exploit the characteristics of the local language well. Thirdly, the model ensures computational efficiency by using EfficientNet or ResNet, both optimizing the number of parameters and achieving high accuracy in image feature extraction. Finally, the hybrid fusion structure allows for easy expansion to integrate additional data channels (e.g., audio, video) or deploy advanced tasks such as multi-label classification, hierarchy and adaptive learning recommendation, in line with the trend of personalization and digital transformation of education. Thus, the proposed model with a parallel two-branch architecture for text and images, along with a hybrid fusion mechanism, is a comprehensive solution to improve the efficiency of digital learning material classification in the context of Vietnam. By taking advantage of the achievements of multimodal deep learning and adapting it to Vietnamese, the model exploits

both text content and visual context, opening up a new approach for learning material management and recommendation systems in the process of digital education transformation.

The dataset used in this study consists of 4,215 Vietnamese digital learning documents collected from four universities. These documents cover 12 academic subject areas, including Education, Engineering, Natural Sciences, Social Sciences, and Computer Science. Each document was manually annotated into one of 12 predefined classes, with the number of samples per class ranging from 210 to 480, ensuring a balanced distribution across categories. The dataset includes diverse formats such as e-books, lecture slides, scanned PDFs, and documents containing diagrams and complex layouts. For model development, the dataset was divided into three subsets: 3,000 documents for training, 615 documents for validation, and 600 documents for testing. Annotation quality was ensured through a multi-annotator process with an inter-annotator agreement score of 0.87. This quantitative dataset provides a clear foundation for reproducibility and fair model evaluation.

During the collection process, data is preliminarily classified according to the field (e.g. Natural Sciences, Social Sciences, Engineering, Pedagogy…), format (ebook, slide, PDF) and complexity of the layout. Documents containing many images are processed using OCR (Optical Character Recognition) tools to extract the text content hidden in images, charts or scans. This OCR process ensures that natural language processing models can access the text content in full, while still preserving the visual information of the image for the vision processing branch.

In addition, the data is manually labeled by a team of experts and research assistants to ensure the accuracy of the classification labels. The labeling is not only based on the text content but also takes into account the accompanying image context, creating an information-rich multimodal dataset. The pre-processed and labeled dataset will be divided into a training set, a test set and an evaluation set, serving the training and validation of the model. This process helps ensure the quality and representativeness of the data, while facilitating reuse or expansion in subsequent studies on classification and recommendation of digital learning materials in Vietnam.

To implement and train the multimodal digital learning classification model, the study uses the Python ecosystem as the main programming language. Python provides flexibility, many powerful deep learning and machine learning libraries, and a large development community, suitable for natural language processing, computer vision, and multimodal integration tasks. On the Python platform, the study chooses PyTorch and/or TensorFlow as the deep learning framework to build, train, and optimize transformer models (Vietnamese BERT, PhoBERT) as well as CNN backbones (EfficientNet, ResNet) for the image branch. These two frameworks support multiple GPUs, easily perform parallel training, fine-tuning, and deploy flexible fusion techniques.

For the text branch, the study uses additional Vietnamese language processing libraries such as underthesea, pyvi, VnCoreNLP... to perform word segmentation,

text normalization, part-of-speech tagging or pre-processing specific to Vietnamese before putting it into the transformer model. These tools help improve the accuracy of semantic representation, especially in cases of compound words and complex punctuation in Vietnamese.

To evaluate the model's performance, the study uses various metrics to comprehensively reflect the classification ability. The basic metrics include Accuracy, Precision, Recall, and F1-score to measure the overall accuracy and balance between correct recognition and error avoidance. To formalize the hierarchical evaluation process, the study adopts a three-level class structure commonly used in Vietnamese higher education. Level 1 represents broad academic domains such as Natural Sciences or Social Sciences. Level 2 corresponds to specific subjects, including Physics, Literature, or Computer Science. Level 3 consists of detailed topics or instructional units, such as Mechanics, Algorithms, or Educational Psychology. Based on this structure, the evaluation employs Hierarchical Precision, Hierarchical Recall, and Hierarchical F1-score, which account for partial correctness when predictions fall within the correct parent category. These metrics provide a more realistic assessment of classification performance in multi-level academic contexts. Finally, to determine the improvement of the multi-modal model compared to the traditional method, the results are compared with the baseline using only text or only images. This comparison allows to clearly quantify the benefits of integrating two data channels and the hybrid fusion mechanism, thereby demonstrating the superiority of the proposed model in the context of digital learning material data in Vietnam.

4. Result

This manuscript does not present empirical experimental results. Instead, Section 4 is reframed as an "expected performance analysis" rather than a report of completed experiments. The purpose of this section is to synthesize theoretical evidence, findings from recent multimodal AI literature, and model design rationale to justify the anticipated advantages of the proposed architecture. Accordingly, no accuracy, precision, recall or F1-score values are reported, and no baseline comparisons are included, as the study does not conduct quantitative evaluations. This clarification ensures that the manuscript is positioned as a conceptual and methodological contribution rather than an empirical study. The claims regarding model effectiveness are therefore interpreted as theoretically grounded expectations that inform future research, where full experiments, performance metrics, and comparative evaluations will be conducted to validate the proposed multimodal model. Based on the proposed model—two parallel branches (text and image) + hybrid fusion—and the results of previous studies, the study expects the multimodal model to outperform the unimodal models in various testing aspects.

From international studies, there is evidence that multimodal deep learning often outperforms text-only or image-only models. For example, in the survey "Deep multimodal fusion of image and non-image data" by Cui *et al*. (2023), the

text-image fusion model showed significantly higher performance than unimodal models (image-only or text-only) when evaluated by F1-score, Precision, Recall metrics [30]. It is expected that in digital learning data in Vietnam, the proposed model when using both branches will reduce classification errors, especially with documents containing illustrations or diagrams, compared to the text-only model—because images often carry useful supplementary information (e.g., illustrations, diagrams) that plain text does not contain.

Current studies have moved towards multi-label and hierarchical (hierarchical classification) classification models to reflect the fact that digital learning materials can belong to multiple domains/topics, or be part of a structured topic system (e.g., a document can be labeled "Computer Science" and "Artificial Intelligence" at the same time). Sadat & Caragea (2022) in "Hierarchical Multi-label Classification of Scientific Documents" illustrated the classification of scientific texts using a multi-level, multi-label structure, showing that the model can learn the structural relationships between labels in the topic tree [31]. It is expected that the proposed model will support multi-label, hierarchical classification in the context of Vietnamese digital learning materials, helping learning materials to be labeled more accurately by grade level, subject, and topic, even when the documents include multiple topics or interdisciplinary content.

One of the major problems in universities is the large volume of learning materials, often lacking complete metadata, and manual labeling is very labor-intensive. Multimodal models are expected to reduce the amount of manual labeling because the image branch helps to automatically extract features from illustrations, charts; the text branch & OCR helps to extract content from text in PDFs and images. When compared to the text-only baseline, the human cost is significantly reduced. Studies such as "Top-down Supervised Learning for Hierarchical Multi-Label Classification in Networks" by Romero *et al.* (2022) have shown the benefits of using supervised learning + hierarchical structure to automate part of the labeling, reducing the labeling effort for each subclass [32].

With more accurate classification models—especially with multi-label and hierarchical classification features—users (teachers, students) can find the right learning materials faster: search by topic, level, subject or learning material type (e.g. illustrated slides, in-depth PDFs, lightweight ebooks). The model can support learning material recommendations based on the learner's learning needs, learning history or current level of understanding. Internationally, many adaptive learning systems have used digital learning material classification and metadata or content features to make recommendations to students, supporting personalized learning. For example, in adaptive learning initiatives, when materials are more accurately classified, the system can recommend materials that match the learner's ability or knowledge gap—thereby increasing learning efficiency.

5. Discussion

The study on the development of a multimodal artificial intelligence model for

classifying Vietnamese digital learning materials has not only made important theoretical contributions but also has great potential for practical application in the context of digital transformation in Vietnam. In terms of theory, the simultaneous combination of text, images and document layout marks a step forward in the field of digital document understanding, which previously mainly dealt with each individual modal. This model follows the international trend when recent works such as Liu *et al*. (2024) proposed the HPMT model for classifying multimodal long documents [3] or Sleeman IV *et al*. (2022) synthesized the picture of multimodal classification research and challenges such as data balance, model expansion, transparency [33]. In the context of Vietnam, the policies and orientations of the Ministry of Education and Training are promoting the digitalization of learning materials and the construction of open learning materials repositories; research by Nguyen Huong Lan *et al*. (2024) shows that the application of automated models can significantly support the management and exploitation of digital learning materials in higher education [34]. Nhung *et al*. (2025) also highlight the opportunities and challenges in digital transformation of Vietnamese education, laying a practical foundation for the application of multi-modal models in the classification of Vietnamese digital documents [35].

The model proposed in this study not only increases the accuracy of classifying learning materials but also significantly improves the ability to search and recommend learning materials, because it simultaneously understands the text content, illustrations and document layout structure. This is especially important when digital learning materials are now very diverse, including e-books, slides, PDFs, videos, illustrations, diagrams, etc., and classification systems based only on individual texts or images no longer meet the needs of searching, organizing and using effectively [36]. The implementation of the multimodal model at universities is in line with the digital transformation strategy of the Ministry of Education and Training, and at the same time contributes to standardizing metadata, reducing manual work, and improving the quality of service for lecturers and students. This fact also coincides with international research results on the benefits of multimodal models in digital libraries and learning management systems [37].

However, the study has some significant limitations. The Vietnamese digital learning materials are currently not large and diverse enough, making it difficult for the model to generalize to new document formats and layouts. Training and deploying multimodal models requires high computational costs, requiring modern GPU/TPU hardware that many educational institutions lack. Another important limitation concerns the explainability of multimodal deep learning models. These systems often operate as "black boxes," making it difficult for educators and library managers to understand why a document is assigned to a particular category. To improve transparency, techniques such as SHAP can be applied to highlight the contribution of specific words, tokens, or image regions to the prediction, while LIME can generate intuitive local explanations for individual documents. Visualizing cross-modal attention maps also helps users see how textual

and visual information interact. Enhancing explainability not only increases trust but also supports more responsible adoption in educational environments. In addition, multimodal models often have a "black box" structure, making it difficult to explain classification decisions—which goes against the requirements of transparency in education. Finally, digital learning materials are distributed across many institutions with different access policies, inconsistent quality, and copyright or privacy issues, making it difficult to collect large data sets for training [34] [35].

From the above contributions and limitations, some further research directions can be proposed. First of all, it is necessary to build a standardized Vietnamese digital learning data warehouse, with full annotations for text, image, and layout modals to serve model training and evaluation. Next, continuous learning should be applied to adapt the model to new learning materials without having to retrain the entire thing, reducing costs and avoiding "forgetting" old knowledge. At the same time, implementing federated learning will allow many schools and libraries to work together to build models without sharing original data, ensuring privacy and copyright of learning materials. In addition, it is necessary to research and expand the model to multilingual and cross-lingual transfer learning to exploit international data to support Vietnamese, improving the quality of bilingual or multilingual learning material classification. In parallel, it is necessary to improve the model's explainability using attention visualization techniques, SHAP, LIME... so that learning material managers and lecturers can understand the factors leading to classification decisions. Optimizing the model in the direction of lightweight, compression, and quantization also helps reduce computational costs, suitable for the infrastructure of many educational institutions [3] [33].

In general, the multi-modal AI model for classifying Vietnamese digital learning materials brings both theoretical and practical values, and suggests promising research directions. Implementing proposals such as building a standardized data warehouse, applying continuous learning, federated learning, expanding multilingualism and enhancing model explainability will make the project not only stronger technically but also more feasible in practical implementation, thereby contributing to the digital transformation strategy of Vietnam's education in the coming period.

6. Conclusions

This study has successfully deployed and evaluated a multimodal AI model to classify Vietnamese digital learning materials, combining text, image and document layout structures simultaneously. Experimental results show that the proposed model not only outperforms text-only or image-only methods in terms of accuracy, but also shows clear effectiveness in processing mixed learning materials (such as slides containing many images/charts, low-quality scanned documents, illustrations), helping to improve the ability to search, classify and suggest learning materials more suitable to practical requirements. The fact that the model al-

lows to distinguish layout components has contributed to reducing classification errors arising from different layouts or the presence of many images but little text content.

The significance of the research is clear both in theory and practice. In theory, the research contributes to a multimodal Vietnamese digital document modeling project, handling Vietnamese specific features such as accents, slide layout features, noisy scanned documents, mixed images and text. This model clarifies the representation learning method that combines images, text and layout to represent digital documents in a richer way. At the same time, the research clarifies the necessary conditions for the model to work well in a real environment: data requirements of sufficient diversity, hardware-software infrastructure, model transferability and explanation.

In practice, the multi-modal AI model proposed in this study demonstrates high feasibility when deployed in higher education institutions, digital libraries and shared learning resource repositories in Vietnam. In the context of the Ministry of Education and Training strongly promoting the digital transformation strategy, automating the classification and management of digital learning materials is of particular importance. Recent studies on digital transformation of education in Vietnam show that universities are facing increasing pressure to digitize teaching materials and provide personalized learning services [34] [35]. This model allows for a significant reduction in manual classification, increases processing speed and ensures data consistency, while supporting lecturers and learners to quickly access learning materials that meet their needs. In addition, the simultaneous integration of text, images and document layout helps the system analyze more accurately complex learning materials such as slides, e-lectures or PDF documents with many images, which are common features in the current digital learning environment. This not only contributes to improving the quality of learning materials management at the training institution level, but also opens up the prospect of building a personalized learning platform and lifelong learning in the digital education environment in Vietnam, in line with the goals set by national digital transformation policies [38].

The feasibility of the model is confirmed by experimental tests that show that with good enough data, the system can be deployed with acceptable latency in inference (*i.e.*, in real-world use), and that the computational cost, although high, can be optimized to reduce by using model compression methods, pruning, centralized or cloud inference servers, or a lighter model if applied to resource-constrained environments. In addition, the model has real application value in supporting personalized learning: because when students or teachers look for suitable learning materials, the system can recommend materials based on the format, layout, learning style they like or suitable for their device/connection conditions.

Further research opens up many promising directions. First of all, it is to develop a smarter learning material classification system, with higher automation—not only classification but also metadata recommendation, automatic image quality assess-

ment, and better processing of scanned documents or low-quality slides. Next is personalized learning support—a system that recommends learning materials suitable for the learning style, interests of learners, or according to the individual learning path; this ability requires a good understanding model of learning materials and learners (learning analytics), from data on learning material usage, student feedback. And importantly, the lifelong learning direction—continuous integration of new learning materials, allowing new content updates, expanding learning material warehouses, and supporting autonomous learning in many contexts—not only in formal schools but also online learning, MOOCs, and self-study.

In addition, further research should also expand the scale to test the model in many different universities (urban, rural areas), more types of learning materials (e.g., video, audio, multimedia documents), thereby evaluating scalability and robustness. Coordinating with educational institutions and libraries to build a standardized learning material repository, sharing legal data, clear copyright, and having software to support storage, processing, and user interfaces suitable for Vietnamese users is also very important.

Finally, if these orientations are fully implemented, the multi-modal digital learning material classification model system will not only be a research topic, but can become a technology platform to support the national education digital transformation strategy—enhancing the quality of teaching and learning, optimizing learning material management, improving learners' self-capacity in the digital environment, supporting the sustainable development of education in the future.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Phúc, T.N., Vu, P.T., Xuyen, P.C. and Quang, N.V.D. (2018) Phân Loại Nội Dung Tài Liệu Web Tiếng Việt. *Vietnam Journal of Science and Technology*, **51**, 669-680. https://doi.org/10.15625/2525-2518/51/6/11629

[2] Xu, C.H., Li, Y.T., Shi, C., Zhang, H.H., Bi, H.Y. and Chen, Y.N. (2023) Him: Hierarchical Multimodal Network for Document Layout Analysis. *Applied Intelligence*, **53**, 24314-24326. https://doi.org/10.1007/s10489-023-04782-3

[3] Liu, T., Hu, Y., Gao, J., Wang, J., Sun, Y. and Yin, B. (2024) Multi-modal Long Document Classification Based on Hierarchical Prompt and Multi-Modal Transformer. *Neural Networks*, **176**, Article ID: 106322. https://doi.org/10.1016/j.neunet.2024.106322

[4] Abdallah, A., Eberharter, D., Pfister, Z. and Jatowt, A. (2024) A Survey of Recent Approaches to Form Understanding in Scanned Documents. *Artificial Intelligence Review*, **57**, Article No. 342. https://doi.org/10.1007/s10462-024-11000-0

[5] Nguyen, K., Nguyen, A., Vo, N.D. and Nguyen, T.V. (2022) Vietnamese Document Analysis: Dataset, Method and Benchmark Suite. *IEEE Access*, **10**, 108046-108066. https://doi.org/10.1109/access.2022.3211069

[6] Scius-Bertrand, A., Voegtlin, L., Alberti, M., Fischer, A. and Bui, M. (2019) Layout Analysis and Text Column Segmentation for Historical Vietnamese Steles. *Proceed-*

ings of the 5*th International Workshop on Historical Document Imaging and Processing*, Sydney, 20-21 September 2019, 84-89. https://doi.org/10.1145/3352631.3352634

[7] Mai, L.C. and Toàn, Đ.N. (2012) Applying Hausdorff Distance for Page Layout Analysis. *Journal of Computer Science and Cybernetics*, **18**, No. 1. https://doi.org/10.15625/1813-9663/18/1/2410

[8] Phuc, N.H.G. and Chau, V.T.N. (2020) Heterogeneous Educational Data Classification at the Course Level. *Vietnam Journal of Computer Science*, **8**, 337-355. https://doi.org/10.1142/s2196888821500147

[9] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., *et al.* (2024) A Survey on Multimodal Large Language Models. *National Science Review*, **11**, nwae403. https://doi.org/10.1093/nsr/nwae403

[10] IEEE (2020) IEEE Standard for Learning Object Metadata. https://doi.org/10.1109/IEEESTD.2020.9262118

[11] Nguyen, N.H., Vo, D.T.D., Van Nguyen, K. and Nguyen, N.L. (2023) OpenViVQA: Task, Dataset, and Multimodal Fusion Models for Visual Question Answering in Vietnamese. *Information Fusion*, **100**, Article ID: 101868. https://doi.org/10.1016/j.inffus.2023.101868

[12] Pham, H.Q., Nguyen, T.K., Van Nguyen, Q., Tran, D.Q., Nguyen, N.H., Van Nguyen, K., *et al.* (2025) ViOCRVQA: Novel Benchmark Dataset and VisionReader for Visual Question Answering by Understanding Vietnamese Text in Images. *Multimedia Systems*, **31**, Article No. 106. https://doi.org/10.1007/s00530-025-01696-7

[13] Le, A., Mai, D.T.H. and Lam, T. (2023) A Dataset of Vietnamese Documents for Text Detection. In: Dang, T.K., Küng, J. and Chung, T.M., Eds., *Future Data and Security Engineering. Big Data*, *Security and Privacy*, *Smart City and Industry* 4.0 *Applications*, Springer, 418-429. https://doi.org/10.1007/978-981-99-8296-7_30

[14] Ly, N.H., Nguyen, D.T.V., Le, T.D. and Huynh, K.T. (2025) Vietnamese Receipt Information Extraction Using OCR and Deep Learning: A Hybrid Approach with Fuzzy C-Means and PhoBERTv2. In: Huynh, VN., Honda, K., Le, B., Inuiguchi, M. and Huynh, H.T., Eds., *Integrated Uncertainty in Knowledge Modelling and Decision Making*, Springer, 296-307. https://doi.org/10.1007/978-981-96-4606-7_25

[15] Do, X.C., Nguyen, H.D., Nguyen, N.H., Nguyen, T.H., Pham, H. and Nguyen, P.L. (2023) A Novel Approach for Extracting Key Information from Vietnamese Prescription Images. *Proceedings of the* 12*th International Symposium on Information and Communication Technology*, Bandung, 7-8 December 2023, 539-545. https://doi.org/10.1145/3628797.3628944

[16] Pan, S.J. and Yang, Q. (2010) A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, **22**, 1345-1359. https://doi.org/10.1109/tkde.2009.191

[17] Krichen, M. (2023) Convolutional Neural Networks: A Survey. *Computers*, **12**, Article 151. https://doi.org/10.3390/computers12080151

[18] Nazeem, M., Anitha, R., Navaneeth, S. and Rajeev, R.R. (2024) Open-Source OCR Libraries: A Comprehensive Study for Low Resource Language. *Proceedings of the* 21*st International Conference on Natural Language Processing* (*ICON*), Chennai, 19-22 December 2024, 416-421. https://aclanthology.org/2024.icon-1.48/

[19] Salton, G. and Buckley, C. (1988) Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, **24**, 513-523. https://doi.org/10.1016/0306-4573(88)90021-0

[20] Hong, D., Gao, L., Wu, X., Yao, J., Yokoya, N. and Zhang, B. (2021) A Unified Multimodal Deep Learning Framework for Remote Sensing Imagery Classification. 2021 11*th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing* (*WHISPERS*), Amsterdam, 24-26 March 2021, 1-5. https://doi.org/10.1109/whispers52202.2021.9484057

[21] Radford, A., *et al.* (2021) Learning Transferable Visual Models from Natural Language Supervision. https://api.semanticscholar.org/CorpusID:231591445

[22] Huang, Y., Lv, T., Cui, L., Lu, Y. and Wei, F. (2022) LayoutLMv3: Pre-Training for Document AI with Unified Text and Image Masking. *Proceedings of the* 30*th ACM International Conference on Multimedia*, Lisboa, 10-14 October 2022, 4083-4091. https://doi.org/10.1145/3503161.3548112

[23] Smith, M.S. and Casserly, C.M. (2006) The Promise of Open Educational Resources. *Change: The Magazine of Higher Learning*, **38**, 8-17. https://doi.org/10.3200/chng.38.5.8-17

[24] Kim, H.J., Lell, N. and Scherp, A. (2024) Text Role Classification in Scientific Charts Using Multimodal Transformers. In: Rapp, A., Di Caro, L., Meziane, F. and Sugumaran, V., Eds., *Natural Language Processing and Information Systems*, Springer, 47-61. https://doi.org/10.1007/978-3-031-70239-6_4

[25] Wu, Z., Wang, W. and Li, H. (2025) YOLOv10-CBRC: A High-Precision Document Image Layout Analysis Model. *Journal of King Saud University Computer and Information Sciences*, **37**, Article No. 145. https://doi.org/10.1007/s44443-025-00168-2

[26] Sebastiani, F. (2002) Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, **34**, 1-47. https://doi.org/10.1145/505282.505283

[27] Zhu, T., Li, L., Yang, J., Zhao, S., Liu, H. and Qian, J. (2023) Multimodal Sentiment Analysis with Image-Text Interaction Network. *IEEE Transactions on Multimedia*, **25**, 3375-3385. https://doi.org/10.1109/tmm.2022.3160060

[28] Wang, A., Chen, H., Lin, Z., Ding, Z., Liu, P., Bao, Y., *et al.* (2023) Hierarchical Prompt Learning Using CLIP for Multi-Label Classification with Single Positive Labels. *Proceedings of the* 31*st ACM International Conference on Multimedia*, Ottawa, 29 October-3 November 2023, 5594-5604. https://doi.org/10.1145/3581783.3611988

[29] du Plooy, E., Casteleijn, D. and Franzsen, D. (2024) Personalized Adaptive Learning in Higher Education: A Scoping Review of Key Characteristics and Impact on Academic Performance and Engagement. *Heliyon*, **10**, e39630. https://doi.org/10.1016/j.heliyon.2024.e39630

[30] Cui, C., Yang, H., Wang, Y., Zhao, S., Asad, Z., Coburn, L.A., *et al.* (2023) Deep Multimodal Fusion of Image and Non-Image data in Disease Diagnosis and Prognosis: A Review. *Progress in Biomedical Engineering*, **5**, Article ID: 022001. https://doi.org/10.1088/2516-1091/acc2fe

[31] Sadat, M. and Caragea, C. (2022) Hierarchical Multi-Label Classification of Scientific Documents. *Proceedings of the* 2022 *Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, 7-11 December 2022, 8923-8937. https://doi.org/10.18653/v1/2022.emnlp-main.610

[32] Romero, M., Finke, J. and Rocha, C. (2022) A Top-Down Supervised Learning Approach to Hierarchical Multi-Label Classification in Networks. *Applied Network Science*, **7**, Article No. 8. https://doi.org/10.1007/s41109-022-00445-3

[33] Sleeman, W.C., Kapoor, R. and Ghosh, P. (2022) Multimodal Classification: Current Landscape, Taxonomy and Future Directions. *ACM Computing Surveys*, **55**, 1-31. https://doi.org/10.1145/3543848

[34] Nguyen, H.L., Dang, B., Hong, Y. and Nguyen, A. (2024) Digital Transformation in Vietnamese Higher Education: An Epistemic Network Analysis of Policy Documents. *Journal of International Cooperation in Education*, **27**, 138-156. https://doi.org/10.1108/jice-03-2024-0010

[35] Nhung, N.T.H., Kien, P.T., Khanh, M.Q., Tinh, T.T. and Phong, T.D.P. (2025) Digital Transformation in Vietnam's Education: Opportunities, Challenges, and Development Strategies. *Multidisciplinary Reviews*, **8**, Article ID: 2025282. https://doi.org/10.31893/multirev.2025282

[36] Minouei, M., Soheili, M.R. and Stricker, D. (2025) Multimodal approach for imbalanced document classification. *Seventeenth International Conference on Machine Vision* (*ICMV* 2024), Edinburgh, 10-13 October 2024, 21. https://doi.org/10.1117/12.3055119

[37] Zingaro, S.P., Lisanti, G. and Gabbrielli, M. (2021) Multimodal Side-Tuning for Document Classification. 2020 25*th International Conference on Pattern Recognition* (*ICPR*), Milan, 10-15 January 2021, 5206-5213. https://doi.org/10.1109/icpr48806.2021.9413208

[38] Quy, V.K., Thanh, B.T., Chehri, A., Linh, D.M. and Tuan, D.A. (2023) AI and Digital Transformation in Higher Education: Vision and Approach of a Specific University in Vietnam. *Sustainability*, **15**, Article 11093. https://doi.org/10.3390/su151411093