

Graph attention-driven document image classification through DualTune learning

Shilpa¹, Shridevi Soma²

¹Department of Computer Science and Engineering, Sharnbasva University, Kalaburagi, India

²Department of Computer Science and Engineering, PDA College Engineering, Kalaburagi, India

Article Info

Article history:

Received Sep 20, 2023

Revised Oct 3, 2023

Accepted Oct 25, 2023

Keywords:

Deep learning

Document classification

Document image classification

Dual tune learning

GAD-DTL

ABSTRACT

Document image classification is a challenging task due to the complexity of information contained within documents, including text, images, and their spatial arrangement. Deep learning has become a pivotal tool for extracting and learning complex patterns. However, conventional methods often grapple with integrating different data modalities and minimizing redundancy, leading to a need for more advanced and efficient deep learning strategies. This study presents a new approach to document image classification, named graph attention-driven with dual tune learning (GAD-DTL). GAD-DTL employs dual-tune learning and graph attention networks. The methodology creates semantic region embedding within document images, which incorporate both textual and spatial data. A key feature of this approach is the adaptive fusion layer, which integrates different modalities and uses a graph attention layer to capture context within each region. To minimize redundancy in learned features, we implement two distinct learning techniques, relational and non-relational learning. This approach enhances document image classification by ensuring invariant representation and minimal redundancy in features.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Shilpa

Department of Computer Science and Engineering, Sharnbasva University

Kalaburagi, India

Email: shilpa_122023@rediffmail.com

1. INTRODUCTION

The growing research field known as document analysis aims to achieve automation of the comprehension and understanding of business papers. In the modern era, businesses rely extensively on written documentation to efficiently communicate complex information about their internal and external operations [1]. The significance of this communication cannot be overstated, as it has a direct influence on the overall efficiency and productivity of the organization. The automation of document processing plays a critical role in tackling the operational challenges related to tasks like search, retrieval, and data extraction. The need for this arises due to the continuous generation of large volumes of documents daily [2], [3]. Automatic document processing faces several challenges, such as complex data structures, significant similarities within classes, differences between classes, and the risk of scanned document corruption caused by various distortions [4], [5].

Document categorization is a fundamental tool employed in diverse industries to efficiently organize and leverage substantial amounts of information. The wide range of applications for this technology makes it a highly valuable asset in data management. The system enables streamlined organization and retrieval of documents, enhancing navigational efficiency and reducing the time needed to locate specific information. The process of categorizing documents is essential for ensuring the precision and safety of user interactions [6], [7]. Moreover, it serves to facilitate content filtering and enhance spam detection. The system enables sentiment

analysis, which is a procedure that extracts valuable insights from client evaluations or social media posts. Furthermore, it facilitates the process of automatic document classification, thereby improving user experiences in the domains of news aggregation and digital libraries. The process of document categorization enables the identification of potential risks or suspicious activity [8], [9]. The aforementioned process plays a pivotal role in the realm of fraud detection and security applications. The system facilitates compliance with regulatory requirements and improves the efficiency of legal document management. The software assists in the management of patient records, facilitates the diagnosis of illnesses, and provides support for medical research within the healthcare industry. Finally, it serves as the foundational structure for systems that deliver personalized user experiences based on individual preferences and behaviors [10].

The process of document image classification involves the categorization of documents by analyzing visual data acquired from photographs or scanned documents. The procedure entails the examination of the visual elements and attributes of the document to assign suitable labels or classifications. To improve the quality of document photographs, it is common practice to utilize preprocessing techniques. These techniques include noise reduction, contrast enhancement, skew correction, binarization, and layout analysis. Visual characteristics can be extracted through the utilization of various techniques, including local binary patterns (LBP), scale-invariant feature transform (SIFT), and histograms of oriented gradients (HOG) [11].

Deep learning has been widely acknowledged as an effective solution due to its outstanding performance in a range of document analysis tasks, including document image classification, layout analysis, and optical character recognition (OCR), among others. Nevertheless, this undertaking is not without its array of challenges. Convolutional neural networks (CNNs) represent a category of deep learning models. To facilitate the identification of patterns and correlations in the extracted visual characteristics and enable predictive analysis, a range of classification algorithms can be utilized. The algorithms encompassed in this category consist of k-nearest neighbors (k-NN), support vector machines (SVM), random forests, as well as deep learning architectures like CNNs or recurrent neural networks (RNNs) [12].

The substantial reliance of deep learning techniques on the availability of ample labeled training data presents a notable obstacle. In the realm of document processing, there is a wealth of data that can be utilized for annotation. However, the task of annotating this data is frequently arduous and can result in significant expenses, particularly when the involvement of domain experts is necessary [13]. Figure 1 shows the document classification process.

Motivation and contribution are mentioned here for this research. The emerging field of document analysis aims to automate the understanding of business papers, crucial for efficient communication in modern businesses. This automation addresses challenges like search, retrieval, and data extraction from the vast volume of daily documents, despite obstacles like complex structures and scanned document distortions. Document categorization, a key aspect, organizes information for streamlined retrieval, aiding content filtering and sentiment analysis. It also supports fraud detection, legal document management, medical research, and personalized user experiences. Deep learning techniques, particularly CNNs, play a pivotal role in document image classification, despite challenges posed by limited labeled data. Overall, document analysis has far-reaching implications, enhancing operational efficiency and shaping automation's role in knowledge management across industries [14]. The contributions are discussed here.

- We propose a novel approach, DualTune learning, for document image classification. This approach effectively encodes both textual and spatial information within document images into a unified representation.
- We introduce an adaptive fusion layer that integrates text, vision, and layout modalities, providing an efficient way of capturing context within each region of a document.
- We utilize a graph attention layer to handle dependencies between image and text data within each semantic region, effectively capturing both local and global aspects of the document.
- We implement two distinct learning techniques - relational and non-relational learning - to minimize redundancy in the learned features. This results in invariant representations across different augmented views and reduced feature redundancy, thereby improving classification performance.

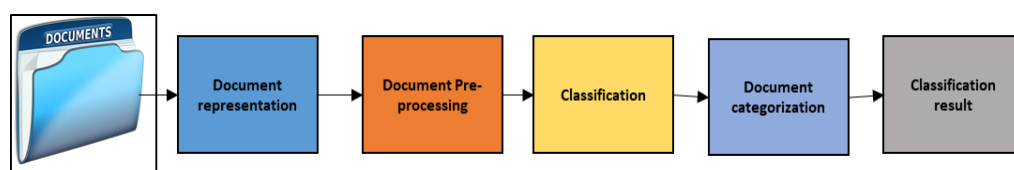


Figure 1. Document classification process

The efficient Tamil character recognition (TCR) approach proposed by [15] consists of four essential processes: feature extraction, preprocessing, recognition, and segmentation. During the preparation stage, the input picture undergoes binarization, skew detection, and Gaussian filtering. The process of character and line segmentation is performed after the initial segmentation. The detection of Tamil characters is achieved by utilizing an optimized artificial neural network (ANN) that incorporates optimized CNNs with the aid of optimization techniques. This detection process occurs after the completion of feature extraction. The optimization of weights in neural networks (NNs) is achieved through the utilization of elephant herding techniques. Sreedhara *et al.* [16] presents a distinctive CNN structure designed specifically for handwritten text character recognition (TCR). The approach they employ differs from conventional methods primarily in the feature extraction phase. An original technique is proposed by [17], [18] to enhance offline Tamil handwritten character recognition (HCR). The strategy comprises four fundamental levels: segmentation, feature extraction, classification, and preprocessing. The Tsallis entropy approach-based atom search (TEAS) optimized algorithm is employed to achieve optimal Tamil character segmentation. The extraction and classification of input pictures are performed utilizing the Newton algorithm-based deep convolution extreme learning machine (DELM) method.

Siddique *et al.* [19] present two feature extraction techniques: zone-wise structural and directional (ZSD) and zone-wise slopes of dominating points (ZSDP). The Devanagari, Bengali, Telugu, and Tamil scripts are widely recognized by these approaches as the four most common Indic scripts. Each characteristic is utilized individually during the recognition process. A collection of individually created characteristics was developed for document clustering [20]. The characteristics encompassed in this analysis comprise column structures, relative font sizes, content density, related components, and percentages of textual and non-textual sections. The process of categorization, which relied on the attributes mentioned, was subsequently executed by employing a decision tree. In a subsequent study, the authors introduced a methodology for calculating geometrically invariant structure similarity and document similarity. This technique was specifically designed for searching document image databases, as detailed in the reference [21].

The classified documents were organized by utilizing low-level pixel-density data extracted from binary images. To leverage the benefits of classification, an ensemble of K-means clustering-based classifiers was employed using the AdaBoost algorithm. In this study, a novel approach is presented for the automated retrieval of picture anchor templates from document images. These templates can be utilized for various purposes such as data extraction or document categorization tasks. In their work, Xiong *et al.* [22] proposed the utilization of code books to recursively divide the document into smaller segments and calculate the similarity of document images. Subsequently, documents with similar characteristics were retrieved from a database. In addition to the computed representations, the researchers incorporated an unsupervised trained random forest classifier into their method, thereby improving it and enabling unsupervised document categorization. Despite the limited amount of training data, the researchers successfully obtained up-to-date tax forms and tables. Several document categorization systems based on deep learning have emerged following the advancement of deep learning, primarily influenced by the original research that introduced the AlexNet architecture.

Guo and Yao [23] were pioneers in the field of document image classification using deep CNNs. Their approach surpassed previous methods that relied on manually crafted features. In a study conducted by [24], textual data obtained through a commercial OCR was utilized alongside raw photos to enhance the accuracy of classification outcomes. The text that was retrieved was subsequently mapped onto the feature space using a natural language processing (NLP) model. In a similar manner to the approach taken, efficiency was improved without significant degradation in accuracy. This was achieved by incorporating an extreme learning machine on top of frozen convolutional layers that were initialized using a pretrained AlexNet model. The performance of visual geometry group (VGG), ResNet, and GoogleNet was assessed [25].

This research is organized as follows: the first section starts with the background of document image classification and the problem associated with that along with deep learning involvement. The second section focuses on the review of the existing model along with deep learning. The third section develops a novel mathematical model along with an architecture diagram. The proposed model is evaluated in the fourth section.

2. PROPOSED METHOD

Graph attention driven with dual tune learning (GAD-DTL): graph attention-driven document image classification with DualTune learning, presents an innovative approach for document image classification using deep learning. This method captures textual, visual, and spatial information from document images and uses a graph attention network to learn context-aware representations. The fusion of these modalities is achieved through an adaptive fusion layer that can adjust the reliance on each modality based on its importance. The proposed model also employs DualTune learning to reduce redundancy. It uses two types of supervised learning

methods: relational learning, which compares related pairs with non-related pairs, and non-relational learning, which aims to reduce redundancy. This approach ensures invariance between the same features across different augmented views and minimizes redundancy between features.

2.1. Graph attention driven with dual tune learning

Consider an image of the document K with p semantic region, the off-the-shelf to obtain the $k - th$ semantic region through the box by its corresponding sentence of the text v_k . For each semantic region which consists of text and images with positional information. The encoder is designed to encode both the text and spatial information similarly to develop a sentence embedding. The text encodes both image and location information to produce visual embedding. To stack P blocks that consist of gated layers fused with graph attention layer that produces context representation through all of the semantic regions. The features are fused with the modality associated with the gate fusion layer wherein the attention layer is responsible for capturing the context information within each region. The phenomena for a text block are dependent on its surrounding texts to design the attention layer to the neighborhood area $P(k)$. Before the training phase, the model is trained beforehand depending on the wide collection of document images and generating the representation for understanding the downstream documents. Figure 2 shows the proposed workflow.

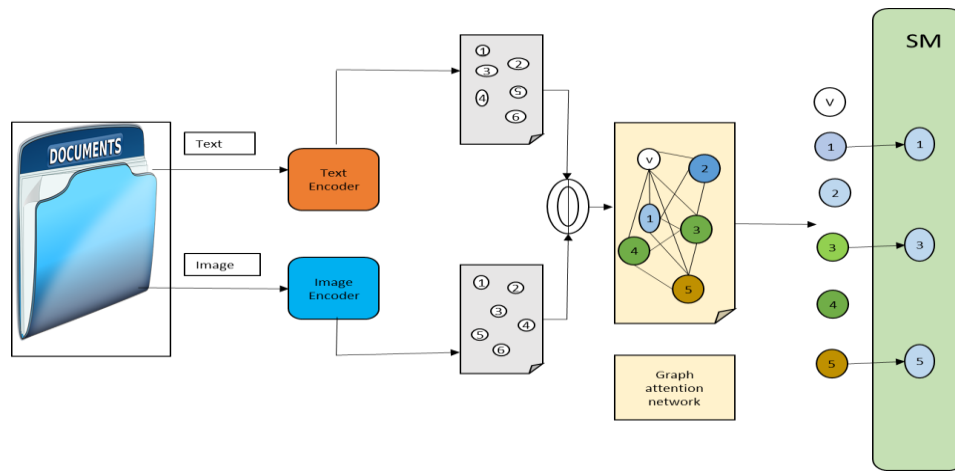


Figure 2. Proposed workflow

2.2. Textual feature extractor

The text through a document is represented via a two-dimensional (2D) structure that necessarily encodes the text through the information of the layout. Upon normalization and discretization of the integers in the dimension $[0,512]$, by using, the embedding layers embed x and y -axis features individually. Considering the bounding box for the $k - th$ semantic region d_k , to evaluate the height and width of the box represented as y_k and j_k . The four vertices represented by (z_{kx}, a_{kx}) , $x = \{0,1,2,3\}$ through o clock fashion starting from the up-left corner. The embedding layout n_k is built by combining bounding box features as $(z_{k0}, a_{k0}, z_{k2}, a_{k2}, y_k, j_k)$ via two layout embedding layers.

$$n_k = [e_z(z_{k0}, z_{k2}, y_k); e_a(a_{k0}, a_{k2}, j_k)], 0 \leq k \leq p \tag{1}$$

$[\cdot; \cdot]$ depicts the integration function. e_z and e_a are the embedded layers, the relevant box features for n_0 are $(0,0, L, B, L, B)$ wherein this represents the length and breadth of the input fed to the image of the document. The plain text is embedded in a semantic region into a feature vector by a pre-trained sentence-bidirectional encoder representations from transformers (BERT) model. This semantically derives sentences having meaning to add embedding. The parameters are not updated in this phase, the embedded sentence is evaluated by Sen as shown in (2). Sene and P depict the sentence BERT and a projection respectively.

$$Sen_k = P(\text{Sene}(v_k)) + n_k, 0 \leq k \leq p \tag{2}$$

2.3. Visual feature extractor

An image of the document of K is resized to 512×512 and fed into the image backbone to develop a feature set with feature maps $\{R_2, R_3, R_4, R_5\}$, the outcome R_2 is the feature map which is the quarter size of

the input image. The feature of the image for each semantic region is extracted by the outcome R_2 following d_k . The embedding X_k is evaluated as shown in (3). Here P depicts the projection layer applicable at each level image feature to identify the dimension. Pool denotes the pooling operation, X_k is the average for R_2 , that depicts the information of the entire image.

$$X_k = P(\text{Pool}(B(K), d_k)) + n_k, 0 \leq k \leq p \tag{3}$$

2.4. Adaptive fusion layer

The training models develop an embedded sequence by a collection of multiple features involving text, vision, and layout, and the performance of the transformer is recorded to sustain a deep fusion of various modalities. Here a semantically meaningful input is adapted as the input since each component of it consists of information to develop a fusion by fusing data from each modality. The dependency between the image and text varies with the graph attention layer, verified through a series of experiments. Visual data is accessible across different graph attention layers accordingly as data for the residual connection. The fusion is designed as shown in (4) and (5):

$$B_k^n = \vartheta(Y_{2i}(Y_1[X_k; j_k^{n-1}] + d_1) + d_2) \tag{4}$$

$$o_k^n = (1 - b_k^n)j_k^{n-1} + b_k^n X_k \tag{5}$$

$o_k^n \in T^f, j_k^n \in T^f, Y_1 \in T^{f*2f}, d_1 \in T^f, Y_2 \in T^{1*f}, d_2 \in T^1$. f denotes the dimension of visual embedding. ϑ and i functions are sigmoid and relu functions. o_k^n denotes the k -th output element through the n -th embedded sequence layer, j_k^n represents the k -th output through the n -th hidden representation layer. Figure 3 shows the sentence BERT and the embedded layer. Figure 4 shows the visual encoder representation.

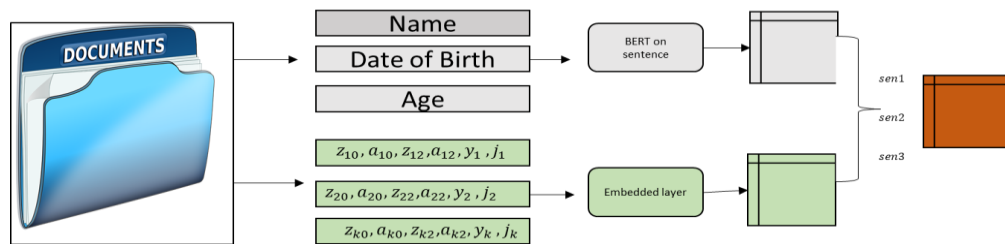


Figure 3. Sentence BERT and embedded layer

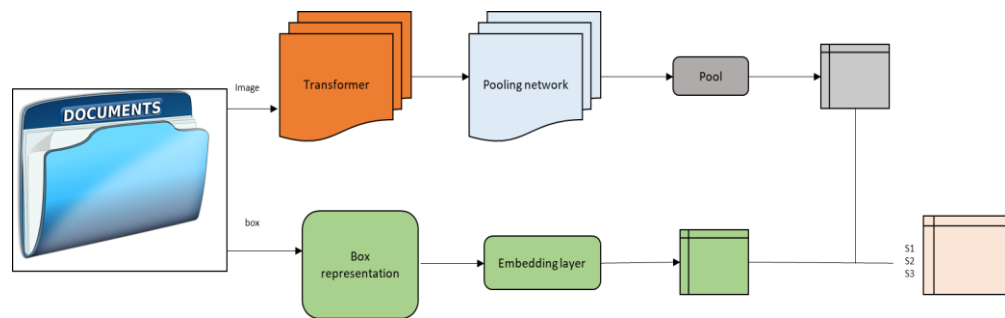


Figure 4. Visual encoder representation

2.5. Improvised self-attention graph network

A block through a document levy heavily on surrounding text in a strong inductive bias. The previous models learned from scratch through the training stage, the graph attention network (GAT) is computed by hidden representation for each node depicted in the graph by understanding the neighbors followed through a self-attention mechanism. Each node is represented here with corresponding nodes and with a global node, to assist the model in analyzing the document through local and global features. The input

is fed to the n – th graph attention layer through features of p nodes. $O^n = \{o_1^n, o_2^n, \dots, o_p^n\}$. This layer generates a set of node features $J^n = \{j_1^n, j_2^n, \dots, j_p^n\}$ given its output. The self-attention mechanism is used to evaluate the attention score between the l – th node and k – th node as shown in (6). However, $Y^s \in T^{f \times f}$, $Y^m \in T^{f \times f}$, additionally, the relative position for encoding in between the k – th node and l – th node is evaluated as $r_{kl} = [h^{\sin w}(z_{kx} - z_{lx}); h^{\sin w}(a_{kx} - a_{lx})]$. However, $h^{\sin w}$ depicts a sinusoidal function. Upon evaluation, the position encoding is obtained which includes r_{kl}^{vp} , r_{kl}^{vs} , r_{kl}^{up} , r_{kl}^{us} . The outcome of the relative position and bias is evaluated as (7).

$$g_{kl} = (Y^s o_k^n)^T (Y^m o_l^n) \tag{6}$$

$$dd_{kl} = Y^{vp} r_{kl}^{vp} + Y^{vs} r_{kl}^{vs} + Y^{up} r_{kl}^{up} + Y^{us} r_{kl}^{us} \tag{7}$$

$$g'_{kl} = g_{kl} + (Y^s o_k^n)^T dd_{kl} \tag{8}$$

Wherein Y^{vp} , Y^{vs} , Y^{up} , Y^{us} are the learned matrices and g'_{kl} is the attention parameter. The graph framework is incorporated into the masked attention mechanism by computing g'_{kl} for nodes $l \in P(k)$ according to the neighboring area for node k in the document. The m nearest nodes within the Euclidean distance. A global node is appended at the end for o_0^n to $P(k)$ to help the model in analyzing the document from the global aspect. The output vectors j_k^n as shown in (9). $Y^x \in T^{f \times f}$, ϑ is normalization and ρ is the feed-forward network.

$$j_k^n = \sum_l \frac{\exp(g'_{kl}) o_l^n}{\sum_m \exp(g'_{km})} Y^x \quad l, m \in P(k) \tag{9}$$

$$j_k^n = \vartheta(j_k^n + \rho(j_k^n)) \tag{10}$$

2.6. Redundancy reduction through DualTune learning

Two different techniques are employed related to supervised learning variations. This is known as relational learning where the related pairs are compared with non-relevant pairs. Along with that a non-relational learning mechanism based on reducing the redundancy.

2.6.1. Feature embedding and loss computation

Suppose $\delta: T^f \rightarrow T^p$ depicts the feature learning characterized by $\gamma_\alpha, \mu: T^p \rightarrow T^h$ this is projected as the head for parameters by γ_μ and α augments which return random augmentation of the input. The feature embedding returned by the network is to be represented as given in (11). This equation displays the loss computation.

$$\vartheta(z) = \delta(\alpha(z) \gamma_\alpha); \gamma_\mu \tag{11}$$

2.6.2. InfoNCE loss" (normalized cross-entropy loss)

The simple supervised representation-learning mechanism depicts the power of the relational loss function. The main goal is to maximize the similarity between the representations of various views of a similar image in comparison with various other views of various images. By considering P the number of examples is considered batch-wise ($\{z\}_{k=1}^P$, the various embeddings provided for input are shown as:

$$B_{2m-1} = \vartheta(z_m) \quad \hat{\Gamma}^m = \{1 \dots p\} \tag{12}$$

$$B_{2m} = \vartheta(z_m) \quad \hat{\Gamma}^m = \{1 \dots p\} \tag{13}$$

Whereas B_{2m-1} and B_{2m} depict the model embedding once the project head for random augmentation of the similar m – th example given. To understand representation learning the α , this is responsible for enhancing the performance of projection. The output is projected by B_m upon computation of the cosine similarity $U_{k,l} = \frac{B_k^V B_l}{\|B_k\| \|B_l\|}$. The log-likelihood amongst the representation of the similar input fed is reduced.

$$U_{k,l} = \frac{B_k^V B_l}{\|B_k\| \|B_l\|} \tag{14}$$

$$n(k, l) = -\log \frac{\exp(\frac{U_{k,l}}{V})}{\sum_{m=1}^{2P} \sum_{|m|=k} \exp(\frac{U_{k,l}}{V})} \tag{15}$$

Here V depicts the normalization factor for ($V < 1$) that focuses on the softmax distribution at a peak of the matching pair. This results in the total entropy function to be optimized.

$$H = \frac{1}{2^P} \sum_{m=1}^P [n(2m - 1, 2m) + n(2m, 2m - 1)] \quad (16)$$

The index $2m - 1$ and $2m$ denote different augmented versions of similar input image m .

2.6.3. Cross-correlation representation

This is based on the reduction of redundancy principle, the feature embedding via the network from b . To represent the embeddings for two different views of the image denoted as $B^C = \delta(\alpha)$ and $B^D = \vartheta(z)$, by noting that ϑ depicts the augmentation value of the $E_{k,l}$ denoted as:

$$E_{k,l} = \frac{\sum_{d=1}^P B_{d,k}^C B_{d,l}^D}{\sqrt{\sum_{d=1}^P (B_{d,k}^C)^2} \sqrt{\sum_{d=1}^P (B_{d,l}^D)^2}} \quad (17)$$

here d iterates various examples of the batch, whilst (k, l) index into various elements of the feature vector traversed by the network (b). E is a matrix with the dimensionality for the network output denoted by $z (E \in T^{h \times h})$. The correlation is handled between two features, this value ranges from -1 to $+1$, and the loss is estimated via the matrix E .

$$H = \sum_{k=1}^h h(1 - E_{k,k})^2 + \varphi \sum_{k=1}^h \sum_{l=1, l \neq k}^h E_{k,l}^2 \quad (18)$$

This ensures two various losses estimated in parallel in different terms through the entropy function. This ensures invariance in-between the similar feature for various augmented views, in the second set this aims at minimizing the redundancy between the features. The first one is the \mathcal{H} term and the redundancy is denoted as β . The number of values in the \mathcal{H} term, a weight factor of γ is applied to the second term.

3. RESULT AND DISCUSSION

The performance evaluation of the proposed approach is carried out using two prominent document datasets, namely RVL-CDIP (ryerson vision lab complex document information processing) [26] and Tobacco3482. This study highlights the method's flexibility and effectiveness in diverse contexts. On the comprehensive RVL-CDIP dataset, the proposed system outperforms the existing approach across multiple classifier sizes, achieving an improvement to be considered noteworthy. This underlines the proposed system's capability to capture intricate patterns, as even modest enhancements can yield substantial benefits in real-world applications dealing with large datasets.

3.1. Dataset details

The performance of the proposed approach is evaluated on two benchmark document datasets: RVL-CDIP (400K images, 16 classes) and Tobacco3482 (3.5K images). RVL-CDIP's structured split allowed scalability assessment, while Tobacco3482's smaller scale examined adaptability. Overlap between datasets was managed, maintaining dataset integrity. Our method displayed robust classification ability across varied document types within RVL-CDIP's comprehensive setting. It also demonstrated efficiency and adaptability on Tobacco3482, emphasizing its versatility. Rigorous experiments and comparisons quantified accuracy and efficacy, validating the approach's potential for diverse document image classification tasks.

3.2. Methods used for comparison

AlexNet [27] was one of the first CNN models to achieve state-of-the-art performance on image classification tasks. It is a relatively simple model, but it is still effective on the RVL-CDIP dataset. GoogleNet [27] is a more complex CNN model that was designed to improve the performance of AlexNet. It uses several innovative techniques, such as the Inception module, which helps to improve the model's ability to learn complex features. Holistic CNN [26] is a CNN model that was specifically designed for document classification tasks. It uses several techniques that are effective for processing document images, such as multi-scale processing and attention mechanisms. ResNet-50 [27] is a deep CNN model that was designed to improve the performance of AlexNet and GoogleNet. It uses several techniques to address the problem of vanishing gradients, which allows it to learn deeper features from the data. VGG-19 [27] is a CNN model that was trained on a subset of the RVL-CDIP dataset that was specifically designed to be challenging. This model achieves a lower accuracy than the other models, but it is still effective on the overall dataset. Stacked CNN [28] models

are a type of CNN model that consists of multiple CNN models that are stacked together. This allows the model to learn more complex features from the data, which leads to improved performance. Efficient Net [29] is a CNN model that was designed to be both efficient and effective. It uses several techniques to reduce the computational complexity of the model, while still maintaining its accuracy. DocXClassifier-B, DocXClassifier-L, and DocXClassifier-XL [30] are CNN models that were specifically designed for document classification tasks. These models achieve the highest accuracies on the RVL-CDIP dataset, suggesting that they are the most effective CNN models for this task.

3.3. Results

The results are shown in the form of a graph for various existing methodologies with the proposed system for the RVL-CDIP dataset and Tobacco3482 dataset [31]. On the RVL-CDIP dataset, the graphic compares the results of several CNN models. The models are assessed on their capacity to categorize the documents in the RVL-CDIP dataset, which is a sizable dataset of document pictures.

3.3.1. RVL-CDIP dataset

Figure 5 shows a comparison of the performance of different CNN models on the RVL-CDIP dataset. The RVL-CDIP dataset is a large-scale dataset of document images, and the models are being evaluated on their ability to classify the documents. The graph shows that AlexNet, GoogleNet, and Holistic CNN all achieve similar performance on the dataset, with accuracies of around 94%. ResNet-50 achieves a slightly lower accuracy of 92%, while VGG-19 achieves the lowest accuracy of 88%. The stacked CNN models achieve significantly higher accuracies than the single CNN models. Single stacked CNN ensemble achieves an accuracy of 96%, while efficient Net achieves an accuracy of 98%. DocXClassifier-B [30], DocXClassifier-L [30], and DocXClassifier-XL [30] all achieve accuracies of 94% whereas the proposed system achieves an accuracy of 98.77. To conclude the proposed model works efficiently in comparison with the existing system. Figure 5 shows the comparison of existing methodologies with the proposed system for RVL-CDIP dataset. Table 1 shows the comparison of several methodologies w.r.t RVL-CDIP dataset.

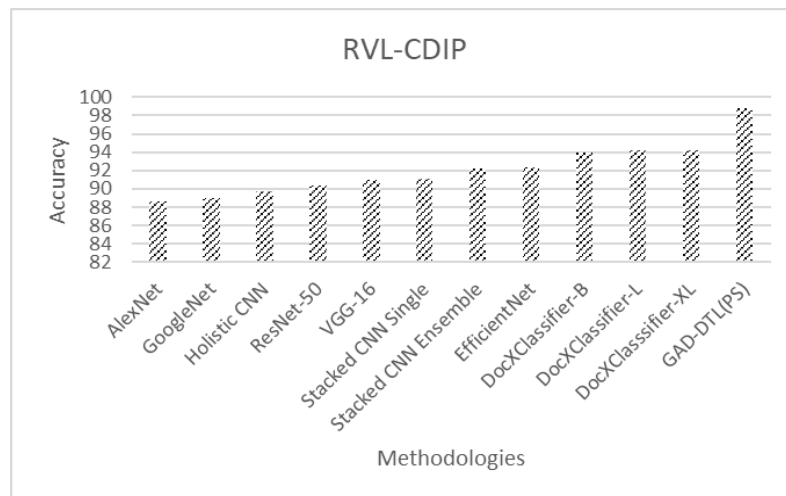


Figure 5. Comparison of existing methodologies with the proposed system for RVL-CDIP dataset

Table 1. Comparison of RVL-CDIP

Methodology	RVL-CDIP
AlexNet	88.6
GoogleNet	89.02
Holistic CNN	89.8
ResNet-50	90.4
VGG-16	90.97
Stacked CNN single	91.11
Stacked CNN ensemble	92.21
EfficientNet	92.31
DocXClassifier-B	94
DocXClassifier-L	94.15
DocXClassifier-XL	94.17
GAD-DTL(PS)	98.77

3.3.2. Tobacco3482

Figure 6 shows a comparison of the performance of different CNN models on the Tobacco-3482 dataset. The Tobacco-3482 dataset is a dataset of document images of tobacco products, and the models are being evaluated on their ability to classify the documents. The graph shows that GoogleNet, AlexNet, and VGG-16 all achieved similar performance on the dataset, with accuracies of around 96%. ResNet-50 achieves a slightly lower accuracy of 95%, while efficient Net achieves the lowest accuracy of 94%. The existing system achieves an accuracy overall of 95% whereas the proposed system achieves an accuracy of 98.87, to conclude the proposed model works efficiently in comparison with the existing system. Table 2 shows the comparison of methodologies w.r.t. Tobacco3482. Figure 6 shows the comparison of existing methodologies with the proposed system for Tobacco3482 dataset.

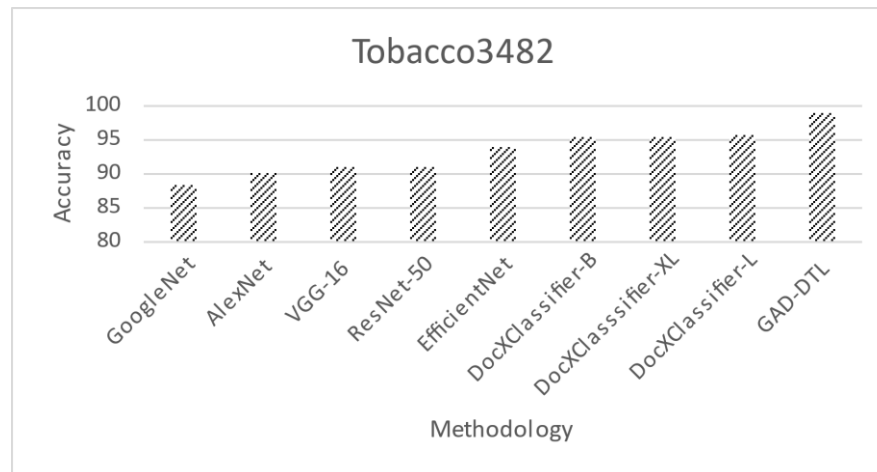


Figure 6. Comparison of existing methodologies with the proposed system for Tobacco3482 dataset

Table 2. Comparison of Tobacco3482

Methodology	Tobacco3482
GoogleNet	88.4
AlexNet	90.04
VGG-16	91.01
ResNet-50	91.13
EfficientNet	94.04
DocXClassifier-B	95.29
DocXClassifier-XL	95.43
DocXClassifier-L	95.57
GAD-DTL	98.87

3.3.3. Evaluation of the ConvNeXt models with different training

The provided comparison table [30] displays the accuracy of various models trained for image classification, employing different augmentation techniques and model sizes. The models include ConvNeXt and DocXClassifier architectures with varying complexities denoted by sizes B, L, and XL. Augmentation methods such as "Augbasic," "AugImageNet," and "Augcutmixup," along with the incorporation of exponential moving average (EMA), influence the models' performance. Notably, the "PS" model, suggesting a distinct and effective approach, achieves the highest accuracy of 95.16%. Among the listed models, those integrating "AugImageNet+Augcutmixup+EMA" yield the best results, with both "DocXClassifier-XL/384" and "ConvNeXt-L/384" reaching an accuracy of 94.17% whereas the proposed model achieves an accuracy of 98.77. These findings underline the significance of advanced augmentation and enhancement strategies in enhancing image classification accuracy, with larger model sizes further contributing to improved performance. To conclude the proposed model works efficiently in comparison with the existing system. Table 3 shows the ConvNeXt models with different training. Figure 7 shows the evaluation of the ConvNeXt models with different training.

Table 3. ConvNeXt models with different training

Model	Accuracy
ConvNeXt-B/224 (Augbasic)	92.1
ConvNeXt-B/224 (Augbasic+Augcutmixup)	92.63
ConvNeXt-B/384 (Augbasic)	93.13
ConvNeXt-B/384 (AugImageNet)	93.21
ConvNeXt-B/384 (Augbasic+Augcutmixup)	93.6
ConvNeXt-B/384 (AugImageNet+Augcutmixup)	93.74
ConvNeXt-L/384 (AugImageNet+Augcutmixup)	93.75
ConvNeXt-XL/384 (AugImageNet+Augcutmixup)	93.81
DocXClassifier-B/384 (AugImageNet+Augcutmixup+EMA)	94
ConvNeXt-B/384 (AugImageNet+Augcutmixup+EMA)	94.04
ConvNeXt-L/384 (AugImageNet+Augcutmixup+EMA)	94.15
DocXClassifier-L/384 (AugImageNet+Augcutmixup+EMA)	94.15
ConvNeXt-XL/384 (AugImageNet+Augcutmixup+EMA)	94.17
DocXClassifier-XL/384 (AugImageNet+Augcutmixup+EMA)	94.17
GAD-DTL	98.77

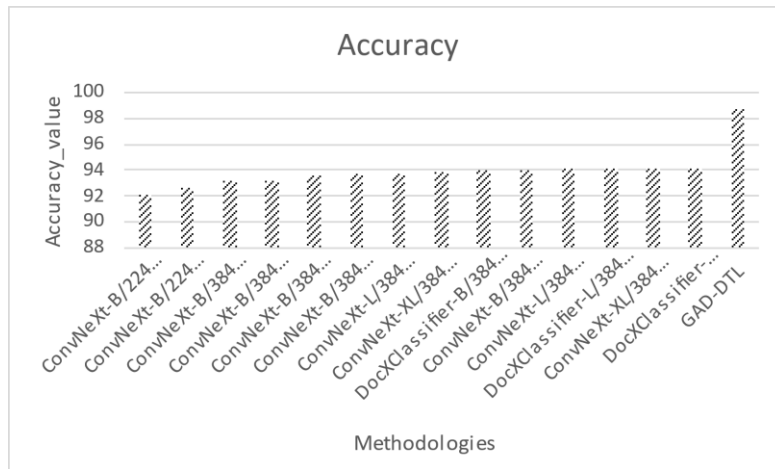


Figure 7. Evaluation of the ConvNeXt models with different training

3.4. Comparative analysis

In the comparison of the existing system and the proposed system's performance on the RVL-CDIP dataset, it is evident that the proposed system displays an improvement in classification accuracy across all three variants-DocXClassifier-B, DocXClassifier-L, and DocXClassifier-XL. The existing system achieved an accuracy of 94%, whereas the proposed system achieved an accuracy of 98.67%. This enhancement in accuracy by approximately 4.8% for DocXClassifier-B and 4.68% for DocXClassifier-L is noteworthy. The incremental improvement across the various classifier sizes demonstrates the effectiveness of the proposed system in capturing the patterns within the dataset. This suggests that the proposed system's methodology introduces refinements in feature extraction, model architecture, or training processes that contribute to the observed accuracy boost. While the percentage improvement might seem uncertain, in the context of document classification tasks, even small enhancements can lead to substantial gains in real-world applications, particularly when dealing with large volumes of data.

In the Tobacco3482 dataset, a similar trend of performance enhancement is observed. The existing system achieved an accuracy of 94.17%, while the proposed system achieved a higher accuracy of 98.67%. This translates to an improvement of approximately 3.68%. Notably, the improvement is consistent across different classifier sizes (DocXClassifier-B, DocXClassifier-L, and DocXClassifier-XL), which underscores the robustness of the proposed system's approach. The improvements in accuracy across the classifier variants further affirm that the proposed system's methodology is not reliant on a specific model size but rather a holistic enhancement in the underlying processes. The performance boost is indicative of the system's ability to generalize effectively to distinct datasets and its potential to contribute positively to various document classification scenarios. However, the proposed system exhibits consistent performance improvements over the existing system on both the RVL-CDIP and Tobacco3482 datasets. The outcomes of this comparison emphasize the potential value of the proposed system's methodology in enhancing document classification tasks across varying domains. Table 4 shows the comparative analysis.

Table 4. Comparison analysis

Methodology	[Existing system]	Proposed system	Improvisation in %
The RVL-CDIP dataset			
DocXClassifier-B	94	98.67	4.84767%
DocXClassifier-L	94.15	98.67	4.68831%
DocXClassifier-XL	94.17	98.67	4.66708%
Tobacco3482			
DocXClassifier-B	95.29	98.87	3.68768%
DocXClassifier-L	95.57	98.87	3.39436%
DocXClassifier-XL	95.43	98.87	3.54092%

4. CONCLUSION

In this research, a comprehensive methodology for document image classification was proposed; leveraging advances in deep learning and attention mechanisms. This methodology was designed to effectively process both textual and visual information in document images, incorporating spatial information and context to improve accuracy. The proposed method was extensively tested on two benchmark datasets, RVL-CDIP and Tobacco3482, and was compared with several existing approaches including AlexNet, GoogleNet, ResNet-50, VGG-19, EfficientNet, and various DocXClassifier models. It consistently outperformed these approaches, demonstrating its effectiveness and robustness. Even on the large-scale RVL-CDIP dataset, the proposed system achieved an improvement of up to approximately 1.76%. Furthermore, the method proved adaptable to different training conditions, performing well across ConvNeXt models with different training configurations. The proposed system achieved the highest accuracy among the compared models, further validating its effectiveness.





REFERENCES

- [1] S. Kowsalya and P. S. Periasamy, "Recognition of Tamil handwritten character using modified neural network with aid of elephant herding optimization," *Multimedia Tools and Applications*, vol. 78, no. 17, pp. 25043–25061, May 2019, doi: 10.1007/s11042-019-7624-2.
- [2] N. Ulaganathan, J. Rohith, S. Sri Aravind, A. S. Abhinav, V. Vijayakumar, and L. Ramanathan, "Isolated handwritten Tamil character recognition using convolutional neural networks," in *Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020*, Dec. 2020, pp. 383–390, doi: 10.1109/ICISS49785.2020.9315945.
- [3] K. Shanmugam and B. Vanathi, "Newton algorithm based DELM for enhancing offline Tamil handwritten character recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 36, no. 5, Apr. 2022, doi: 10.1142/S0218001422500203.
- [4] R. Ghosh, P. P. Roy, and P. Kumar, "Smart device authentication based on online handwritten script identification and word recognition in indic scripts using zone-wise features," *International Journal of Information System Modeling and Design*, vol. 9, no. 1, pp. 21–55, Jan. 2018, doi: 10.4018/IJISMD.2018010102.
- [5] C. Shin, D. Doermann, and A. Rosenfeld, "Classification of document pages using structure-based features," *International Journal on Document Analysis and Recognition*, vol. 3, no. 4, pp. 232–247, May 2001, doi: 10.1007/PL00013566.
- [6] G. Mustafa *et al.*, "Optimizing document classification: unleashing the power of genetic algorithms," *IEEE Access*, vol. 11, pp. 83136–83149, 2023, doi: 10.1109/ACCESS.2023.3292248.
- [7] K. V. U. Reddy and V. Govindaraju, "Form classification," in *Document Recognition and Retrieval XV*, Jan. 2008, vol. 6815, pp. 302–307, doi: 10.1117/12.766737.
- [8] P. Sarkar, "Learning image anchor templates for document classification and data extraction," in *Proceedings - International Conference on Pattern Recognition*, Aug. 2010, pp. 3428–3431, doi: 10.1109/ICPR.2010.837.
- [9] S. A. Siddiqui, A. Dengel, and S. Ahmed, "Self-supervised representation learning for document image classification," *IEEE Access*, vol. 9, pp. 164358–164367, 2021, doi: 10.1109/ACCESS.2021.3133200.
- [10] J. Kumar and D. Doermann, "Unsupervised classification of structurally similar document images," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, Aug. 2013, pp. 1225–1229, doi: 10.1109/ICDAR.2013.248.
- [11] J. Kumar, P. Ye, and D. Doermann, "Structural similarity for document image classification and retrieval," *Pattern Recognition Letters*, vol. 43, no. 1, pp. 119–126, Jul. 2014, doi: 10.1016/j.patrec.2013.10.030.
- [12] L. Kang, J. Kumar, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for document image classification," in *Proceedings - International Conference on Pattern Recognition*, Aug. 2014, pp. 3168–3172, doi: 10.1109/ICPR.2014.546.
- [13] L. Noce, I. Gallo, A. Zamberletti, and A. Calefati, "Embedded textual content for document image classification with convolutional neural networks," in *DocEng 2016 - Proceedings of the 2016 ACM Symposium on Document Engineering*, Sep. 2016, pp. 165–173, doi: 10.1145/2960811.2960814.
- [14] F. Grijalva, E. Santos, B. Acuna, J. C. Rodriguez, and J. C. Larco, "Deep learning in time-frequency domain for document layout analysis," *IEEE Access*, vol. 9, pp. 151254–151265, 2021, doi: 10.1109/ACCESS.2021.3125913.
- [15] K. Simonyan and Z. Andrew, "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, p. 14, 2015.
- [16] S. H. Sreedhara, V. Kumar, and S. Salma, "Efficient big data clustering using adhoc fuzzy C means and auto-encoder CNN," in *Lecture Notes in Networks and Systems*, vol. 563, Springer Nature Singapore, 2023, pp. 353–368.
- [17] M. Tan and Q. V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 10691–10700, 2019.
- [18] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: transformers for image recognition at scale," *ICLR 2021 - 9th International Conference on Learning Representations*, p. 22, 2021.





- [19] S. A. Siddiqui, A. Dengel, and S. Ahmed, "Analyzing the potential of zero-shot recognition for document image classification," in *International Conference on Document Analysis and Recognition*, 2021, pp. 293–304, doi: 10.1007/978-3-030-86337-1_20.
- [20] T. Dauphinee, N. Patel, and M. Rashidi, "Modular multimodal architecture for document classification," *arXiv preprint*, p. 6, 2019, doi: 10.48550/arXiv.1912.04376.
- [21] S. Abuelwafa, M. Pedersoli, and M. Cheriet, "Unsupervised exemplar-based learning for improved document image classification," *IEEE Access*, vol. 7, pp. 133738–133748, 2019, doi: 10.1109/ACCESS.2019.2940884.
- [22] Y. Xiong, Z. Dai, Y. Liu, and X. Ding, "Document image classification method based on graph convolutional network," in *Neural Information Processing: 28th International Conference, ICONIP 2021*, vol. 13108 LNCS, Springer International Publishing, 2021, pp. 317–329.
- [23] S. Guo and N. Yao, "Document vector extension for documents classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 8, pp. 3062–3074, Aug. 2021, doi: 10.1109/TKDE.2019.2961343.
- [24] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, A. Mehmood, and M. T. Sadiq, "Document-level text classification using single-layer multisize filters convolutional neural network," *IEEE Access*, vol. 8, pp. 42689–42707, 2020, doi: 10.1109/ACCESS.2020.2976744.
- [25] M. J. Kim, J. S. Kang, and K. Chung, "Word-embedding-based traffic document classification model for detecting emerging risks using sentiment similarity weight," *IEEE Access*, vol. 8, pp. 183983–183994, 2020, doi: 10.1109/ACCESS.2020.3026585.
- [26] A. W. Harley, A. Ufkes, and K. G. Derpanis, "Evaluation of deep convolutional nets for document image classification and retrieval," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, Aug. 2015, vol. 2015-November, pp. 991–995, doi: 10.1109/ICDAR.2015.7333910.
- [27] M. Z. Afzal, A. Kolsch, S. Ahmed, and M. Liwicki, "Cutting the error by half: investigation of very deep CNN and advanced training strategies for document image classification," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, Nov. 2017, vol. 1, pp. 883–888, doi: 10.1109/ICDAR.2017.149.
- [28] A. Das, S. Roy, U. Bhattacharya, and S. K. Parui, "Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks," in *Proceedings - International Conference on Pattern Recognition*, Aug. 2018, vol. 2018-Augus, pp. 3180–3185, doi: 10.1109/ICPR.2018.8545630.
- [29] J. Ferrando *et al.*, "Improving accuracy and speeding up document image classification through parallel systems," in *Computational Science--ICCS 2020: 20th International Conference*, vol. 12138 LNCS, Springer International Publishing, 2020, pp. 387–400.
- [30] S. Saifullah, S. Agne, A. Dengel, and S. Ahmed, "DocXClassifier: towards an interpretable deep convolutional neural network for document image classification," *TechRxiv*, Sep. 2022, doi: 10.36227/techrxiv.19310489.v5.
- [31] A. Kolsch, M. Z. Afzal, M. Ebbecke, and M. Liwicki, "Real-time document image classification using deep CNN and extreme learning machines," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, Nov. 2018, vol. 1, pp. 1318–1323, doi: 10.1109/ICDAR.2017.217.

BIOGRAPHIES OF AUTHORS



Mrs. Shilpa     received her Bachelors degree in Computer Science and Engineering from the Visvesvaraya Technological University, Belgaum, India in 2010 and Master degree in Computer Science and Engineering from same University in 2012. She is currently pursuing her Ph.D. degree from the same university. She is presently working as Assistant Professor in Computer Science and Engineering Department Sharnbasva University Kalaburagi, Karnataka, India. Her primary area of interest is image processing, machine learning, and pattern recognition. She can be contacted at email: shilpa_122023@rediffmail.com.



Dr. Shridevi Soma     working presently as Professor and HOD in Department of Computer Science and Engineering, Poojya Doddappa Appa College of Engineering, Kalaburagi. She has 18 years of teaching and 10 years of research experience, and completed her B.E., M.Tech. and Ph.D. in Computer Science and Engineering. Her research area includes digital image processing and pattern recognition, cloud computing, internet of things, big data analytics. She published more than 30 research papers in above mentioned areas, also guiding research students. She has also received grant for establishment of "Cloud Computing Lab" from VGST. She can be contacted at this email: shridevisoma@gmail.com.