machine learning & knowledge extraction

MDPI

Article

VisFormers—Combining Vision and Transformers for Enhanced Complex Document Classification

Subhayu Dutta [1], Subhrangshu Adhikary [2,*] and Ashutosh Dhar Dwivedi [3]

[1] Department of Computer Science & Engineering, Dr. B.C. Roy Engineering College, Durgapur 713206, West Bengal, India; duttasuvo90@gmail.com
[2] Department of Research & Development, Spiraldevs Automation Industries Pvt. Ltd., Raiganj 733123, West Bengal, India
[3] Cyber Security Group, Aalborg University, DK-2000 Copenhagen, Denmark; addw@es.aau.dk
* Correspondence: subhrangshu.adhikary@spiraldevs.com; Tel.: +91-700-179-7064

Abstract: Complex documents have text, figures, tables, and other elements. The classification of scanned copies of different categories of complex documents like memos, newspapers, letters, and more is essential for rapid digitization. However, this task is very challenging as most scanned complex documents look similar. This is because all documents have similar colors of the page and letters, similar textures for all papers, and very few contrasting features. Several attempts have been made in the state of the art to classify complex documents; however, only a few of these works have addressed the classification of complex documents with similar features, and among these, the performances could be more satisfactory. To overcome this, this paper presents a method to use an optical character reader to extract the texts. It proposes a multi-headed model to combine vision-based transfer learning and natural-language-based Transformers within the same network for simultaneous training for different inputs and optimizers in specific parts of the network. A subset of the Ryers Vision Lab Complex Document Information Processing dataset containing 16 different document classes was used to evaluate the performances. The proposed multi-headed VisFormers network classified the documents with up to 94.2% accuracy, while a regular natural-language-processing-based Transformer network achieved 83%, and vision-based VGG19 transfer learning could achieve only up to 90% accuracy. The model deployment can help sort the scanned copies of various documents into different categories.

Keywords: VGG19; Transformers; multi-headed neural network; optical character reader; complex document classification

Citation: Dutta, S.; Adhikary, S.; Dwivedi, A.D. *VisFormers*—Combining Vision and Transformers for Enhanced Complex Document Classification. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 448–463. https://doi.org/10.3390/make6010023

Received: 9 December 2023
Revised: 9 February 2024
Accepted: 14 February 2024
Published: 16 February 2024

1. Introduction

Document classification is a vital process in the field of information management and retrieval [1]. It involves categorizing documents into predefined classes or categories based on their content, enabling the efficient organization, search, and analysis of vast amounts of textual data. The digital revolution and the explosion of online material have significantly increased the importance of this process in recent years [2]. Document classification is a potent technique for streamlining information access, as information overload is a persistent problem. The urgency of this work is evident as the volume of digital information continues to explode. In recent news, it has been reported that over 2.5 quintillion bytes of data are generated every day, and by 2025, it is estimated that global data will reach 163 zettabytes [3]. Most of these data require more organization, making it hard to determine what is essential. Our work attempts to sort and organize these data, facilitating better and quicker decision making using modern technologies.

Complex documents contain a mix of text, images, labels, tables, and various elements, which present a formidable challenge in document classification. These documents can include memos, newspapers, letters, and more, and effectively categorizing them is critical

for the rapid digitization of diverse content [4]. The difficulty arises from the fact that these scanned images look very similar as they share standard page colors, text characteristics, and paper textures. Their classification based on so few contrasting characteristics further increases the challenge.

Optical character recognition (OCR) and natural language processing (NLP) are two transformative technologies that have revolutionized the way we interact with and extract information from textual content [5]. OCR is a technology that permits the conversion of printed or handwritten text into machine-encoded text, as its name suggests. It scans text-containing physical documents or photos, analyzes the data, and translates the characters into a digital format that computers can understand [6]. On the other hand, NLP is a branch of artificial intelligence that focuses on the interaction between computers and human language. It empowers computers to understand, interpret, and generate human language in a way that is both meaningful and contextually relevant [7]. In addition to facilitating human–computer interaction, NLP is essential for data analysis because it enables computers to interpret and extract information from massive amounts of unstructured text data.

Transfer vision is the term used to describe a machine learning model's capacity to transfer knowledge from one task or domain to another, notably in computer vision [8]. This idea is crucial to increasing the adaptability and versatility of AI systems. Transfer vision uses previously trained models and their ingrained properties to effectively complete new, related tasks [9]. For instance, a model initially trained to identify items in photographs can be fine-tuned to identify particular species of animals in shots of wildlife. Compared to training a model from scratch for every new task, this technique saves time and computational resources [10]. A wide range of applications, such as driverless vehicles, medical imaging, and security monitoring, are significantly impacted by transfer vision.

In the realm of enhanced complex document classification, several powerful algorithms play pivotal roles, each exhibiting distinct characteristics. InceptionV3 (Incv3) employs a sophisticated architecture with multiple parallel pathways, enabling it to capture intricate details in diverse document structures. VGG16, renowned for its simplicity and effectiveness, utilizes a deep convolutional network to discern hierarchical features, making it adept at recognizing complex document patterns. ResNet50, known for its deep residual learning, excels in handling intricate relationships between document elements, enhancing accuracy [11]. MobileNet, designed for efficiency on mobile devices, balances accuracy and speed in document classification tasks. Lastly, the Transformer model, renowned for its attention mechanism, captures long-range dependencies within documents, offering a unique approach to understanding contextual relationships [12]. Collectively, these algorithms contribute to the advancement of enhanced complex document classification by addressing various challenges and nuances inherent in diverse document structures.

Commercially, advanced technologies such as intelligent recognition framework (IRF), enterprise resource planning (ERP), and customer relationship management (CRM) systems play a crucial role in managing document circulation and optimizing workflow efficiency. These systems analyze and process various types of data, including scanned documents, to create organized databases. While ERP and CRM systems excel in managing structured data and streamlining workflow processes, they often face limitations when handling unstructured data, such as printed complex documents. These systems may struggle to accurately classify and process documents containing handwritten annotations, diverse elements like figures and tables, or varying layouts [13]. According to various reports, organizations manage an average of 300,000 documents per employee per year, and knowledge workers spend about 20% of their time searching for information across different documents and systems [14]. This large number of documents, often in printed form and possibly with handwritten notes, can make it difficult for advanced systems to understand them fully [15]. As a result, document circulation and information management efficiency may be compromised, leading to delays and errors in workflow operations. The need for

a specialized document classification model becomes apparent given the prevalence of complex printed documents in various industries [16].

The state-of-the-art works perform complex document classification, but there is a big challenge regarding the fact that all documents have similar color gradients and texture [17]. This challenge limits the performance of the state-of-the-art models. Furthermore, only some attempts have been made to use OCR for this purpose [18]. However, that also requires better classification performance, given that many documents might have similar kinds of written text. However, the document classes can be difficult to classify based on their spatial information. Therefore, a challenge exists: **"Can we develop a highly efficient model that can process both vision and language properties of the complex document simultaneously to provide faster and more accurate classification than the state of the art?"**. This paper attempts to answer this question, and the primary contributions of the work are:

- To develop a fast and efficient complex document classification model.
- To combine computer vision and natural language processing networks in a single model for complex document classification.
- Facilitating different types of inputs and enabling different optimizers in different parts of the network while training.
- To benchmark the performance with the state-of-the-art methods using the standard complex document classification dataset RVL-CDIP.

The paper's structure is as follows: In Section 2, recent efforts in document classification and related works are summarized. Section 3 elaborates on the steps taken in experimentation. Section 4 presents results for the proposed vision Transformer. Section 5 focuses on results obtained from individual vision and NLP methods. Finally, Section 6 summarizes the key findings and concludes the research.

2. Related Works

Many researchers have worked on classifying documents using convolutional neural networks (CNNs) and transfer learning [19]. The study by Abdullah et al. [20] addresses Arabic document classification using machine learning. Convolutional neural networks (CNN) with a character-level model attained the greatest accuracy of 98% throughout the authors' experiments with various algorithms, outperforming earlier techniques and proving their usefulness, particularly in social media environments. Shuo et al. [21] address the growing need for efficient technical document classification in technology organizations by highlighting the availability of multimodal information in various papers. In their study, they provide TechDoc, a brand-new multimodal deep-learning architecture that blends relationships between text, images, and documents. This strategy provides scalable solutions for document management in tech firms, outperforming unimodal systems and current benchmarks.

Various other researchers have focused on organizing documents through optical character recognition (OCR). Despite the widespread use of the language in India, Srinivasa Rao et al. [22] addresses the need for more progress in Telugu optical character recognition (OCR) systems. They introduce the ITP-STTR model, leveraging deep learning for improved Telugu text recognition. With challenging character combinations and a dearth of recent OCR advancements, their research fills a critical gap in Telugu OCR development, yielding superior results. In addressing the urgent problem of road safety and car theft in India, Anuj et al. [23] strongly emphasize the significance of automatic number plate detection (ANPD). They use TensorFlow for training and testing, using a variety of datasets to detect plates with an accuracy of 85%.

Certain researchers emphasize the utilization of natural language processing (NLP) and Transformers for document classification. Authors like Irfan et al. [24] address employers' challenges in selecting suitable job applicants from a large pool of resumes. Their study introduces a natural language processing (NLP) and machine learning (ML)-based resume classification system (RCS) to automate and expedite the categorization process. Multiple

ML algorithms and NLP techniques are evaluated, with support vector machine (SVM) classifiers achieving over 96% accuracy in multi-class resume classification. Aroush et al. [25] use deep learning and already-trained language models to handle the complex and time-consuming task of patent classification. Using datasets like USPTO-2M and M-patent, their work investigates the fine-tuning of models like BERT, XLNet, RoBERTa, and ELECTRA to achieve cutting-edge performance on multi-label patent categorization. Iqra et al. [26] address the task of multi-label emotion classification in short social media posts. Using transfer learning, they investigate the utilization of single- and multiple-attention processes in LSTM and Transformer networks (such as XLNet, DistilBERT, and RoBERTa). Their innovative technique exceeds current benchmarks, with the RoBERTa-MA model, in particular, obtaining 62.4% accuracy on the English SemEval-2018 E-c dataset.

Another group of studies delves into the notable progressions in Transformer-based models for natural language processing (NLP) and document classification. Yang et al. [27] propose a novel model, D2BFormer, utilizing the Vision Transformer framework to address the critical task of degraded document binarization. This end-to-end trainable model introduces a dual-branched encoding feature fusion module, effectively combining components from both Vision Transformer and deep convolutional neural networks, leading to improved binarization quality and reduced computational complexity. Rahali et al. [28] developed Transformer-based (TB) models in natural language processing (NLP), emphasizing their expressiveness through self-attention mechanisms. The paper categorizes TB models, compares their architectures, and discusses limitations, offering insights to boost innovation in NLP applications and AI-powered products. Pilicita et al. [29] investigate the utility of five BERT-based pre-trained models in classifying mobile educational applications. Leveraging a dataset enriched with descriptions and categories from the Google Play Store, the study demonstrates the effectiveness of these models, achieving notable accuracy rates ranging from 76% to 81%.

There are various models designed for tasks related to documents, demonstrating how Transformer architectures can be used effectively in various areas. Nasi et al. [30] present PharmKE, a knowledge extraction platform for pharmaceutical texts. Leveraging transfer learning, the platform achieves a 96% F1-score in named entity recognition tasks, outperforming fine-tuned BERT and BioBERT models. The open-source modular architecture promotes reproducibility, accessibility, and potential integration into mobile systems, empowering patients with relevant medication information. In the study by Meshrif et al. [31], the authors address the challenge of classifying Arabic tweets by developing ARABERT4TWC, a BERT-based text classification model. The paper highlights the model's effectiveness in achieving high classification results across various datasets, outperforming other deep learning and conventional techniques. The proposed model showcases the potential of transfer learning with BERT for automating the classification of Arabic tweets. Tang et al.'s [32] work introduces Universal Document Processing (UDOP), a groundbreaking document AI model that unifies text, image, and layout modalities, achieving high-quality document editing and content customization. Leveraging a vsion–text–layout Transformer, UDOP sets the state of the art in eight document AI tasks, showcasing its versatility and dominance in the field.

Table 1 highlights several significant contributions that have been made in the state of the art across diverse domains. The various contributions include Arabic text classification, where machine learning models, particularly CNNs with character-level models, exhibit superior performance [20]. Nonetheless, this research could benefit from a more comprehensive dataset size and scope discussion. In the realm of tech document classification, the TechDoc architecture has proven its efficiency, enhancing categorization and scalability, particularly for tech companies [21]. However, its domain-specific focus limits broader applicability. Similarly, digitizing handwritten Devanagari text with CNN-based DHTR models preserves ancient knowledge but is confined to the Devanagari script [22]. Indian number plate detection using CNN-based models shows promise in enhancing road safety and theft prevention but necessitates scalability and diverse datasets [23]. Resume classifica-

tion with NLP and ML algorithms improves accuracy and automation but requires further exploration of scalability and broader applications [24]. Patent document classification, enhanced by transfer learning, shows improved performance [25], yet it remains confined to patent documents and classification tasks. Lastly, digitizing handwritten Devanagari script effectively preserves it, but broader applications warrant further study [33]. Therefore, a knowledge gap exists throughout the literature for creating an optimized accurate vision and Transformer combined model for classifying scanned copies of various documents into different categories.

Table 1. A summary of the current state of the art in document classification, including an overview of recent relevant studies and their associated limitations.

Source	Objective	Data Type	Algorithm	Remarks	Limitations
[20]	Arabic text classification using ML	Text data	MNB, BNB, SGD, LR, SVC, CNN	CNN with character-level model outperforms; applicability in various domains, particularly social media	Limited dataset size taken for classification.
[21]	Efficient tech document classification	Technical text	TechDoc architecture.	Improved tech document categorization and scalability for large tech companies.	Specific to tech documents; limited domain applicability
[22]	Digitization of handwritten Devanagari text	Text and image	CNN-based DHTR	Automated Devanagari script digitization; preservation of ancient knowledge	Specific to Devanagari script
[23]	Indian scenario number plate detection using TensorFlow	Image data	CNN-based model	Significant potential for road safety and theft prevention; real-world applications	Limited scalability, dataset size, and diversity
[24]	Resume classification using NLP and ML	Text data	Various ML algorithms, NLP	Efficient automation of resume categorization; improved accuracy and reliability	Focuses on resumes; limited scalability
[25]	Patent document classification using transfer learning	Text data	BERT, XLNet, RoBERTa, ELECTRA	Enhanced patent classification; improved state-of-the-art performance.	Limited to patent documents; specific to classification task
[26]	Multi-label emotion classification in texts	Text data	LSTMs, Transformers	Improved accuracy in multi-label emotion classification; outperforms existing benchmarks.	Focuses on social media text; limited to emotion classification
[33]	Digitization of handwritten Devanagari text	Text and image	CNN-based DHTR	Automated Devanagari script digitization; preservation of ancient knowledge	Specific to Devanagari scripts

3. Methodology

To carry out this experiment, we followed the process outlined in Figure 1. This process involved combining two main steps. In the first step, we classified the document

images using transfer learning for vision tasks. In the second step, we used OCR (text extraction) and Transformer models to classify the text. After that, we combined vison and OCR to create a strong network for classification.

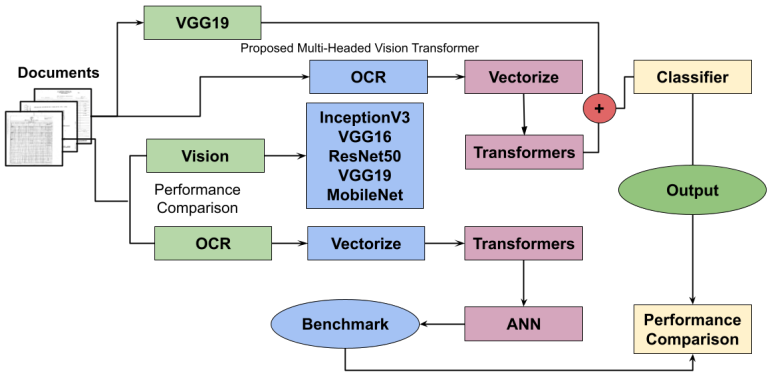

Figure 1. The framework workflow includes two key steps: document image classification using transfer learning, followed by OCR and Transformer-based text classification, ultimately integrating vision and OCR for robust classification.

3.1. Data Collection

The RVL-CDIP (Ryerson Vision Lab Complex Document Information Processing) dataset [34] is a substantial collection of grayscale images comprising 400,000 samples, meticulously organized into 16 distinct document categories. Each category letter, form, email, handwritten, advertisement, scientific report, scientific publication, specification, file folder, news article, budget, invoice, presentation, questionnaire, resume, and memo is numerically labeled from 0 to 15. The dataset has been thoughtfully partitioned into subsets, with 320,000 images allocated for training, 40,000 for validation, and an additional 40,000 for testing. Notably, the images have been resized to ensure that their largest dimension does not exceed 1000 pixels, ensuring uniformity in data representation. This dataset provides a robust foundation for developing and evaluating document classification and information processing models.

3.2. Optical Character Recognition

In the context of our proposed model, optical character recognition (OCR) is applied to all distinct documents. This preprocessing performs noise reduction, contrast enhancement, and resizing tasks to ensure optimal recognition accuracy. Following preprocessing, it uses the deep learning model, which identifies characters, words, the spatial position of the characters, and even complex layouts within the image [35]. Once the text is extracted from the image, postprocessing techniques are applied to enhance the accuracy of the recognized text, such as spell-checking and formatting correction. The final machine-readable text is generated, which is later converted to vectors.

3.3. Text Preprocessing

Text preprocessing is a vital step in natural language processing tasks, enhancing the quality and efficiency of text analysis. Techniques like stemming and lemmatization are initially applied to the text, standardizing words to their root forms and reducing complexity [36]. A vocabulary size of 10,000 words was used for text processing. Additionally, common stopwords are removed to focus on meaningful content. This list can contain 200 common words. The processed text is then tokenized, breaking it into individual words or tokens. An average token count per document of 300 was used. To prepare the text for machine learning, it is converted into numerical form using count vectorization, where each word is represented as a vector with its frequency [37]. The Adam optimizer was used to update the weights of this portion of the network.

3.4. Image Preprocessing

All distinct document images are initially standardized to a fixed size, such as 224 × 224 pixels, through resizing [38]. Following this, various image augmentation techniques are applied, including altering parameters like rotation, zoom, and brightness to create multiple versions of the same image, which helps reduce overfitting and improve model generalization [39]. Once the images have been standardized and augmented, they are converted into numerical vectors. For image classification, a typical neural network architecture includes an input layer with an image size of 224 × 224 pixels, convolutional layers with 32 to 256 filters, max-pooling layers, fully connected layers with 128 to 1024 nodes, dropout layers with a dropout rate of 0.2 to 0.5, ReLU activation functions, and 16 output layers with a softmax activation function for 16 distinct class probabilities [40]. The RMSprop optimizer was used to update network weights during training.

3.5. Proposed Multi-Headed Vision–Transformer Network

The proposed VisFormers model is created by combining the vision and Transformer networks. The network is summarized in Figure 2. The preprocessed vectorized language data are used as the input to the Transformer model with up to 200 features per sample. The output of this network is a 16-node vector, with each node corresponding to a particular category of document. On the other hand, the pre-trained VGG-19 model has been used for the vision network with ImageNet weights. This ensured that the spatial and visual features of the documents were also utilized while training the model. The VGG-19 was intended to create a reduced feature map that can complement the Transformer head's decision making. The parameters of this network were not updated during the training to retain the transfer learning properties. Images of 224 × 224 pixels were used as the input to the network, which was then flattened and passed through a dense layer of 64 nodes [38]. The target of the vision network was a reshaped and resized version of the same input. The vectors obtained from both the vision and the Transformer's output layers were then concatenated and passed through a series of layers until the final output layer of 16 nodes was reached. Adam optimizer with categorical cross-entropy loss was used to train both the tails of the network and the Transformer part of the network. The tail of the vision part of the network was optimized with a root mean squared propagation (RMSprop) optimizer using mean squared error loss [41]. The proposed network, therefore, has a total of three objective functions. The first one is the Transformer head, which is used to process the textual information; the second one is the VGG-19 head, which is used to process the visual information; and the third one is the tail of the network, which combines the information received from the Transformer and VGG-19 heads.

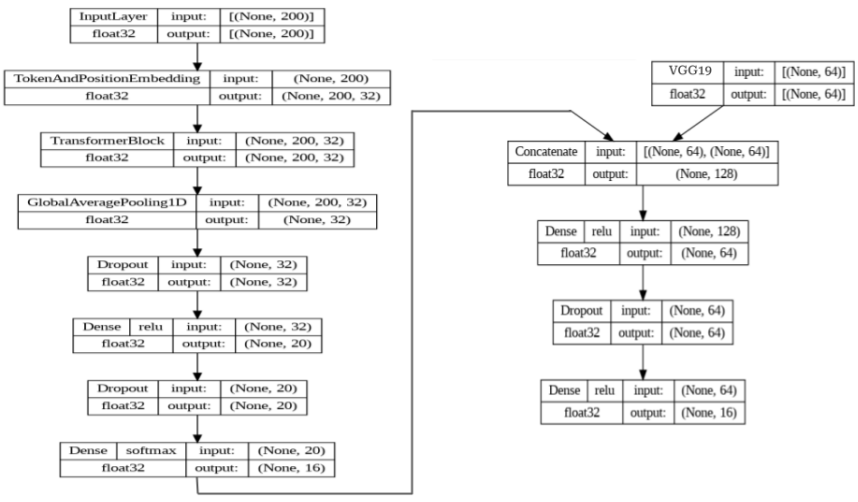

Figure 2. The VisFormers model combines a Transformer and a pre-trained VGG-19 network for document classification with specific architecture details.

Let the text vector be given by the tensor

$$t = \sum_{i=1}^{200} a_i \vec{x}_i \tag{1}$$

Here, a_i indicates the intensity of the vector, and $\vec{x}_i$ represents the ith component. This, according to Einstein's summation notation, can be given by

$$t = a_i \vec{x}_i \tag{2}$$

This is fed into the input layer of a neural network n_1 given by

$$n_1 = \sum_{i=1}^{200} \left(\sigma_1 (w_i a_i \vec{x}_i) \right) \tag{3}$$

Here, σ_1 indicates the activation function rectified linear unit, and w represents the weight matrix. Following this, a Transformer block is introduced. Let τ_1 be the target variable for the text network, and V represent the values in the word vector.

$$A(a_i \vec{x}_i, \tau_1, V) = \sigma_2 \left(\frac{a_i \vec{x}_i \tau_1^T}{\sqrt{d_k}} \right) V \tag{4}$$

From this, global average pooling was used to extract the necessary information. This is given by

$$F_k = \sum_{x,y} f_k(x, y) \tag{5}$$

where $f_k(x, y)$ denotes the feature map, and k represents the depth of the map. Let n_1 be the neural network updated according to these steps and d represent a dropout layer. Therefore, the neural network is updated as

$$n_1 = \sum_{k=1}^{16} \left(d_{0.5} \sum_{j=1}^{20} d_{0.2} \left(\sum_{i=1}^{200} \left(\sigma_1 (w_i a_i \vec{x}_i) \right) \right)_j \right)_k \tag{6}$$

This part of the neural network is optimized by Adam optimizer.

On the other hand, the image vector can be represented by the tensor

$$i = \sum_{j=1}^{m} \sum_{i=1}^{3} a_{i,j} \vec{x_{i,j}} \tag{7}$$

where a represents the intensity of the pixel ranging from 0 to 255, $\vec{x_m}$ represents the position of the pixel, and i indicates the three pixel planes, namely red, green, and blue.

Let n_2 represent the structure of the VGG-19 network [42]. This part of the network is optimized with RMSprop. Therefore, the concatenation is given by

$$n = n_1 \otimes n_2 \tag{8}$$

This is followed by a dense and dropout layer. This completes our objective function, which is given by

$$n \longleftarrow \sum_{j=1}^{15} \left(d_{0.2} \sum_{i=1}^{64} n_i \right) \tag{9}$$

This part of the network is updated using RMSprop using the mean squared error loss function.

3.6. Comparitive Analysis with Transfer Learning and Natural Language Processing

To compare performance, we used state-of-the-art methods like transfer learning and NLP [43]. Document images are initially resized to a fixed 224 × 224 pixel size and are then converted to numerical vectors. Then, we employ a diverse set of popular transfer learning networks for our document image classification task, including InceptionV3 (IncV3), VGG16, ResNet50, VGG19, and MobileNet. For the NLP, we extract text from images using optical character recognition. Then, text preprocessing is performed, which involves stemming, lemmatization, and removing stop words [44]. The processed text is tokenized and converted into vectors using count vectorization. The vectors are then passed to a Transformer for classification. Then, a comparative study is obtained using various classification metrics between transfer learning and NLP. Subsequently, we integrate both transfer learning and NLP to create a robust hybrid network for classification, harnessing the strengths of both approaches to enhance our overall classification capabilities.

4. Results

We used heatmap visualization and document classification performance for our Vision Transformer model. The results of our experiments are shown in the tables below and are further discussed in this section.

4.1. Heatmap Visualization for the Proposed Vision Transformer

The Grad-CAM (gradient-weighted class activation mapping) heatmap visualization, as shown in Figure 3 for our suggested Vision Transformer applied to 16 different kinds of document images, offers invaluable insights into the model's decision-making process. Grad-CAM aids in both model interpretability and identification of the salient features for precise categorization by emphasizing the regions of interest in these images, enabling us to grasp which parts are crucial for classification. The striking red regions in the heatmap indicate the focal points that the transfer learning models rely on for classification. These regions encapsulate the most discriminative features crucial for distinguishing the different document categories. Essentially, they serve as the model's attention zones, highlighting text patterns, shapes, or other visual cues pivotal for accurate classification. The areas the algorithm considers less significant for its categorization choice are represented by the parts of other colors, which are unremarkable in the heatmap. These areas could include background noise, pointless text, or less instructive visual components. The model maximizes its emphasis on the essential components of the image by ignoring these less important regions during classification, improving its capacity to make accurate document category predictions. By distinguishing between crucial and less relevant regions, Grad-CAM empowers the model to make informed decisions, ultimately improving its classification performance.

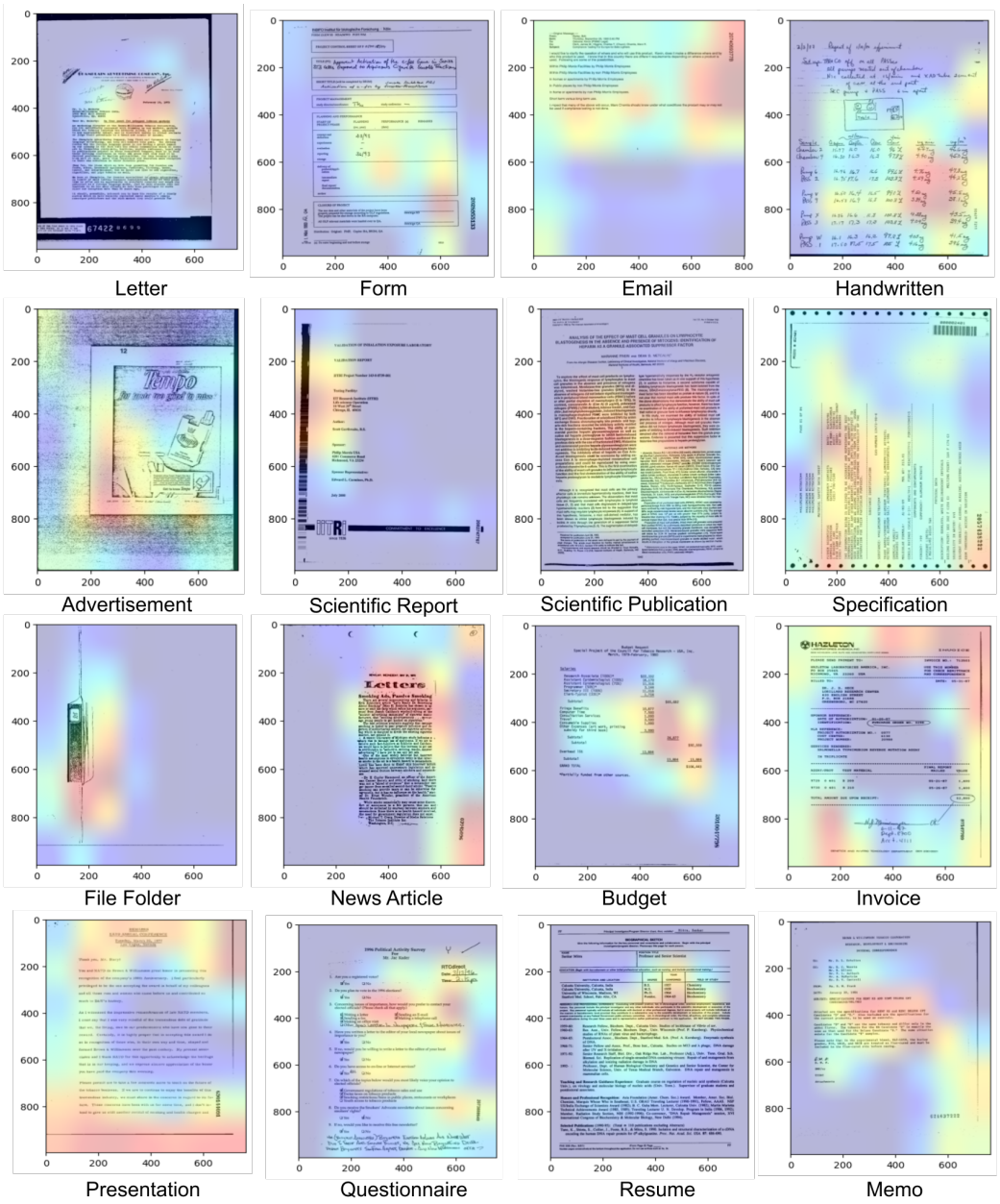

Figure 3. The Grad-CAM heatmap visualization highlights critical regions in document image classification. The reddish color indicates a higher probability density of having decision-making features.

4.2. Document Classification Performance for the Proposed Vision Transformer

The proposed Vision Transformer, a hybrid model combining NLP Transformer and transfer learning CNN, performs strongly in document classification tasks, as demonstrated in Table 2. The model shows an outstanding accuracy of 94.2%

Regarding computational efficiency, the model requires 951 s (approximately 15.85 min) for training, which, while being time consuming, is acceptable given the strong performance. With a test duration of 287 ms, the testing phase is significantly quicker, making it ideal for real-time or nearly real-time document classification applications. Compared to individual models, the Vision Transformer exhibits superior performance, surpassing both the standalone NLP Transformer and transfer learning models regarding accuracy and other classification metrics.

Table 2. Performance overview for the proposed Vision Transformer model, showcasing key metrics and execution times for evaluation.

Metrics	Proposed VisFormers Performance
Accuracy	**0.942**
Precision	0.913
Recall	0.935
F1 Score	0.924
Train Time (s)	951
Test Time (ms)	287

5. Discussion

In this section, we delve into the document classification performance for Transformers with optical character recognition and natural language processing and the document classification performance for vision-based transfer Learning. Additionally, we provide a comparative analysis with the state-of-the-art methods, offering insights into the effectiveness of our approach.

5.1. Document Classification Performance for Transformers with Optical Character Recognition and Natural Language Processing

The performance of Transformers in document classification, aided by optical character recognition (OCR) and natural language processing (NLP), is quite promising, as demonstrated in Table 3. The accuracy of this model is 0.83, meaning that it accurately categorizes 83% of the dataset's documents. With a precision score of 0.87, the model is highly accurate 87% of the time when classifying documents, a measure of the model's ability to prevent false positives. The recall score, in this case, is likewise vital at 0.85. The F1 score, which balances precision and recall, is 0.85, demonstrating the model's sturdiness. In terms of efficiency, the training time is 58 s, which suggests that the model is relatively quick to train. This can be a crucial consideration in real-world applications, especially when dealing with large datasets or the need for frequent model updates. The test time is also quite reasonable at 64 ms, implying that the model can make rapid predictions once trained. These results underscore the effectiveness of combining OCR and NLP with Transformers for document classification tasks. With high accuracy, precision, and recall, this model demonstrates its potential in accurately categorizing documents.

Table 3. Performance metrics for the state-of-the-art OCR and NLP-enhanced document classification model.

Metrics	SOTA OCR + NLP Performance	Proposed VisFormers
Accuracy	0.83	**0.942**
Precision	0.87	**0.913**
Recall	0.85	**0.935**
F1 Score	0.85	**0.924**
Train Time (s)	58	**951**
Test Time (ms)	64	**287**

5.2. Document Classification Performance for Vision-Based Transfer Learning

Table 4 presents a comprehensive analysis of the document classification performance achieved through vision-based transfer learning using five different pre-trained convolutional neural network (CNN) models: InceptionV3 (IncV3), VGG16, ResNet50, VGG19, and MobileNet. With accuracy scores of 0.89 and 0.90, VGG16 and VGG19 are in first place. This suggests that 89% and 90% of the time, respectively, these algorithms classify documents into the correct categories. Because of their great accuracy, they are likely to be excellent at differentiating across document classes and are a good fit for applications where accuracy is crucial. The model's capacity for avoiding false positives is measured by precision. With

a precision score of 0.91, VGG19 outperforms the competition in this area since it accurately predicts a document's class 91% of the time. This characteristic is essential in situations where incorrect classifications can have serious repercussions.

Table 4. Document classification performance using different state-of-the-art pre-trained CNN transfer learning models.

Metrics	IncV3	VGG16	Res Net50	VGG19	Mobile Net	VisFormers
Accuracy	0.79	0.89	0.87	0.9	0.85	**0.94**
Precision	0.81	0.85	0.88	0.91	0.88	**0.91**
Recall	0.79	0.83	0.86	0.87	0.86	**0.94**
F1 Score	0.79	0.83	0.86	0.88	0.86	**0.92**
Train Time (s)	785	802	792	855	779	951
Test Time (ms)	124	231	144	254	112	287

Here, VGG16 exhibits the highest recall at 0.83, indicating its proficiency in capturing true positives while minimizing false negatives. A high recall is vital when missing relevant documents can be detrimental. F1 score, which balances precision and recall, also showcases VGG19's superiority with a score of 0.88, highlighting its well-rounded performance. Beyond classification accuracy and precision, the practicality of these models depends on their training and testing times. In this regard, MobileNet stands out as the most efficient model, with a training time of 779 s and a rapid testing time of 112 ms. This makes MobileNet particularly suitable for real-time applications and scenarios with limited computational resources. VGG16, although accurate, requires more training time (802 s) and testing time (231 ms). The other models fall in between in terms of computational demands. In summary, VGG16 and VGG19 offer high accuracy and precision, making them ideal for applications demanding precise categorization.

5.3. Comparison with the State of the Art

There are several advantages of our proposed VisFormers model relative to the state-of-the-art approaches. The work conducted by [45,46] achieves an accuracy of less than 90% and requires over 900–1200 s for training using primary image classification neural networks. Nevertheless, our proposed model is designed by combining the transfer learning image classification model and Transformers and obtained an accuracy of over 96%. In the paper [47], the author has only used OCR and classified them using Transformers, achieving an accuracy of less than 80%, which is very low compared to our proposed model. The work conducted by the authors of [48,49] used both vision as well as NLP for classification, but the accuracy they achieved was less than 90%, and training time was also very high. In contrast, our model obtained an accuracy of over 96% and a training time of 951 s, surpassing all other models. In the paper [29], the author used BERT-based pre-trained models for classifying documents. Their model achieved an accuracy of 81%, which is very low compared to our proposed vision Transformer. The D2BFormer model developed by the author in their paper [27,28] excels in degraded document binarization. One limitation is the potential sensitivity to variations in document types. Our work surpasses this limitation, achieving superior performance across diverse document datasets. Universal Document Processing, developed by the author in their paper [30,32], excels in document-related tasks; one limitation is its potential computational intensity during training. However, our model surpasses this challenge, achieving superior performance in terms of computational efficiency and overall accuracy. In the paper [31], the effectiveness of the ARABERT4TWC model in classifying Arabic tweets is showcased. One limitation lies in the lack of exploration into domain and cross-domain pre-training for BERT. Our work surpasses this limitation by demonstrating superior performance in tokenized vectors and learning rate fine-tuning across three datasets. The provided Table 5 shows a comprehensive summary of the performance comparisons. In summary, the results indicate that

our model outperforms existing approaches in accuracy and efficiency, positioning it as a valuable contribution to the state-of-the-art visual document classification domain.

Table 5. Performance of the proposed Vision Transformer model compared to other state-of-the-art models.

Source	Vision	OCR	Accuracy	Train Time (s)
Tensmeyer et al. [45]	✓	X	<90%	>900
Siddiqui et al. [46]	✓	X	<85%	>1200
Larson et al. [47]	X	✓	<80%	>500
Kanchi et al. [48]	✓	✓	<90%	>1500
Bakkali et al. [49]	✓	✓	<92%	>1000
Proposed VisFormers	✓	✓	>**94%**	<=**951**

5.4. Limitations of the Proposed VisFormers Model

Despite its success in achieving a high accuracy of 94.2% on the RVL-CDIP dataset, the proposed VisFormers model has certain limitations. The model's performance may be sensitive to variations in document characteristics not well-represented in the training set, potentially leading to misclassifications in real-world scenarios. Additionally, the current implementation focuses on a fixed set of 16 document classes, limiting its adaptability to diverse or evolving document categories. Furthermore, the model's effectiveness may vary when applied to documents with highly unconventional layouts or visual elements not encountered during training.

6. Conclusions

Complex document classification from scanned images is challenging but essential for rapid digitization. Given the prevalence of printed complex documents in various industries, the need for a specialized document classification model is apparent. Our proposed model addresses this gap by leveraging optical character recognition (OCR) to extract text and employing a multi-headed VisFormers network for simultaneous vision-based transfer learning and natural-language-based Transformers. By combining these approaches, our model accurately classifies complex documents, filling the void left by traditional ERP and CRM systems. Integrating our model into existing software solutions enhances their capability to manage and process diverse document types effectively, ultimately improving organizational workflow efficiency and information management. In the proposed model, the vision network receives image input of the data, and the Transformer network receives vectorized OCR data of the same image. Therefore, different parts of the network receive different inputs; they are trained differently using different loss functions and different optimizers. The proposed VisFormers network is tested against a standard complex document classification dataset called RVL-CDIP, and it outperforms the state-of-the-art works by achieving an accuracy score of 94.2%.

The model is tested for 16 document classes, which can be increased in future works. The model can be deployed at several workplaces to digitize paper documents.

Author Contributions: Conceptualization, S.D. and S.A.; methodology, S.D. and S.A.; validation: S.D., S.A. and A.D.D.; formal analysis: S.D. and S.A.; investigation: S.D., S.A. and A.D.D.; resources: S.D., S.A. and A.D.D.; data curation: S.A.; writing—original draft preparation: S.D. and S.A.; writing—review and editing: S.D., S.A. and A.D.D.; visualization: S.D.; supervision: S.A. and A.D.D.; project administration: S.A.; funding acquisition: S.A. and A.D.D. All authors have read and agreed to the published version of the manuscript.

Funding: The research was funded by Spiraldevs Automation Industries Pvt. Ltd. with grant number: SDAI/RND/GRANT/202312091.

Data Availability Statement: Data for the experiment can be made available from the authors upon request.

Acknowledgments: The work is supported by Aerosys Defence and Aerospace Pvt. Ltd., GS Enterprise, Spiraldevs Automation Industries Pvt. Ltd., Wingbotics LLP, and the Gyanam Multi-Disciplinary Research Foundation.

Conflicts of Interest: There are no conflicts of interest regarding the paper to declare.

Code Availability: The code for the work is made available at https://doi.org/10.5281/zenodo.10641331.

Ethical Statement: The work adheres to all ethical guidelines provided by the journal.

References

1. Audebert, N.; Herold, C.; Slimani, K.; Vidal, C. Multimodal deep networks for text and image-based document classification. In Proceedings of the Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, 16–20 September 2019; pp. 427–443.
2. Adhikari, A.; Ram, A.; Tang, R.; Lin, J. Docbert: Bert for document classification. *arXiv* **2019**, arXiv:1904.08398.
3. Kim, D.; Seo, D.; Cho, S.; Kang, P. Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. *Inf. Sci.* **2019**, *477*, 15–29. [CrossRef]
4. Bhagat, R.; Thosani, P.; Shah, N.; Shankarmani, R. Complex Document Classification and Integration with Indexing. In Proceedings of the 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 4–6 August 2021; pp. 1477–1484. [CrossRef]
5. Biten, A.F.; Tito, R.; Gomez, L.; Valveny, E.; Karatzas, D. Ocr-idl: Ocr annotations for industry document library dataset. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23 October 2022; pp. 241–252.
6. Aydin, Ö. Classification of documents extracted from images with optical character recognition methods. *Comput. Sci.* **2021**, *6*, 46–55.
7. Jiang, M.; Hu, Y.; Worthey, G.; Dubnicek, R.C.; Underwood, T.; Downie, J.S. Impact of OCR quality on BERT embeddings in the domain classification of book excerpts. *Ceur Proc.* **2021**, *1613*, 0073.
8. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A comprehensive survey on transfer learning. *Proc. IEEE* **2020**, *109*, 43–76. [CrossRef]
9. Banerjee, S.; Akkaya, C.; Perez-Sorrosal, F.; Tsioutsiouliklis, K. Hierarchical transfer learning for multi-label text classification. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 6295–6300.
10. Behera, B.; Kumaravelan, G.; Kumar, P. Performance evaluation of deep learning algorithms in biomedical document classification. In Proceedings of the 2019 11th International Conference on Advanced Computing (ICoAC), Hawaii, HI, USA, 18 March 2019; pp. 220–224.
11. Zhao, Z.; Yang, S.; Zhao, D. A new framework for visual classification of multi-channel malware based on transfer learning. *Appl. Sci.* **2023**, *13*, 2484. [CrossRef]
12. Baniata, L.H.; Kang, S. Transformer Text Classification Model for Arabic Dialects That Utilizes Inductive Transfer. *Mathematics* **2023**, *11*, 4960. [CrossRef]
13. Singh, R.; Gildhiyal, P. An Innovation Development of Document Management and Security Model for Commercial Database Handling Systems. In Proceedings of the 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS), Raichur, India, 24–25 February 2023; pp. 1–6. [CrossRef]
14. Finances Online. 71 Cloud File & Document Management Statistics You Must Know: 2021 Data Analysis & Market Share. 2024. Available online: https://financesonline.com/cloud-file-document-management-statistics (accessed on 10 February 2024).
15. Pandey, M.; Arora, M.; Arora, S.; Goyal, C.; Gera, V.K.; Yadav, H. AI-based Integrated Approach for the Development of Intelligent Document Management System (IDMS). *Procedia Comput. Sci.* **2023**, *230*, 725–736. [CrossRef]
16. Dutta, S.; Goswami, S.; Debnath, S.; Adhikary, S.; Majumder, A. If Human Can Learn from Few Samples, Why Can't AI? An Attempt On Similar Object Recognition with Few Training Data Using Meta-Learning. In Proceedings of the 2023 IEEE North Karnataka Subsection Flagship International Conference (NKCon), Belagavi, India, 19–20 November 2023; pp. 1–6. [CrossRef]
17. Sajadfar, N.; Abdollahnejad, S.; Hermann, U.; Mohamed, Y. Text detection and classification of construction documents. In Proceedings of the ISARC, International Symposium on Automation and Robotics in Construction, Banff, AL, Canada, 24–24 May 2019; Volume 36, pp. 446–452.
18. Guha, A.; Samanta, D. Real-time application of document classification based on machine learning. In Proceedings of the Intelligent Computing Paradigm and Cutting-Edge Technologies (ICICCT 2019), Istanbul, Turkey, 30–31 October 2019; pp. 366–379.
19. Adhikary, S.; Dutta, S.; Dwivedi, A.D. Secret learning for lung cancer diagnosis—A study with homomorphic encryption, texture analysis and deep learning. *Biomed. Phys. Eng. Express* **2023**, *10*, 015011. [CrossRef]
20. Muaad, A.Y.; Kumar, G.H.; Hanumanthappa, J.; Benifa, J.B.; Mourya, M.N.; Chola, C.; Pramodha, M.; Bhairava, R. An effective approach for Arabic document classification using machine learning. *Glob. Transit. Proc.* **2022**, *3*, 267–271. [CrossRef]
21. Jiang, S.; Hu, J.; Magee, C.L.; Luo, J. Deep learning for technical document classification. *IEEE Trans. Eng. Manag.* **2022**, *71*, 1163–1179. [CrossRef]

22. Dhanikonda, S.R.; Sowjanya, P.; Ramanaiah, M.L.; Joshi, R.; Krishna Mohan, B.; Dhabliya, D.; Raja, N.K. An efficient deep learning model with interrelated tagging prototype with segmentation for telugu optical character recognition. *Sci. Program.* **2022**, *2022*, 1059004. [CrossRef]
23. Tote, A.S.; Pardeshi, S.S.; Patange, A.D. Automatic number plate detection using TensorFlow in Indian scenario: An optical character recognition approach. *Mater. Today Proc.* **2023**, *72*, 1073–1078. [CrossRef]
24. Ali, I.; Mughal, N.; Khand, Z.H.; Ahmed, J.; Mujtaba, G. Resume classification system using natural language processing and machine learning techniques. *Mehran Univ. Res. J. Eng. Technol.* **2022**, *41*, 65–79. [CrossRef]
25. Haghighian Roudsari, A.; Afshar, J.; Lee, W.; Lee, S. PatentNet: Multi-label classification of patent documents using deep learning based language understanding. *Scientometrics* **2022**, *127*, 207–231. [CrossRef]
26. Ameer, I.; Bölücü, N.; Siddiqui, M.H.F.; Can, B.; Sidorov, G.; Gelbukh, A. Multi-label emotion classification in texts using transfer learning. *Expert Syst. Appl.* **2023**, *213*, 118534. [CrossRef]
27. Yang, M.; Xu, S. A novel Degraded Document Binarization model through vision transformer network. *Inf. Fusion* **2023**, *93*, 159–173. [CrossRef]
28. Rahali, A.; Akhloufi, M.A. End-to-end transformer-based models in textual-based NLP. *AI* **2023**, *4*, 54–110. [CrossRef]
29. Pilicita, A.; Barra, E. Using of Transformers Models for Text Classification to Mobile Educational Applications. *IEEE Lat. Am. Trans.* **2023**, *21*, 730–736. [CrossRef]
30. Jofche, N.; Mishev, K.; Stojanov, R.; Jovanovik, M.; Zdravevski, E.; Trajanov, D. Pharmke: Knowledge extraction platform for pharmaceutical texts using transfer learning. *Computers* **2023**, *12*, 17. [CrossRef]
31. Alruily, M.; Manaf Fazal, A.; Mostafa, A.M.; Ezz, M. Automated Arabic long-tweet classification using transfer learning with BERT. *Appl. Sci.* **2023**, *13*, 3482. [CrossRef]
32. Tang, Z.; Yang, Z.; Wang, G.; Fang, Y.; Liu, Y.; Zhu, C.; Zeng, M.; Zhang, C.; Bansal, M. Unifying vision, text, and layout for universal document processing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19254–19264.
33. Pande, S.D.; Jadhav, P.P.; Joshi, R.; Sawant, A.D.; Muddebihalkar, V.; Rathod, S.; Gurav, M.N.; Das, S. Digitization of handwritten Devanagari text using CNN transfer learning–A better customer service support. *Neurosci. Inform.* **2022**, *2*, 100016. [CrossRef]
34. Harley, A.W.; Ufkes, A.; Derpanis, K.G. Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015.
35. Jayoma, J.M.; Moyon, E.S.; Morales, E.M.O. OCR based document archiving and indexing using PyTesseract: A record management system for dswd caraga, Philippines. In Proceedings of the 2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), Manila, Philippines, 3–7 December 2020; pp. 1–6.
36. Camastra, F.; Razi, G. Italian text categorization with lemmatization and support vector machines. In *Neural Approaches to Dynamics of Signal Exchanges*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 47–54.
37. Wendland, A.; Zenere, M.; Niemann, J. Introduction to text classification: Impact of stemming and comparing TF-IDF and count vectorization as feature extraction technique. In Proceedings of the Systems, Software and Services Process Improvement: 28th European Conference, EuroSPI 2021, Krems, Austria, 1–3 September 2021; pp. 289–300.
38. Adhikary, S. Fish Species Identification on Low Resolution-A Study with Enhanced Super Resolution Generative Adversarial Network (ESRGAN), YOLO and VGG-16. *Res. Sq.* **2022**. [CrossRef]
39. Groleau, A.; Chee, K.W.; Larson, S.; Maini, S.; Boarman, J. Augraphy: A data augmentation library for document images. *arXiv* **2022**, arXiv:2208.14558.
40. Rhanoui, M.; Mikram, M.; Yousfi, S.; Barzali, S. A CNN-BiLSTM model for document-level sentiment analysis. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 832–847. [CrossRef]
41. Dutta, S.; Adhikary, S. Evolutionary Swarming Particles To Speedup Neural Network Parametric Weights Updates. In Proceedings of the 2023 9th International Conference on Smart Computing and Communications (ICSCC), Kochi, India, 17–19 August 2023; pp. 413–418. [CrossRef]
42. Dey, N.; Zhang, Y.D.; Rajinikanth, V.; Pugalenthi, R.; Raja, N.S.M. Customized VGG19 Architecture for Pneumonia Detection in Chest X-Rays. *Pattern Recognit. Lett.* **2021**, *143*, 67–74. [CrossRef]
43. Liu, R.; Shi, Y.; Ji, C.; Jia, M. A survey of sentiment analysis based on transfer learning. *IEEE Access* **2019**, *7*, 85401–85412. [CrossRef]
44. Pappagari, R.; Zelasko, P.; Villalba, J.; Carmiel, Y.; Dehak, N. Hierarchical transformers for long document classification. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 838–844.
45. Tensmeyer, C.; Martinez, T. Analysis of convolutional neural networks for document image classification. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; Volume 1, pp. 388–393.
46. Siddiqui, S.A.; Dengel, A.; Ahmed, S. Self-supervised representation learning for document image classification. *IEEE Access* **2021**, *9*, 164358–164367. [CrossRef]
47. Larson, S.; Lim, Y.Y.G.; Ai, Y.; Kuang, D.; Leach, K. Evaluating Out-of-Distribution Performance on Document Image Classifiers. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 11673–11685.

48. Kanchi, S.; Pagani, A.; Mokayed, H.; Liwicki, M.; Stricker, D.; Afzal, M.Z. EmmDocClassifier: Efficient multimodal document image classifier for scarce data. *Appl. Sci.* **2022**, *12*, 1457. [CrossRef]
49. Bakkali, S.; Ming, Z.; Coustaty, M.; Rusiñol, M. Visual and textual deep feature fusion for document image classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2020; pp. 562–563.