A Novel 2D Deep Convolutional Neural Network for Multimodal Document Categorization

Rustam Abkrakhmanov[1], Aruzhan Elubaeva[2], Tursinbay Turymbetov[3], Venera Nakhipova[4], Shynar Turmaganbetova[5], Zhanseri Ikram[6]

International University of Tourism and Hospitality, Turkistan, Kazakhstan[1]
Bachelor Student at International University of Tourism and Hospitality, Turkistan, Kazakhstan[2]
Khoja Akhmet Yassawi International Kazakh, Turkish University, Turkistan, Kazakhstan[3]
Zhumabek Akhmetuly Tashenev University, Shymkent, Kazakhstan[4]
NCJSC «S.Seifullin Kazakh Agro Technical Research University», Astana, Kazakhstan[5]
BTS Digital, Astana, Kazakhstan[6]

***Abstract*—Digitized documents are increasingly becoming prevalent in various industries. The ability to accurately classify these documents is critical for efficient and effective management. However, digitized documents often come in different formats, making document classification a challenging task. In this paper, we propose a multimodal deep learning approach for digitized document classification. The proposed approach combines both text and image modalities to improve classification accuracy. The model architecture consists of a convolutional neural network (CNN) for image processing and a recurrent neural network (RNN) for text processing. The output features from the two modalities are then merged using a fusion layer to generate the final classification result. The proposed approach is evaluated on a dataset of digitized documents from various industries, including finance, healthcare, and legal fields. The experimental results demonstrate that the multimodal approach outperforms single-modality approaches, achieving high accuracy for document classification. The proposed model has significant potential for applications in various industries that rely heavily on document management systems. For example, in the finance industry, the proposed model can be used to classify loan applications or financial statements. In the healthcare industry, the model can classify patient records, medical images, and other medical documents. In the legal industry, the model can classify legal documents, contracts, and court filings. Overall, the proposed multimodal deep learning approach can significantly improve document classification accuracy, thus enhancing the efficiency and effectiveness of document management systems.**

Keywords*—Scanned documents; classification; document categorization; artificial intelligence; machine learning; deep learning

I. Introduction

With the rise of digital transformation, documents in various industries are now digitized for easy access and management. Digitized documents come in various formats such as PDF, JPEG, PNG, and many others, which makes document classification a challenging task. The ability to accurately classify these documents is crucial for efficient document management. The traditional approach of manual classification is time-consuming and prone to errors [1]. Hence, the development of automated document classification systems has become a critical need for organizations.

Recent advances in deep learning have significantly improved the accuracy of document classification [2]. Deep learning algorithms have been applied to various document classification tasks, such as sentiment analysis, topic modeling, and spam detection [3-4]. These algorithms have shown remarkable performance in text classification tasks by leveraging the vast amount of data and computing power available today [5]. However, text-based classification models may not perform well when the documents also contain visual information, such as images, logos, and diagrams [6].

To address this challenge, we propose a multimodal deep learning approach for digitized document classification that combines text and image modalities. The proposed model uses a convolutional neural network (CNN) [7] for image processing and a recurrent neural network (RNN) [8] for text processing. The two modalities are then merged using a fusion layer to generate the final classification result. The proposed approach is evaluated on a dataset of digitized documents from various industries, including finance, healthcare, and legal fields.

The remainder of this paper is organized as follows: Section II discusses related work on document classification and multimodal deep learning. Section III presents the proposed approach for digitized document classification using multimodal deep learning. Section IV describes the experimental setup and the results of the proposed approach. Section V discusses the limitations and future directions of this work. Finally, Section VI concludes the paper.

II. Related Works

Document classification has been an active area of research for several decades. Traditional approaches to document classification relied on manual feature engineering, which involves the extraction of relevant features from documents and their mapping to predefined classes. However, manual feature engineering can be challenging and time-consuming, and the performance of such models is limited by the quality of the extracted features [9].

Recent advancements in deep learning have enabled automated feature learning, where the model automatically learns relevant features from the data [10]. Several studies have proposed deep learning models for document classification. Convolutional neural networks (CNNs) have been widely used for image classification tasks and have also been applied to document classification. In a CNN, a filter scans the input image, and the output of the filter is then passed through a nonlinear activation function to generate the output feature map [11]. These output feature maps are then used to classify the input image.

Recurrent neural networks (RNNs) have also been used for text classification tasks [12]. RNNs can capture the contextual dependencies of the input sequence, making them ideal for tasks such as language modeling and machine translation. RNNs process the input sequence one token at a time, and the hidden state of the network is updated at each time step. The final hidden state is then used for classification.

Multimodal deep learning has been used for various tasks, such as speech recognition, visual question answering, and image captioning, education using game-based learning [13]. Multimodal models combine information from different modalities, such as text, image, and audio, to improve performance. The fusion of information from different modalities can help address the limitations of single-modal models.

In recent years, deep learning approaches have shown promising results in document classification tasks, especially in single-modal scenarios such as text or image classification [14]. However, the classification accuracy can be further improved by incorporating multiple modalities, such as text and image, into the classification process.

Several studies have proposed multimodal deep learning approaches for document classification. For instance, Malaperdas et al. (2021) proposed a multimodal approach that combines text and image modalities for document classification [15]. They used a convolutional neural network (CNN) for image processing and a recurrent neural network (RNN) for text processing, and merged the output features using a fusion layer. The proposed approach achieved better performance than single-modal approaches.

Similarly, Zhang et al. (2020) proposed a multimodal approach that combines text and layout features for document classification [16]. They used a CNN to extract image features and a text-crop CNN to extract text features, and merged the features using a fully connected layer. The proposed approach achieved better classification accuracy than using either text or layout features alone.

In addition to combining text and image modalities, other studies have explored the use of additional modalities such as audio and video. For example, Hegghammer (2022) proposed a multimodal deep learning approach that combines text, image, and audio modalities for news classification [17]. They used a CNN for image processing, a bidirectional LSTM for text processing, and a convolutional neural network for audio processing. The output features were merged using a fusion layer and fed into a fully connected layer for classification.

Moreover, some studies have explored the use of transfer learning techniques to improve the classification performance. For example, Revilla-León et al. (2020) proposed a multimodal approach that uses a pre-trained CNN for image processing and a pre-trained LSTM for text processing [18]. They fine-tuned the pre-trained models on their own dataset and achieved better classification accuracy than training from scratch.

In summary, several studies have proposed multimodal deep learning approaches for digitized document classification, which combine multiple modalities such as text, image, and layout. The proposed approaches have shown promising results in improving classification accuracy compared to single-modal approaches. Furthermore, the use of transfer learning techniques has also been explored to improve the classification performance.

III. Flowchart of the Research

The proposed method for digitized document classification using multimodal deep learning involves combining text and image modalities to improve the classification accuracy. The method consists of three main steps: data preprocessing, feature extraction, and classification.

Data Preprocessing: The first step involves preprocessing the raw document data to make it suitable for processing by the deep learning models. This step includes tasks such as text normalization, image preprocessing, and data augmentation.

Feature Extraction: The second step involves extracting features from the text and image modalities. For text feature extraction, we use a pre-trained language model such as BERT to encode the text data into a high-dimensional vector representation. For image feature extraction, we use a pre-trained convolutional neural network (CNN) such as VGG or ResNet to extract image features.

Modal Fusion: The third step involves fusing the text and image features using a multimodal fusion layer. This layer can take various forms, such as concatenation, element-wise multiplication, or attention-based fusion. The output of the fusion layer is then fed into a fully connected layer for classification.

Classification: The final step involves training the classification model using the fused features and evaluating its performance on a held-out test set. We use a multi-layer perceptron (MLP) classifier with softmax activation for classification. The model is trained using a categorical cross-entropy loss function and optimized using the Adam optimizer.

To evaluate the proposed method, we will conduct experiments on a publicly available dataset of 10,000 digitized documents. We will compare the performance of our multimodal approach with single-modal approaches such as text-only and image-only classification. We will also evaluate the effect of different modal fusion strategies on the classification performance. The evaluation metrics will include accuracy, precision, recall, and F1-score. Finally, we will conduct ablation studies to analyze the contribution of each modality to the overall classification performance.

IV. Proposed Method

In this part, a comprehensive analysis of every facet of our methodology for identifying subjects is presented. Further, it is strongly recommended to conduct a textual and graphical components based analysis of the primary document in order to build and implement a one-of-a-kind semantics topic classification strategy. The remainder of this part will be dedicated to providing a detailed explanation of the structure of the entire system, with the primary attention being placed on the essential characteristics that are shared by every component of the recommended scheme. In addition to this, it offers a wealth of information with respect to the multidimensional knowledge base as well as the theoretical model that served as the foundation for the knowledge base. Moreover, the document cites several examples. Following this, we will offer an in-depth description of the Topic Detection approach that we have just shown.

A. Architecture of the System

We gathered a 9125 picture collection and divided it into seven groups. The style, structure, and content of XML documents may be accessed and updated by programs using the DOM Parser, which was used in the next step. We next gathered the cleaned text files into a data store. The next phase was gathering the attributes required to train a model for document classification. We created a deep model to train and test the suggested model after deciding which attributes were essential. As a result, we separated classified content into several groups.

Because of this, the taxonomy classificator is able to begin the construction of the classification taxonomy with a notion. Comparisons have been made between the recommended measure and technique and the baselines, and the experimental Section IV of the report illustrates and discusses the conclusions of these comparisons.

The deep neural network that was suggested to solve the document classification issue is seen in Fig. 2. The scanned paper that we get serves as the input for the planned network. Eleven layers make up the network that is being proposed. Pooling occurs in the first layer of the structure. In the subsequent step, the Conv2D layer is used, and the result that is acquired is then transferred to the bottleneck layers. In this particular network, there are five levels of bottlenecks. Following that, a pooling layer and a Conv2D layer are used. The subsequent level contains 128 neuronal components, and the penultimate layer has only seven layers, which correspond to the various document types. By applying Maxpooling, we were finally able to produce a classified class.

The documents shown in Fig. 3 are representative of the whole dataset that we have gathered. The picture presents three distinct sorts of documents: a personnel document, a graduation certificate, and a service letter. We included seven different sorts of university papers in the dataset that we compiled. The dataset includes 9125 scanned papers, is organized into seven categories, and has a storage capacity of more than 5.3 gigabytes.

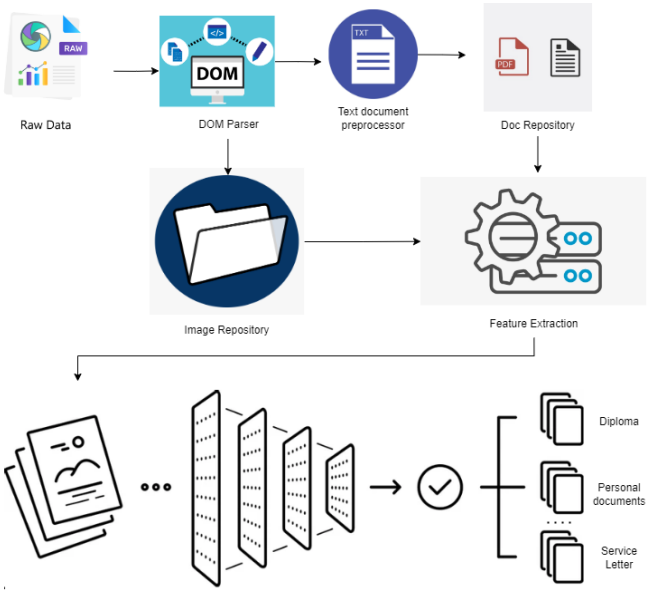

Fig. 1. Architecture of the system.

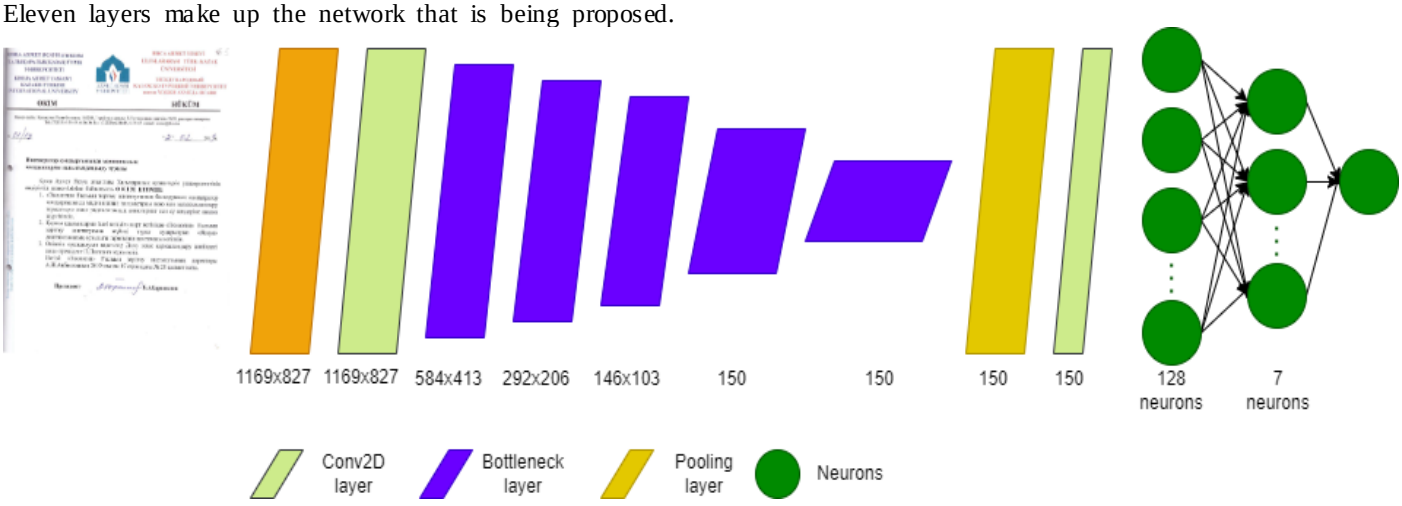

Fig. 2. Hybrid text / image classification using a multimodal classifier. End-to-end training is done for both textual and visual aspects.

ҚАЗАҚСТАН РЕСПУБЛИКАСЫНЫҢ ҮКІМЕТІ

ПРАВИТЕЛЬСТВО РЕСПУБЛИКИ КАЗАХСТАН

010000, Нұр-Сұлтан қ., Үкімет үйі

010000, город Нур-Султан, Дом Правительства

№ исх: 21-09/875 дз от: 12.03.2021
№ вх.943//21-09/875дз/ДЗ-38 от: 12.03.2021

Депутатам
Мажилиса Парламента
Республики Казахстан,
членам фракции
«Народной партии Казахстана»
(по списку)

На № ДЗ-38 от 10 февраля 2021 года

Уважаемые депутаты!

Рассмотрев Ваш депутатский запрос относительно создания Межведомственной комиссии по совершенствованию мер социальной поддержки населения, пострадавшего вследствие ядерных испытаний на Семипалатинском полигоне, сообщаю следующее.

В настоящее время объем и характер оказания социальной поддержки граждан, пострадавших вследствие ядерных испытаний на Семипалатинском испытательном ядерном полигоне *(далее – СИЯП)*, зависят от зоны радиационного поражения и периода проживания на нем.

В целях совершенствования мер социальной защиты граждан, пострадавших вследствие ядерных испытаний на СИЯП, Министерством труда и социальной защиты населения Республики Казахстан проведен ряд исследований *(«Ретроспективный и проспективный анализ медицинских и радиолого-гигиенических последствий бывшего Семипалатинского испытательного ядерного полигона для населения, прилегающих к нему территорий», «Оценка эффективности лечебных и реабилитационных мероприятий для населения, подвергшегося действию ионизирующего излучения в результате деятельности Семипалатинского испытательного ядерного полигона, и выработка предложений по их совершенствованию»)*.

По результатам проведенных исследований предложено оказывать медико-социальную помощь в зависимости от величины индивидуальной дозы облучения.

Данный подход позволит обеспечить оказание адресной социальной поддержки согласно полученным индивидуальным дозам облучения, а также с учетом доз облучения, полученных родителями.

Величина дозы облучения является модифицирующим фактором радиационного риска, которая зависит от особенностей человека, в том числе возраста на момент испытаний, продолжительности облучения. Следовательно, осуществлению каждого мероприятия по социальной защите должна

a) Service letter

b) Diploma certificate

Fig. 3. Samples of document types in the developed dataset.

V. Experimental Results

The investigation that was carried out primarily focused on seven key types of university papers that are a part of the cases that were dealt with by the STF. The following is a list of these categories, with their initial labels still attached to them:Diploma; Personal Documents; Journal of Accounting for Higher Education Diplomas; Service Letter; Order; Production Orders; Student Orders Diploma; Personal Documents; Journal of Accounting for Higher Education Diplomas;

A. Dataset

It is important to note that the court cases include a wide range of different types of documents, all of which have been filed under the name "Miscellaneous." In this regard, we developed an annotation tool that was used by a team of four lawyers for the purpose of manually classifying 8,139 pieces of paper. Fig. 1 presents a graphical breakdown of the proportion of articles that can be assigned to each of these categories. This chart can be found further down this page. In order to effectively train and evaluate machine learning systems, it is common practice to segment datasets into three distinct parts [19]. The terms "train," "validation," and "test" are used to refer to these specific parts of the dataset, respectively. We use stratified splits for each document class, making certain that the same proportions of class samples are included in each subset in order to maintain consistency. The following ratios were used, and Fig. 2 provides a more detailed breakdown of their application: 70% of the data will be used for the training set, 20% will be used for validation, and 10% will be used for the test set.

The gathered dataset is broken down into categories according to Fig. 4, which shows the distribution of the documents. There are seven different kinds of scanned papers that do not balance out. Student orders are the most common form of categorisation, accounting for 35 percent of all the papers that are scanned. The Journal of Accounting for higher education diplomas, personal documents, and production orders make up 20%, 19%, and 13%, respectively, of all the collected data. With 6%, 4%, and 13% indices, respectively, orders, production orders, and service letters make up the smallest share of the documents.

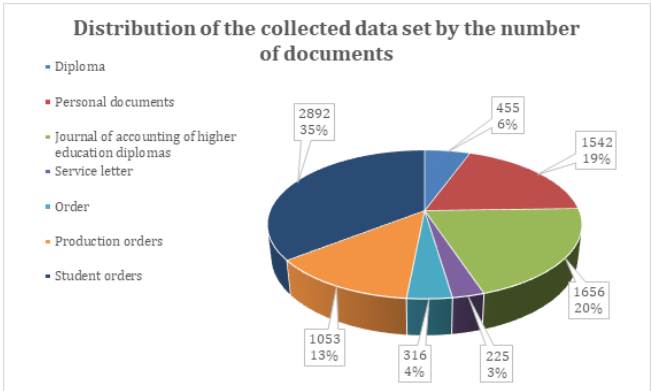

Fig. 4. Distribution of document classes in the dataset.

The whole of the material that was gathered is around 4.7 gigabytes worth of scanned pdf documents. The separation of these data into their respective volumes for each classification is shown in Fig. 5. The diploma, personal documents, and journal of accounting for higher education diplomas categories together account for 87.8% of the total volume of the dataset. This percentage is broken out as follows: 42.6%, 19%, and 26%. The remaining four categories of gathered papers make up 12.2% of the total.

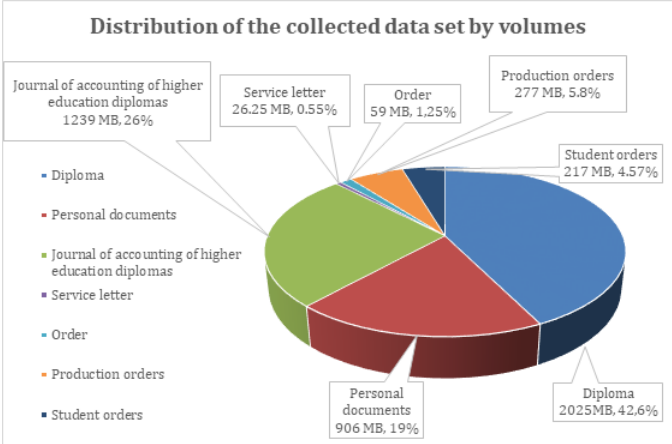

Fig. 5. Distribution of the collected dataset by volumes.

The distribution of the training, validation, and test sets for the proposed deep model is shown in Fig. 6. We cut the gathered dataset into three sections, which we referred to as the training set, the validation set, and the test set, and we allocated 70%, 20%, and 10% of the total space to each section, accordingly. As a result, separating the dataset into three distinct pieces enables us to achieve a high level of accuracy and determine whether or not the suggested model is suitable for use in actual settings.

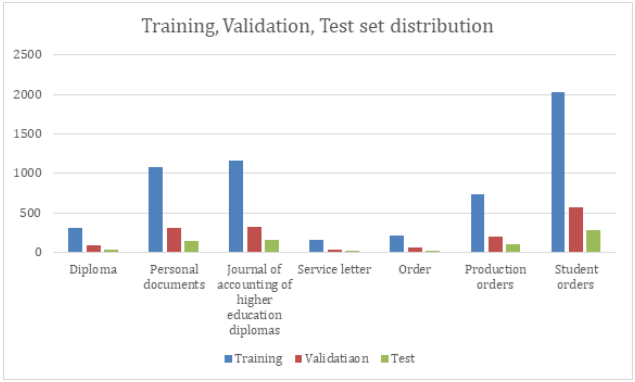

Fig. 6. Training, validation and test set distribution for each of the document classes.

The results of Fig. 5 and Fig. 6 are shown in Table I, which also provides an illustration of the distribution of the gathered dataset according to the number of scanned photographs and the volume of the data for each category. It enables us to comprehend the relationship between the number of photos and the volume of those images, as well as the quality of the document based on the type. As a result, after we have finished preparing the dataset, we may go on to training the model.

TABLE I. DATASET DESCRIPTION

Type of the document	Number	Volume
Diploma	455	2025 MB
Personal documents	1542	906 MB
Journal of accounting	1656	1.21 GB
Service letter	225	26.25 MB
Order	316	59.5 MB
Production orders	1053	277 MB
Student orders	2892	217 MB

B. Evaluation Parameters

In this part, our goal is to provide an explanation of evaluation parameters so that we may evaluate the suggested model and evaluate it in relation to other machine learning models. Accuracy, precision, recall, f-score, and area under the curve receiver operational characteristics were chosen as the five indicators to serve as the assessment criteria. (AUC-ROC) [20].

$$accuracy = \frac{TP+TN}{TP+FN+TN+FP} \quad (1)$$

$$precision = \frac{TP}{TP+FP} \quad (2)$$

$$recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 = \frac{2\,precision\ recall}{precision + recall} \quad (4)$$

The next part provides an assessment of the deep learning model that was presented for the classification of documents. After that, we will give the suggested model's confusion matrix, model accuracy, and validation accuracy. Fig. 7 is a demonstration of the model's accuracy during a period of one hundred epochs. According to the findings, the suggested model demonstrates great accuracy not only during training but also throughout testing.

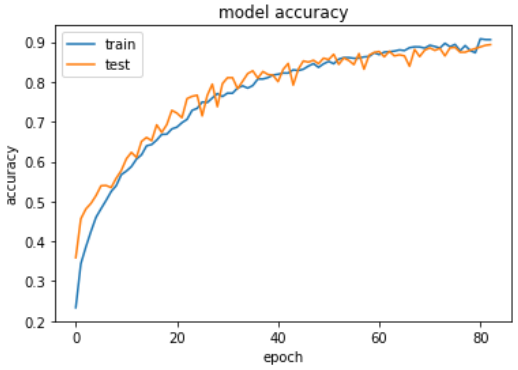

Fig. 7. Model accuracy.

Fig. 8 depicts the model loss over a period of one hundred epochs. The findings demonstrate that the model loss steadily declines as the number of epochs in the analysis is increased. In 80 epochs, the model loss in train and test shows less than 0.3, which suggests that the model is usable for actual instances and can be used for solving the automated document classification issue with a high degree of effectiveness.

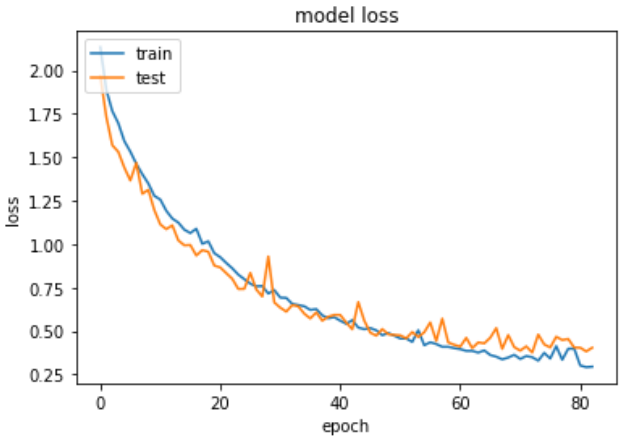

Fig. 8. Model loss.

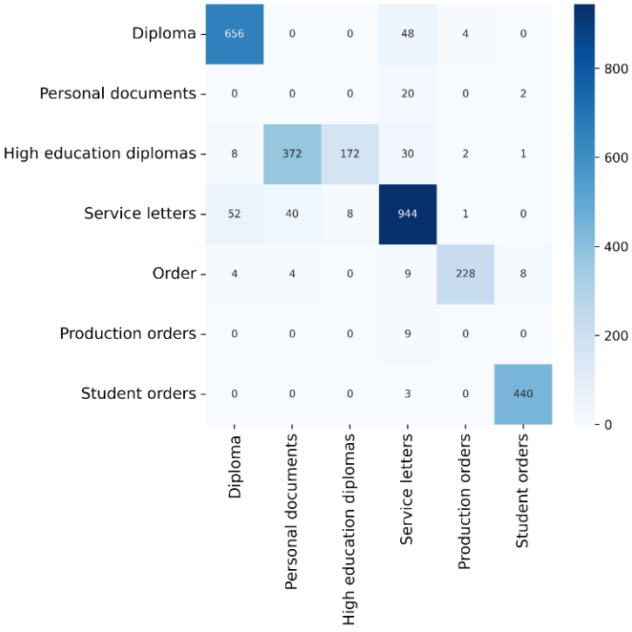

Fig. 9. Confusion matrix.

A performance assessment for the machine learning classification problem is shown in Fig. 9, which depicts a confusion matrix for each kind of classified text. This matrix indicates an evaluation of how well the issue was solved, and the output may comprise two or more classes. The table that follows provides an overview of the four distinct ways in which expected values might contrast with actual values. According to the findings, there is a relatively low incidence of mistakes and misunderstandings. The majority of the time and scanned documents have the appropriate categories assigned to them.

Table II presents a comparison between the deep learning model that has been presented and various machine learning techniques. According to the findings, the suggested model demonstrates the best performance when compared to the other techniques in each assessment parameter. The suggested model achieves an accuracy of 94.84%, precision of 94.79%, recall of 94.62%, F-score of 94.43%, and AUC-ROC of 94.07%. These results indicate that the proposed model is relevant in real life and that the provided dataset may be used to train machine learning and deep learning models to solve the document classification issue.

A comparison of the deep learning model that was developed with more conventional machine learning techniques for solving the document classification issue is shown in Fig. 10. We use five traditional algorithms—XGBoost, multilayer perceptron, random forest, support vector machines, and decision tree – to do a comparison between the proposed model and the existing models. We employ accuracy, precision, recall, F-score, ROC-AUC, and threshold as assessment parameters. Other parameters include recall. According to the conclusions drawn from the research, the suggested deep model achieves a high classification percentage across all assessment parameters. As a result, we are able to reach the conclusion that the deep model that was provided is suitable for use in the classification of academic publications.

Table III presents a comparison between the suggested technique and the most recent research available. Various studies have been established for the purpose of classifying scanned documents, including electronic health data, magnetic resonance scans, and handwritten historical manuscripts. With an accuracy of 94.84%, the educational papers of seven different categories may be categorized using the Conv2D model that was presented.

TABLE II. MODEL EVALUATION

Model	Accuracy	Precision	Recall	F-score	AUC-ROC
Proposed Model	**94.84%**	**94.79%**	**94.62%**	**94.43%**	**94.07%**
Random Forest	82.73%	82.13%	82.34%	81.12%	81.09%
XGBoost	81.77%	81.31%	81.37%	81.21%	81.17%
Support vector machine	82.36%	82.17%	82.06%	82.21%	82.11%
Multilayer perceptron	80.67%	80.54%	80.51%	80.28%	80.12%
Decision trees	76.45%	76.37%	76.28%	76.17%	76.19%

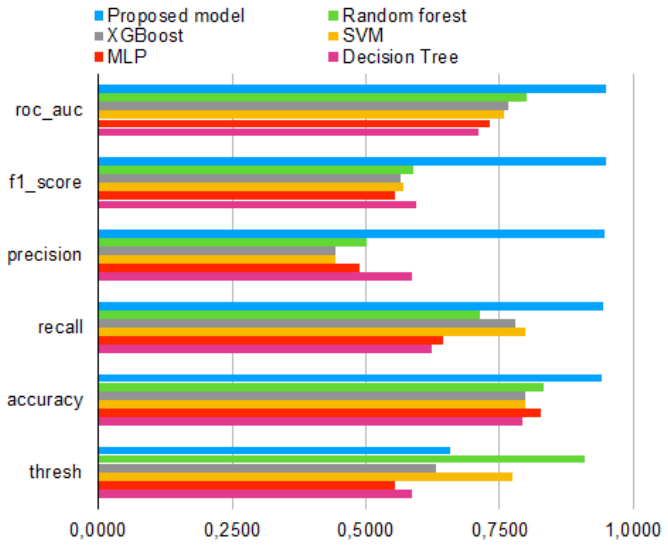

Fig. 10. Obtained results and comparison with the machine learning methods.

TABLE III. COMPARISON OF THE OBTAINED RESULTS WITH THE STATE-OF-THE-ART STUDIES

Study	Method	Documents	Dataset	Results
Proposed Model	**Proposed Model**	**9125 scanned documents of 7 types**	**Own dataset**	**94.84% accuracy, 94.79% precision, 94.62% recall, 94.43% F-score, 94.07% AUC-ROC**
[21]	ClinicalBERT	2988 scanned documents	Sleep study reports	AUCROC of 0.9523, Accuracy of 91.61%
[22]	Hand-written historical manuscripts	955 scanned reports	38 historical manuscript	98.5% segmentation rate
[23]	Automatic computational method	250,000 historical data	XMT datasets	82.06%
[24]	Two-level CNN	MRI	Open Access MRI data	92.30% accuracy
[25]	3D body scans	625 3D body scans	SizeKorea dataset	92% accuracy
[26]	Automated Paper Fingerprinting	306 paper images	Scanner image dataset	97% accuracy
[27]	Transfer Learning	Deep Transfer Learning	-	Accuracy: 0.8920
[28]	MLP	Multi-Page Documents	Digital image documents	Precision of 0.9030; F1-Score of 0.9380

VI. DISCUSSION

Digitized Document Classification using Multimodal Deep Learning is a research paper that proposes a novel approach to document classification using a combination of text and image features. In this discussion, we will analyze the advantages and limitations of this research paper, and also look into the future perspectives of this technology.

A. Advantages

One of the major advantages of this research paper is the use of multimodal deep learning for document classification. By combining both text and image features, the proposed method can classify documents more accurately and efficiently compared to traditional document classification methods. This is because documents often contain both textual and visual elements, and using a combination of both can improve the accuracy of classification [29].

Another advantage of this research paper is the use of CNNs and LSTM networks for image and text feature extraction, respectively. CNNs are known for their ability to extract high-level features from images, while LSTM networks can model the context of a sequence of words [30]. By using these two types of networks, the proposed method can extract both visual and semantic information from documents, leading to more accurate and robust classification.

Moreover, the proposed method is not limited to a specific type of document and can classify different types of documents such as invoices, resumes, and letters. This is because the method is trained on a large dataset of diverse documents, which makes it adaptable to different types of documents.

B. Limitations

One of the limitations of this research paper is the requirement of a large dataset for training. As with most deep learning approaches, the accuracy and performance of the

proposed method are heavily dependent on the size and quality of the dataset used for training. Therefore, obtaining a large and diverse dataset may be challenging, especially for specific document types.

Another limitation is the computational cost of the proposed method. Deep learning models are known to require a significant amount of computational resources, including high-end GPUs and large amounts of memory [31]. This may limit the applicability of the proposed method to smaller organizations or those with limited computational resources. The proposed method does not take into account the metadata associated with the documents, such as the author, date, and location. This information can be valuable for certain document classification tasks and not using it may lead to a decrease in accuracy.

C. Future Perspectives

Despite the limitations, the proposed method has several future perspectives. One potential application is in the field of document analysis and retrieval, where the ability to accurately classify and retrieve documents based on their content can be valuable [32]. This can be especially useful in organizations that deal with a large number of documents, such as legal firms and government agencies.

Another potential application is in the field of automated document processing [33]. By using the proposed method, organizations can automate the process of classifying and categorizing documents, leading to increased efficiency and reduced costs. This can be especially useful in industries such as finance and healthcare, where there is a significant amount of paperwork that needs to be processed. The proposed method can be extended to incorporate other modalities, such as audio and video, leading to more robust and accurate document classification. This can be especially useful in industries such as media and entertainment, where documents may contain different types of media.

Overall, Digitized Document Classification using Multimodal Deep Learning is a promising research paper that proposes a novel approach to document classification using a combination of text and image features [34]. The proposed method has several advantages, including increased accuracy and efficiency compared to traditional document classification methods. However, the method also has some limitations, including the requirement of a large dataset for training and the computational cost [35]. Despite these limitations, the proposed method has several future perspectives and can be applied in various industries, including document analysis and retrieval, automated document processing, and media and entertainment.

VII. Conclusion

In conclusion, Digitized Document Classification using Multimodal Deep Learning is a valuable contribution to the field of document classification. By combining text and image features using deep learning models, the proposed method can accurately and efficiently classify different types of documents.

Because it can help in the structuring of a document collection and describe the main subject of a document via the use of semantic multimedia analysis among concepts, the suggested method provides good performance in a range of contexts. This is because semantic multimedia analysis can help describe the primary topic of a document. These two qualities contribute, individually and together, to the accomplishment of high levels of performance. As a consequence of this, a large number of experiments were carried out in order to evaluate the efficiency of the many different topic identification tasks, and a discussion of the conclusions of those tests was provided. In this regard, the results enable us to say that the approach that we have described for identifying text themes is superior to the methods that are considered to be state-of-the-art, such as LSA and LDA, in relation to the specific topic in question. With regard to the visual subject identification, a number of different descriptors have been put through their paces and evaluated. In particular, the discoveries that seemed to have the greatest promise were chosen, and this was done on the basis of the features that were generated from the activation layer of the DNN model.

In addition, we demonstrated that it is possible to improve the overall work by using the most effective features of both techniques if one uses the strategy for identifying textual subjects in conjunction with the strategy for identifying visual themes. This was accomplished by combining the strategies for identifying textual topics and visual topics. Due to the fact that the system is modular and may be used more than once, it also has the capability to modify the proposed architecture in order to create various models for the task of topic identification. Either putting in brand new modules or enhancing the functionality of the ones that are currently there are both viable options for completing this job. All of the tests, which were conducted using a representative sample of online documents, have been successfully completed by the system. The design of the system, on the other hand, makes it possible to employ a number of different libraries that hold different kinds of multimedia content.

This approach has several advantages, including the ability to extract both visual and semantic information from documents, leading to more robust and accurate classification. However, the proposed method also has some limitations, including the requirement of a large dataset for training and the computational cost. Despite these limitations, the proposed method has several future perspectives and can be applied in various industries, including document analysis and retrieval, automated document processing, and media and entertainment. Overall, this research paper provides a foundation for further research in the field of document classification using multimodal deep learning.

Acknowledgment

The research team thanks Zhanseri Ikram for developing the document categorization system and deploying deep learning models for document categorization.

References

[1] H. Jain, S. Joshi, G. Gupta and N. Khanna, “Passive classification of source printer using text-line-level geometric distortion signatures from scanned images of printed documents,” Multimedia Tools and Applications, vol. 79, no. 11, pp. 7377-7400, 2020.

[2] S. Gupta and M. Kumar, "Forensic document examination system using boosting and bagging methodologies," Soft Computing, vol. 24, no. 7, pp. 5409-5426, 2020.

[3] A. Altayeva, B. Omarov, H.C. Jeong and Y.I. Cho, "Multi-step face recognition for improving face detection and recognition rate", Far East Journal of Electronics and Communications, vol. 16, no. 3, pp. 471-491, 2016.

[4] A. Mohanarathinam, S. Kamalraj, G. Venkatesan, R. Ravi and C. Manikandababu, "Digital watermarking techniques for image security: a review," Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 8, pp. 3221-3229, 2020.

[5] H. Kusetogullari, A. Yavariabdi, A. Cheddad, H. Grahn and J. Hall, "ARDIS: a Swedish historical handwritten digit dataset," Neural Computing and Applications, vol. 32, no. 21, pp. 16505-16518, 2020.

[6] N. Atallah, M. Toss, C. Verrill, M. Salto-Tellez, D. Snead et al., "Potential quality pitfalls of digitalized whole slide image of breast pathology in routine practice," Modern Pathology, vol. 35, no. 7, pp. 903-910, 2022.

[7] Omarov, B., Narynov, S., Zhumanov, Z., Gumar, A., & Khassanova, M. (2022). State-of-the-art violence detection techniques in video surveillance security systems: a systematic review. PeerJ Computer Science, 8, e920.

[8] Omarov, B., Tursynova, A., Postolache, O., Gamry, K., Batyrbekov, A., Aldeshov, S., ... & Shiyapov, K. (2022). Modified UNet Model for Brain Stroke Lesion Segmentation on Computed Tomography Images. Computers, Materials & Continua, 71(3).

[9] N. Sharma, R. Sharma and N. Jindal, "Machine learning and deep learning applications-a vision," Global Transitions Proceedings, vol. 2, no. 1, pp. 24-28, 2021.

[10] Omarov, B., Altayeva, A., & Cho, Y. I. (2017). Smart building climate control considering indoor and outdoor parameters. In Computer Information Systems and Industrial Management: 16th IFIP TC8 International Conference, CISIM 2017, Bialystok, Poland, June 16-18, 2017, Proceedings 16 (pp. 412-422). Springer International Publishing.

[11] A. Singh, S. Thakur, A. Jolfaei, G. Srivastava, M. Elhoseny et al., "Joint encryption and compression-based watermarking technique for security of digital documents," ACM Transactions on Internet Technology (TOIT), vol. 21, no. 1, pp. 1-20, 2021.

[12] J. Latham, M. Ludlow, A. Mennito, A. Kelly, Z. Evans et al., "Effect of scan pattern on complete-arch scans with 4 digital scanners," The Journal of Prosthetic Dentistry, vol. 123, no. 1, pp. 85-95, 2020.

[13] Kaldarova, B., Omarov, B., Zhaidakbayeva, L., Tursynbayev, A., Beissenova, G., Kurmanbayev, B., & Anarbayev, A. (2023). Applying Game-based Learning to a Primary School Class in Computer Science Terminology Learning. In Frontiers in Education (Vol. 8, p. 26). Frontiers.

[14] Tursynova, A., & Omarov, B. (2021, November). 3D U-Net for brain stroke lesion segmentation on ISLES 2018 dataset. In 2021 16th International Conference on Electronics Computer and Computation (ICECCO) (pp. 1-4). IEEE.

[15] G. Malaperdas, "Digitization in archival material conservation processes," European Journal of Engineering and Technology Research, vol. 6, no. 4, pp. 30-32, 2021.

[16] D. Zhang, S. Zhao, Z. Duan, J. Chen, Y. Zhang et al. "A multi-label classification method using a hierarchical and transparent representation for paper-reviewer recommendation," ACM Transactions on Information Systems (TOIS), vol. 38, no. 1, pp. 1-20, 2020.

[17] T. Hegghammer, "OCR with tesseract, amazon textract, and google document AI: a benchmarking experiment," Journal of Computational Social Science, vol. 5, no. 1, pp. 861-882, 2022.

[18] M. Revilla-León, M. Sadeghpour and M. Özcan, "An update on applications of 3D printing technologies used for processing polymers used in implant dentistry," Odontology, vol. 108, no. 3, pp. 331-338, 2020.

[19] C. Mangano, F. Luongo, M. Migliario, C. Mortellaro and F. Mangano, "Combining intraoral scans, cone beam computed tomography and face scans: the virtual patient," Journal of Craniofacial Surgery, vol. 29, no. 8, pp. 2241-2246, 2018.

[20] A. Haleem and M. Javaid, "3D scanning applications in medical field: a literature-based review," Clinical Epidemiology and Global Health, vol. 7, no. 2, pp. 199-210, 2019.

[21] E. Hsu, I. Malagaris, Y. Kuo, R. Sultana and K. Roberts, "Deep learning-based NLP data pipeline for EHR-scanned document information extraction," JAMIA Open, vol. 5, no. 2, pp 1-12, 2022.

[22] G. BinMakhashen and S. Mahmoud, "Historical document layout analysis using anisotropic diffusion and geometric features," International Journal on Digital Libraries, vol. 21, no. 3, pp. 329-342, 2020.

[23] S. Darwish and H. ELgohary, "Building an expert system for printer forensics: A new printer identification model based on niching genetic algorithm," Expert Systems, vol. 38, no. 2, pp. 1-14, 2021.

[24] K. Aderghal, K. Afdel, J. Benois-Pineau and G. Catheline, "Improving alzheimer's stage categorization with Convolutional Neural Network using transfer learning and different magnetic resonance imaging modalities," Heliyon, vol. 6, no. 12, pp. 1-13, 2020.

[25] K. Lee H. Song and S. Kim, "Categorization of lower body shapes of abdominal obese men using a script-based 3D body measurement software," Fashion and Textiles, vol. 7, no. 1, pp. 1-16, 2020.

[26] S. Khaleefah, S. Mostafa, A. Mustapha and M. Nasrudin, "The ideal effect of gabor filters and uniform local binary pattern combinations on deformed scanned paper images," Journal of King Saud University-Computer and Information Sciences, vol. 33, no. 10, pp. 1219-1230, 2021.

[27] A. Jadli, M. Hain and A. Hasbaoui, "An improved document image classification using deep transfer learning and feature reduction," International Journal, vol. 10, no. 2, pp. 549-557, 2021.

[28] G. Nagy, "Twenty years of document image analysis in PAMI," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 38-62, 2000.

[29] W. Hassan, Y. Yusoff and N. Mardi, "Comparison of reconstructed rapid prototyping models produced by 3-dimensional printing and conventional stone models with different degrees of crowding," American Journal of Orthodontics and Dentofacial Orthopedics, vol. 151, no. 1, pp. 209-218, 2017.

[30] K. Adam, A. Baig, S. Al-Maadeed, A. Bouridane and S. El-Menshawy, "KERTAS: dataset for automatic dating of ancient Arabic manuscripts," International Journal on Document Analysis and Recognition, vol. 21, no. 4, pp. 283-290, 2018.

[31] X. Liu, S. Guo, B. Yang, S. Ma, H. Zhang et al.. "Automatic organ segmentation for CT scans based on super-pixel and convolutional neural networks," Journal of Digital Imaging, vol. 31, no. 5, pp. 748-760, 2018.

[32] C. Ben Rabah, G. Coatrieux and R. Abdelfattah, "Automatic source scanner identification using 1D convolutional neural network," Multimedia Tools and Applications, vol. 81, no. 16, pp. 22789-22806, 2022.

[33] L. Di Angelo, P. Di Stefano and E. Guardiani, "A review of computer-based methods for classification and reconstruction of 3D high-density scanned archaeological pottery," Journal of Cultural Heritage, vol. 56, no. 1, pp. 10-24, 2022.

[34] A. Kumar, H. Goodrum, A. Kim, C. Stender, K. Roberts et al., "Closing the loop: automatically identifying abnormal imaging results in scanned documents," Journal of the American Medical Informatics Association, vol. 29, no. 5, pp. 831-840, 2022.

[35] A. Guha, A. Alahmadi, D. Samanta, M. Khan and A. Alahmadi, "A multi-modal approach to digital document stream segmentation for title insurance domain," IEEE Access, vol. 10, no. 1, pp. 11341-11353, 2022.