International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 11, No. 10, October, 2024, pp. 2809-2816
Journal homepage: http://iieta.org/journals/mmep

Enhancing Document Image Classification with Discrete Wavelet Transform and Swin Transformer

Ahmed Hussein Salman[1*], Waleed Al-Jawher[2]

[1] Iraqi Commission for Computers & Informatics, Informatics Institute of Postgraduate Studies, Baghdad 10001, Iraq
[2] Department of Electronic and Communication, Uruk University, Baghdad 10001, Iraq

Corresponding Author Email: phd202110677@iips.edu.iq

https://doi.org/10.18280/mmep.111023

Received: 5 April 2024
Revised: 26 May 2024
Accepted: 23 July 2024
Available online: 31 October 2024

Keywords:
classification, Discrete Wavelet Transform (DWT), image document, Swin Transformer

ABSTRACT

Document image classification has experienced significant advancements in recent years, via the development of transformer-based models. This research contributes to the present literature on techniques for classifying document images and emphasizes the significance of combining several methods to improve the accuracy of classification. The proposed method combines DWT and Swin Transformer architectures. By using both hierarchical and multi-scale features, DWT improves the Swin Transformer's ability to tell the difference between complex patterns in images. This is achieved by combining the advantages of the two methods. In addition, the main objective is to increase the effectiveness of the classification model by relying on the features extracted from the wavelet transform and the strength of the swing as a classifier. The proposed method was applied to the dataset "Tobacco3482-jpg." The images in (the spatial domain) are fed into a wavelet transform, which decomposes into four bands. We rely exclusively on the approximate information band because it includes the most important features without considering other bands. Therefore, datasets will be modified to include approximate information about images, and this modified dataset in (frequency domain) fed to Swin Transformer, as a classifier. The proposed method's performance was successful, indicating its strong potential in image classification. Accuracy reached 90%.

1. INTRODUCTION

The digitization of paper documents, the use of digital devices, and the expansion of internet content are expanding document images dramatically. Document images include scanned copies of printed documents, searchable and non-searchable PDFs, and handwritten notes. Automatic classification of document pictures is needed for digital archiving, information retrieval, automated indexing, and content analysis. In addition, many classifications of documents can be text data or images, each of which has its own role, especially within systems that require information retrieval and management, where the classification aspect is important within this field where huge amounts of digital documents can be extracted. Systems do not work unless there are advanced techniques to classify them without the need to use the manual method [1-3]. Classifying document images by content and structure is document image classification. They are necessary to understand and operate with a large number of digital documents for searching, navigating, and automated processing processes. In the legal and medical sectors, document classification is crucial to keeping records organized and accessing essential information promptly. In libraries and archives, document image categorization can preserve and find priceless historical documents. The emergence of the era of technology has led to an important tool for creating effective methods for classifying images due to the rapid increase in digital documents, where it is possible to build and develop an effective model capable of classifying documents, including images, in terms of their content and visual features, and this represents the main goal for classifying images [4-6].

Many challenges can be involved in classifying image documents and this is related to differences in document layout and noise in image documents [7-9]. Images of documents frequently show all sorts of noise, from background patterns to stains, through to characteristic scanning or photocopying artifacts. This noise can saturate discriminative features and limit the suitability of traditional classification methods. Document images are very diverse in terms of layouts, containing text in different sizes, fonts, orientations, or arrangements in multiple columns, which can be interspersed with figures, tables, and so on. Such diversity is quite challenging for conventional models to capture and classify the features. The various types of content present in documents, such as text, graphics, and images, complicate feature extraction and classification. Common practices find it difficult to retain the balance between performing in-detail local feature extraction and at the same time understanding global context and structure. This current study addressed the limitations found in image classification methods in previous research through the document image classification has been considerably advanced in accuracy and efficiency. However, existing algorithms mainly sum up the text instances and do not consider the layout information. more these methods generally achieve lower performance for complex layouts and images with different resolutions. Bridging these two realms, the fusion of Swin Transformer and DWT offers an innovative design to overcome these bottlenecks due to both the Swin Transformer's hierarchical vision capacity and the multi-resolution analysis expertise of DWT. The reason for choosing DWT for this method is that it can decompose images into four different frequency sub-bands, the spatial relationships and edge detail needed to distinguish document elements such as text, shapes, tables, and texture capture, despite differences in document quality and noise strength within the image document.

The complex structure and pattern within any document can easily be identified when using the Swin converter. The reason is that it uses self-attention mechanisms and remote correlations in image correction by capturing important contextual information necessary to classify various types and features of documents, whether they are text or image documents. This work investigates the integrated reorganization of these 2 methods for improved robustness and accuracy in document image classification to address the lack of unified and flexible manipulation over complex and diverse document structures.

2. RELATED WORK

Recently, there have been several research studies that have dealt with the aspect of classifying image documents due to the increasing amount of digital documents in the recent period, and the reason is due to the development of the latest technologies that can accurately classify image documents, which are characterized by their diverse layouts, content hierarchies, and writing styles. In this part, we describe the most important previous research works through which we can lay the foundation for the proposed method based on DWT and Swin transforms for document image classification. Al-Anzi and AbuZeina [10] presented different methods for Arabic text classification, starting with extracting textual features based on singular value analysis and latent semantic indexing, including TF-IDF, and then using the cosine similarity method within this scope for classification. There are several methods, including KNN and SVM methods, in the classifier. Six problems related to cosine measurements. The accuracy and cosine similarity of the LSI classifiers were 82.5%. Liu et al. [11] proposed a hierarchical compiler based on the Swin compiler, which serves as a general-purpose backbone for computer vision. In addition, it addresses the challenges in adapting the converter from language to vision by relying on this converter with variable windows to increase efficiency. The method used modeled several parameters for the models, in addition to linear computational complexity concerning image size. Through this method based on the Swin Transformer, the model outperformed previous models in image classification, semantic segmentation tasks, and object detection. Multi-class text classification for Uzbek texts is proposed by Rabbimov and Kobilov [12]. A dataset was created from ten types of Uzbek "Daryo" internet news items. Various machine learning methods were utilized for multi-class text categorization, including SVM, DTC, RF, LR, and MNB. The text classification functional scheme and software development stages are described technologically. The TF-IDF technique and word- and character-level n-gram models extracted features. Text classification hyperparameters were defined using 5-fold cross-validation. The best experiment accuracy was 86.88%. The models and approaches in this study can be utilized to classify Uzbek literature and conduct future research. Text classification functionality and software were described by Tian et al. [13]. The study extracted features using word- and character-level n-gram models and TF-IDF. Text categorization hyperparameters were obtained by 5-fold cross-validation. An innovative picture demising method uses a multi-stage CNN with wavelet transform. MWDCNN uses a dynamic convolutional block, two wavelets transform and augmentation blocks, and a residual block. The MWDCNN algorithm balances decreasing performance and computing costs. It optimizes denoising by reducing noise and improving features. Testing shows that the MWDCNN outperforms popular denoising methods in quantitative and qualitative analysis. Dhar and Abedin [14] suggested a Tree-Based Pipeline Optimization Tool method and the vector control method along with TF-IDF with Logistic Regression, KNN, Support Vector Machine (SVM), Naïve Byes, and AdaBoost for classifying Bengali news headlines. This resulted in a very successful Bengali text classification system. (TPOT) using machine learning. In this investigation, the upgraded ML pipeline outperformed ordinary ML with an average accuracy of 81%. Guo et al. [15] presented a new method for image classification based on a multi-attention fusion network (MAFN). The MAFN method involves multiple spatial and zonal interest units to mitigate the effect of redundant bands and overlapping pixels. The benefit of the proposed methodology is the integration of information in addition to the reuse of features from several levels, which in turn generates more representative features. Liu et al. [16] presented an innovative network called Spectral SWIN for hyperspectral image (HSI) classification. The network is capable of efficiently converting HSI data into a representation of spatial and spectral features through the use of a novel oscillation spectral module (SSM). Here it serves as the backbone of the computer-designed converter for HSI. Several experiments using the proposed method were conducted on two distinct HSIs and its efficiency was proven. Mahmoud et al. [17] presented a method for classifying image documents based on the use of SVM and Gradient Boosting, with great care in pre-processing the image data, and the highest image optimization level is determined to obtain the best image accuracy.

To classify Bengali news headlines, Shahin et al. [18] used ANN, LSTM, and Bi-LSTM for this work, with the best accuracy rates reaching 70.94%, 82.20%, and 85.14% using these methods.

The main contribution of this paper can be detailed as follows:

Noise reduction vs. feature preservation: Utilizing the low-frequency component, which exhibits reduced noise levels while preserving major features of the image.

Enhanced feature extraction: Taking advantage of the approximation information band provided by DWT to extract significant patterns or features from the signal, significantly improving the performance of the classifier by focusing on essential information.

Uses self-attention processes in Swin Transform to examine relations and long-range dependencies between image patches, which makes it possible to accurately identify intricate patterns and layouts.

Efficiency and effectiveness: Reducing processing time and complexity of the model and increasing the effectiveness of the classification model by relying on the features extracted from Wavelet Transform and the strength of the Swin as a classifier.

The structure of this paper is as follows. The proposed methodology provided in Section 2 is illustrated. Section 3 provides an explanation of the results and a discussion of classification. Lastly, Section 4 will delve into the discussion of the conclusion and prospects.

3. PROPOSED METHOD

Document image classification method combines Discrete Wavelet Transform and Swin Transformer to exploit the best of both worlds, for A brief description of these methodologies and how these methods work and interact with each other shown in the main block diagram of the proposed structure is given in the Figure 1.

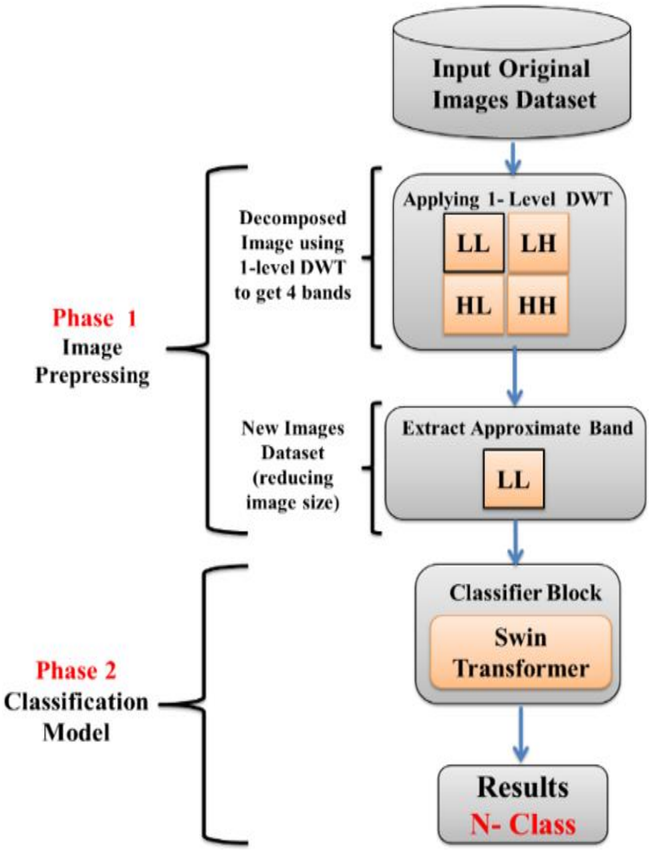


Figure 1. Flowchart of the proposed method

It consists of the following procedures Initially, the image is fed into Wavelet Transform as an image preprocessing by reducing the image size by decomposing the image into four bands (LL, LH, HL, HH): where three bands are ignored and LL is kept because it represents approximate information, The first stage of preprocessing is an application of the DWT to document images for features extraction. DWT represents an image at multiple frequency bands at different resolutions. This allows to analysis of both localized patterns and the frequency information of these localized patterns over the whole image and owing to DWT, document images are analyzed at multiple resolutions. Given that the document may have various sizes, DWT is used to adjust those sizes uniformly. and then the modified image is entered into the Swin transform as a classifier model. Swin Transformer is a hierarchical vision approach that employs local attention, non-overlapping window-based preclusion of the whole image, and shift to capture local context. Swin Transformer thus makes robust by employing a preclusion mechanism to effectively learn feature representations via local and shift-based mechanisms to maintain the global context of the image. Combining the high-level features from the Swin Transformer and low-level features from DWT improves both sensitivity and specificity when compared to each other. The L2-normalized output of the Swin Transformer is input to this class to project onto a trainable prototype classification space for exploiting both features. Each of these components will be explained in detail in the succeeding subsections.

3.1 The dataset

In this study, we utilize a publicly accessible dataset consisting of images extracted from scanned documents originating from USA Tobacco firms. The collection was provided by Legacy Tobacco Industry papers and was generated by the University of California San Francisco (UCSF). Tobacco3482 is a dataset consisting of 3482 document images and 10 categories of documents. Despite its limited size, it remains highly popular for the task of document image classification. These datasets accurately depict the challenges that businesses and organizations may encounter, taking into consideration the quality and type of the classes. The Tobacco-3482 dataset comprises document images categorized into 10 classes, including letter, form, email, resume, memo, and others. The dataset contains a total of 3482 images.

3.2 Phase one: Image preprocessing

In the process of image preprocessing, we use DWT transformer, which entails applying the DWT to the images to extract significant features or information that can be beneficial for classification purposes. There are two steps involved in the image preprocessing phase when working with the "Tobacco3482-jpg" image dataset.

3.2.1 Decomposed image using 1-level DWT

In this step, decomposed images are divided into four bands of frequency sub bands, which capture texture, spatial connections, and edge features that are crucial for differentiating document elements such as text, figures, and tables. By applying DWT because it has an important role in dividing the image into coefficients represented by approximation or low-frequency components and coefficients that provide detailed or high-frequency information, this can be achieved by implementing a series of filtering and sampling procedures taking into account.

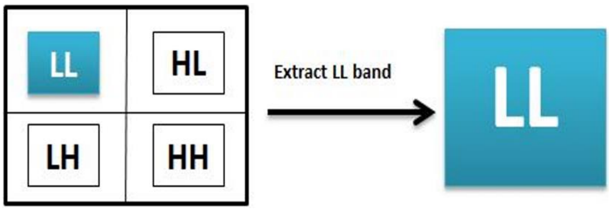


Figure 2. Sub-bands formed after 1-level DWT

The approximation coefficients correspond to the less detailed or lower frequency parts of the signal, whereas the detail coefficients capture the more intricate or higher frequency information [19-21]. DWT decomposed the image into 4 bands (LL, LH, HL, and HH) as shown in Figure 2.

3.2.2 Extract approximate band

The one-level decomposition of DWT using a filter. The LL band contains substantial information gathered from the original image. The LH, HL, and HH bands include the vertical, horizontal, and diagonal components of the original image. The original image can be reconstructed by exclusively considering the LL band image and disregarding any other irrelevant information from the other bands.

During this step of the suggested approach, the LL band is selected while the other bands are disregarded. This happens because the LL band holds the most important information about the features of the image. The LL band is essentially a smaller-sized representation of the original image. Consequently, the subsequent image processing will be performed on this reduced-sized image. The LL band is extracted from the initial images and used to create a new dataset of images. This dataset includes the LL information from the original images and is then used as input for the classification model as shown in Figure 3.

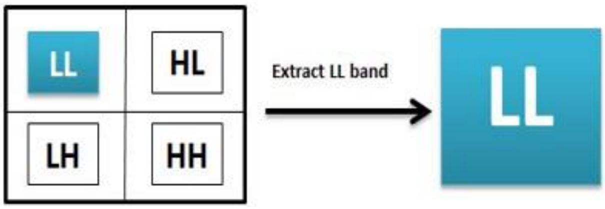


Figure 3. Extract approximate band

3.3 Phase two: Classification model

In this part, the model is built to classify image documents using the proposed method, the Swin Transformer, as a model for classifying images. To build any model, pre-processing of the image must be done, and here I used DWT for this task. DWT is a mathematical tool that transforms a signal into its constituent parts, making it easier to analyze and process. It works through decomposition, downsampling, and repetition to decompose the signal into approximation and detail coefficients. DWT is used in signal denoising, image compression, and feature extraction in machine learning for classification and recognition tasks. The benefit of it is to reduce the size of the image and extract important features included in the field of LL. For greater visibility in the computer world, a key role was played by the Swin Transformer, also referred to as the other window transformer, which uses hierarchical feature maps created by embedding image patches in deeper layers. This improvement has led to higher performance in many fine-grained definition tasks. Among them are object detection, semantic segmentation, image classification, and other things [22-24]. The process of dividing images in a non-overlapping manner into sections and restricting self-attention processes to each window. The Swin Transformer had a major role in this field. By applying the proposed method, there was a clear improvement in computational efficiency, as the computational cost scales linearly with the size of the input image. In addition, this work is suitable due to Swin Transformers' efficient representation, scalability, and feature extraction capabilities. They capture local and global features, making them ideal for document image classification and real-world applications with high-resolution data. Figure 4 shows the diagram of the Swin Transformer.

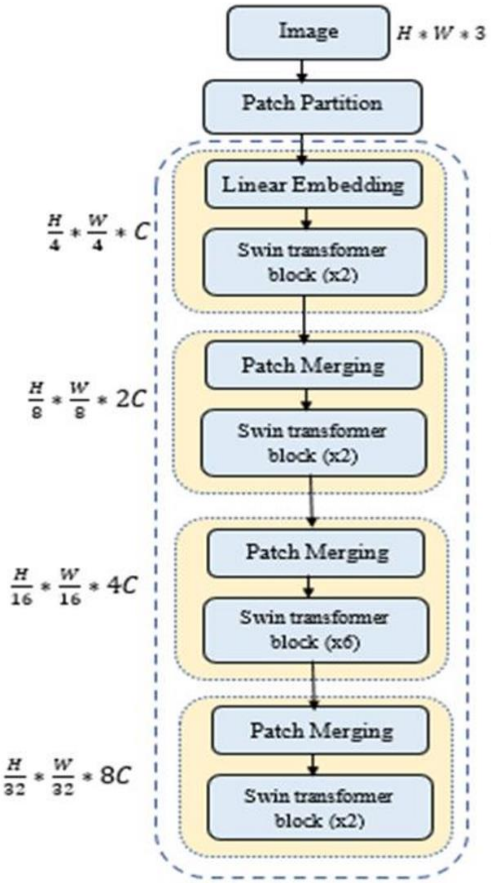


Image
$H*W*3$
Patch Partition
Linear Embedding
Swin transformer block (x2)
$\frac{H}{4}*\frac{W}{4}*C$
Stage 1
Patch Merging
Swin transformer block (x2)
$\frac{H}{8}*\frac{W}{8}*2C$
Stage 2
Patch Merging
Swin transformer block (x6)
$\frac{H}{16}*\frac{W}{16}*4C$
Stage 3
Patch Merging
Swin transformer block (x2)
$\frac{H}{32}*\frac{W}{32}*8C$
Stage 4

Figure 4. Swin Transformer

An overview of each stage is shown in Figure 4.

A patch partition: image input is split into patches.

Linear embedding: Embed patches into vectors linearly.

Stage 1: A pair of Swin Transformer blocks process sequential patches using self-attention with two processing steps.

Patch Merging: Combination of patches to form larger patches by decreasing the spatial dimensions.

Stage 2: The merged patches are treated with another set of Swin Transformer blocks.

Stages 3 - 4: Merging the patches and blocks of the Swin Transformer will step to further reduce the resolution and enhance the features representation.

It efficiently processes large images with high performance and is suitable for a wide range of tasks in the visual data analysis category.

By applying ST and due to the hierarchical structure of this model, small image patches can be integrated into deeper transformer layers, which in turn will facilitate the ability to model different scales for classification of image metadata. The algorithm demonstrates linear computing complexity by partitioning the image into non-overlapping windows and confining self-attention actions to each window. The Swin Transformer's ability to scale effectively makes it well-suited for a wide range of computer vision (CV) applications. Figure 5 illustrates how the suggested strategy uses the Swin Transformer as a classifier.

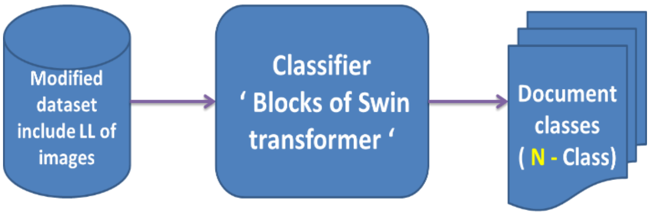

Modified dataset include LL of images
Classifier ' Blocks of Swin transformer '
Document classes (N - Class)

Figure 5. Using Swin Transformer

The combination of DWT with Swin Transformer provides a synergistic advantage in extracting features, resulting in a more holistic comprehension of document information.

Our approach helps in overcoming the limitations associated with existing methods and the integration of DWT and Swin Transformer for document image classification is a unique way. The method integrates the multi-resolution and hierarchical feature extraction abilities of DWT, which can be able to extract the overall properties of substructures at different scales and levels of detail from the angle. Indeed, it also improves its noise and variability resistance, lessens the effect of noise, and captures complex patterns and long dependencies. The Swin Transformer is a design specifically designed to be computationally efficient and scalable while being easy to use without increasing time or complexity. It will also improve the ability of pattern recognition to accurately recognize complex patterns and structures. It is the first attempt at document image classification, where DWT and Swin Transformer are introduced together and results in outperforming works. This multiplier effect addresses many challenges.

In this paper, we have studied the integration of DWT and swing transformers for faster document image classification. This work emphasizes feature extraction with DWT, to extract both high-frequency and low-frequency components, and a fair hierarchical representation with Swin Transformers, that deal with large images efficiently and capture-mid-level spatial hierarchy. We evaluate the performance of the combined DWT and Swin Transformer with the existing traditional and state-of-the-art methods. Some important areas covered are methodology, data preprocessing, model training and optimization, experimental results, and functional applications. The paper ignores any nonvisual data analysis techniques, other machine learning models, or any real-time document classification systems. In this paper, we will delve deeper into the use of DWT and Swin Transformers and establish their efficacy in documentation.

4. RESULTS AND DISCUSSION

In this section, the most important results were achieved by applying DWT with a Swin Transformer described in this part, to detect and predict image documents, as shown in the Detecting Image documents using (Swin-T). The metrics of precision, recall, f-measure, accuracy, sensitivity, and specificity will be calculated using Eqs. (1)-(6) to assess the performance of a classifier [25].

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Positive instance recall or sensitivity are highly impacted by the true positives rate (TP) and false positive rate.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

The accuracy, false-negative rate (FN) and percentage of accurate predictions are computed using the following formula.

$$Accuracy = \frac{TP + FP}{TP + TN + FP + FN} \tag{3}$$

True negative is denoted by "TN", whereas "sensitivity" refers to the quantity of positive records that yield the intended outcome.

$$Sensitivity = TP/TP + FN \tag{4}$$

Accurately sorting positive records from each positive paper is what is meant by particularity.

$$Specificity = \frac{TN}{TN} + FP \tag{5}$$

The F-measure analyzes measurements and performs numerous data recovery accuracy norms.

$$F1\,Score = 2 * (Recall * Precision)/(Recall + Precision) \tag{6}$$

FN is used in erroneous classification, whilst TP and FP are used in proper classification. The sensitivity and specificity of a test define how accurately it can classify documents. The results are likely to predominantly reflect the accuracy of true positives when the emphasis is skewed and false positives are neglected. On the other hand, recall will be reflected in the scores if false positives are highlighted at the expense of actual positives. Success can also be measured quantitatively through the use of several methods from through compare the performance of the proposed method with currently used state-of-the-art methods on the above metrics. I will be counting the increase in posts with greater accuracy, precision, recall, and F1 scores as success, all without making the calculations slower. Cross Validation K-fold cross-validation will be used to validate the results. In this technique, we split the dataset in k parts where we train the model in k-1 parts and validate it in the remaining part. You do this k times and then average over the performance across all the iterations. Statistical Significance Testing will use statistical tests, such as paired t-tests, to verify that using the proposed method, concerning baseline methods, indeed results in significant performance gains.

This part evaluates the scalability and generalizability of the method so that the proposed method can process large datasets and can generalize for other.

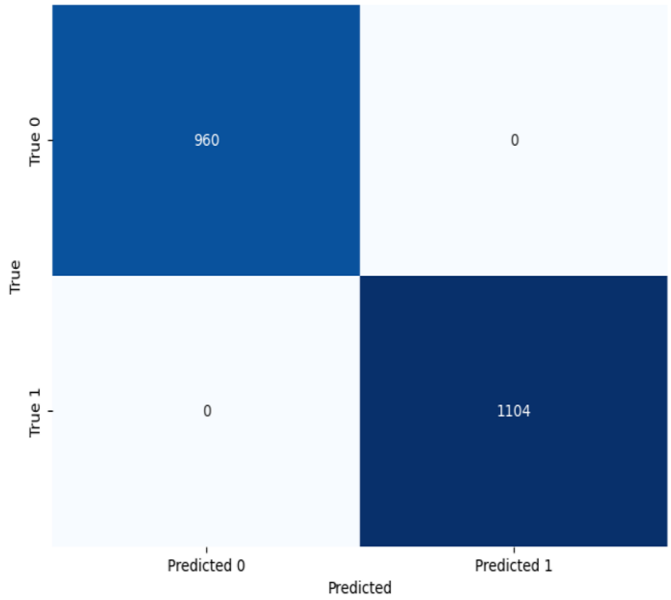
The depicted image in Figure 6 is a confusion matrix, a commonly employed tabular representation that describes the effectiveness of a classification model on a certain dataset of test data, where the actual values are already known. The x-axis of this confusion matrix reflects the model's anticipated classifications, specifically labeled as "predicted 0" and "predicted 1". The y-axis (vertical) depicts the true classifications, labeled as "True 0" and "True 1". Each cell in the matrix displays the count of predictions generated by the model, as indicated:

-The model accurately predicted the negative class 960 times, specifically in the class of true negatives (TN).

-The model achieved perfect accuracy in predicting the positive class, with zero errors in the form of false positives (FP).

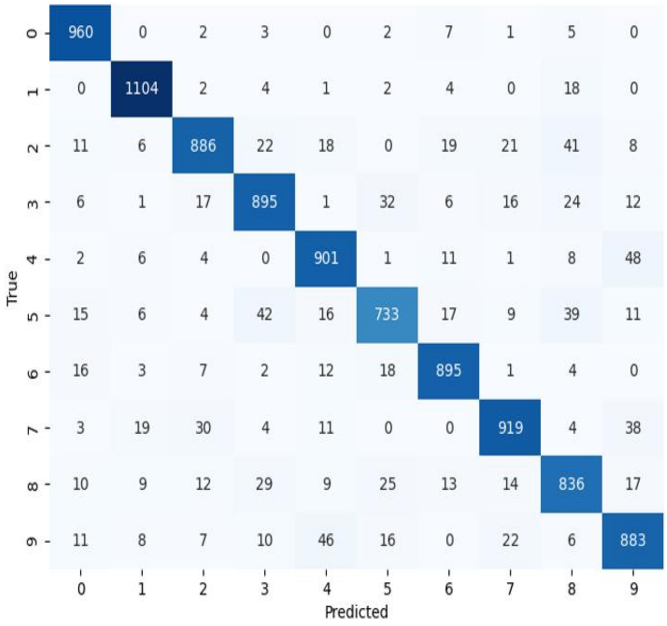
-The model achieved perfect accuracy in predicting the negative class, with 0 false negatives (FN).

-The model accurately predicted the positive class 1104 times. Part of the data is correctly identified as true positives (TP).



	Predicted 0	Predicted 1
True 0	960	0
True 1	0	1104

Figure 6. Detecting image documents using Swin-T



	0	1	2	3	4	5	6	7	8	9
0	960	0	2	3	0	2	7	1	5	0
1	0	1104	2	4	1	2	4	0	18	0
2	11	6	886	22	18	0	19	21	41	8
3	6	1	17	895	1	32	6	16	24	12
4	2	6	4	0	901	1	11	1	8	48
5	15	6	4	42	16	733	17	9	39	11
6	16	3	7	2	12	18	895	1	4	0
7	3	19	30	4	11	0	0	919	4	38
8	10	9	12	29	9	25	13	14	836	17
9	11	8	7	10	46	16	0	22	6	883

Figure 7. Predicted image documents using Swin-T

The confusion matrix demonstrates a flawless classifier, with no occurrences of false positives or false negatives. A confusion matrix for a situation with several classes is shown in Figure 7 and Table 1. This matrix differs from the last one in that it has many classes, with predicted and true values ranging from 0 to 9, as opposed to a binary classifier. When there are more than two potential outcomes to anticipate in a classification problem, this is the norm. For each class, the number of model predictions is displayed in the corresponding cell of the matrix: Each class's number of accurate predictions is shown by the diagonal cells, which run from top to bottom. For each category, these are called true positives. As an example, the model got class 0 960 times, class 1 1104 times, class 2 886 times, and so on right. Incorrect predictions, in which the model incorrectly predicted a class that did not exist, are displayed in the cells that are not on the diagonal. The model incorrectly classified 11 cases from class 0 as class 2, 6 cases from class 1 as class 2, etc. In the majority of classification tasks, the objective is to minimize the number of off-diagonal predictions and maximize the number of diagonal predictions, which represent the right predictions. More occurrences of that specific outcome are usually indicated by a deeper color, which signifies a higher value in this representation. If there are notable values outside the diagonal, it means the model made a mistake in its classification. Class 2 was erroneously labeled as class 8 41 times and class 4 as class 9 48 times; this could indicate that the model is confused about the similarities between these classes.

Table 1. Classification image documents using Swin-T

No.	Precision	Recall	F1-score	Support
0	0.93	0.98	0.95	980
1	0.95	0.97	0.96	1135
2	0.91	0.86	0.88	1032
3	0.89	0.89	0.89	1010
4	0.89	0.92	0.90	982
5	0.88	0.82	0.85	892
6	0.92	0.93	0.93	958
7	0.92	0.89	0.90	1028
8	0.85	0.86	0.85	974
9	0.87	0.88	0.87	1009
accuracy			0.90	10000
macro avg	0.90	0.90	0.90	10000
weighted avg	0.90	0.90	0.90	10000

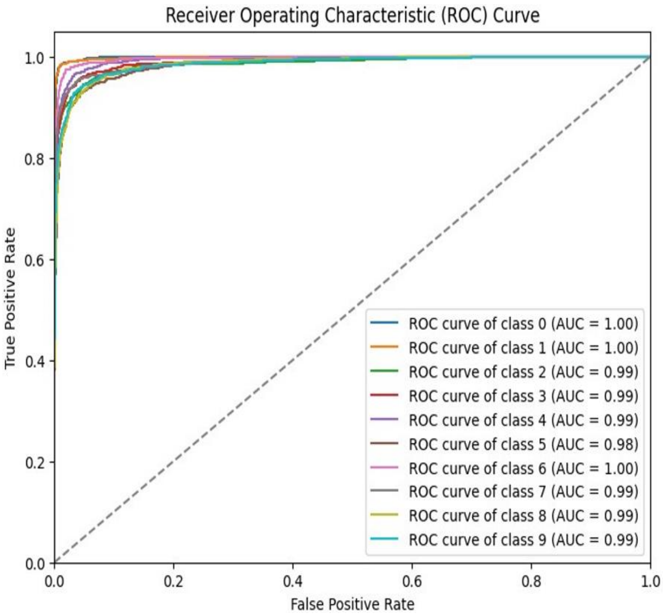


Figure 8. A receiver operating characteristic (ROC) curve for the suggested approach

As shown in Figure 8, ROC curve is a useful tool for evaluating the performance of a binary classifier system when its discrimination threshold is adjusted. It displays the TP rate in comparison to the FP rate at various threshold settings. A distinct class—a multi-class classification—is present in the figure. An Area Under the Curve (AUC) value, which indicates how well the classifier can distinguish between classes, is assigned to each curve. AUC values of 0.5 imply no discriminative power, whereas AUC values of 1.0 are ideal. The curves have a strong correlation with the upper left corner, signifying a high true positive rate and a low false positive rate. This suggests that the classifier operates exceptionally well across all classes. The AUC for classes 0 and 1 is 1.00, meaning that there are no false positives or false negatives, suggesting perfect categorization. The AUC values for the remaining classes (2–9) fall between 0.98 and 0.99, which is still regarded as excellent. A no-skill classifier is represented by the dashed line; a classifier with random prediction will, for instance, have a ROC curve that falls along this line (with an AUC of 0.5). All things considered, the classifier under test in this ROC curve is operating at a very high level in every class. Consistent performance across classes is indicated by the relatively tiny variances in AUC values.

Finally, compare the proposed method to similar studies from other authors, as shown in Table 2.

Table 2. Comparative analysis of the proposed method with other studies

Ref. No.	Method for Selecting Features	Classifier Approach	Highest Accuracy Performance
[4]	Latent Semantic Indexing (LSI) and cosine similarity: We employed a weighing technique called TF.IDF	SVM: Naïve Bayes, K-nearest Neighbors, Neural Networks, Random Forests, and classification trees	- The accuracy of the cosine similarity classifier with LSI features is 82.5%. - SVM attains an accuracy of 84.75%.
[6]	-Word-level and character-level n-gram models and the TF-IDF algorithm were used for feature extraction. -Hyperparameters for text categorization were established using 5-fold cross-validation.	Machine learning methods include SVM, DTC, RF, LR, and MNB.	The maximum SVM accuracy was 86.88% utilizing the radial basis function kernel.
[8]	Vector control method and TF-IDF	Machine learning methods include LR, KNN, SVM, Naïve Byes, AdaBoost, and TPOT	TFIDF and counter vectorizer achieve 81% TPOT macro average accuracy with LR's f1 score.
[11]	Use VGG-19 for feature extraction	SVM and gradient-boosting Methods - multimodal mech learning multi-classifier for document image classification	SVM outperforms Adaboost with 0.964. 0.853.
[12]	The paper did not indicate it	Using different models Bi-LSTM, LSTM, and ANN	The accuracy of BiLSTM is 85.14 percent.
The proposed mothed	Feature extraction using DWT	DWT and Swin Transformer	A median accuracy of 90.1% was attained by the suggested method on the standard Tobacco3482.

5. CONCLUSION AND FUTURE WORK

In this research we apply DWT with Swin Transformer in document image classification offers a promising way to enhance document classification tasks by utilizing the power of DWT for improved document quality Utilizing the approximate band presents a lower number of coefficients in comparison to other bands after the image decomposition procedure, resulting in enhanced efficiency of the classifier's operations and expedited feature analysis and classification performance, resistance to interference changes, and improved Swin transform feature extraction using self-attention processes to assess long-term connections and relationships between picture patches, enabling precise recognition of complicated layouts and patterns in documents. and gathers contextual data for categorizing document types and components. DWT and Swin Transformer work synergistically to extract characteristics, improving content comprehension. With advanced methods, domain-specific expertise, and an emphasis on practical applications, the obtained results were found to be consistent with previous searches.

ACKNOWLEDGEMENT

The authors would like to thank Informatics Institute of Postgraduate Studies Iraqi Commission for Computers & Informatics (https://iips.edu.iq/): Baghdad-Iraq, for its support in the present work.

REFERENCES

[1] Alaei, F., Alaei, A., Blumenstein, M., Pal, U. (2016). A brief review of document image retrieval methods: Recent advances. In 2016 International Joint Conference on Neural Networks (IJCNN): Vancouver, BC, Canada, pp. 3500-3507. https://doi.org/10.1109/IJCNN.2016.7727648

[2] El Barbary, O.G. (2020). Document classification in information retrieval system based on neutrosophic sets. Infinite Study.

[3] Wu, F., Ji, Y., Shi, W. (2022). Design of a computer-based legal information retrieval system. Computational Intelligence and Neuroscience, 2022(1): 6942773. https://doi.org/10.1155/2022/6942773

[4] Liu, L., Wang, Z., Qiu, T., Chen, Q., Lu, Y., Suen, C.Y. (2021). Document image classification: Progress over two decades. Neurocomputing, 453: 223-240. https://doi.org/10.1016/j.neucom.2021.04.114

[5] Chen, J.A., Hou, J.C., Tsai, R.T.H., Liao, H.M., Chen, S.P., Chang, M.C. (2024). Image classification for historical documents: A study on Chinese local gazetteers. Digital Scholarship in the Humanities, 39(1): 61-73. https://doi.org/10.1093/llc/fqad065

[6] Kölsch, A., Afzal, M.Z., Ebbecke, M., Liwicki, M. (2017). Real-time document image classification using deep CNN and extreme learning machines. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, pp. 1318-1323. https://doi.org/10.1109/ICDAR.2017.217

[7] Das, A., Roy, S., Bhattacharya, U., Parui, S.K. (2018). Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks. In 2018 24th International Conference on Pattern Recognition (ICPR): Beijing, China, pp. 3180-3185. https://doi.org/10.1109/ICPR.2018.8545630

[8] Mohsenzadegan, K., Tavakkoli, V., Kyamakya, K. (2021). A deep-learning based visual sensing concept for arobust classification of document images under real-world hard conditions. Sensors, 21(20): 6763. https://doi.org/10.3390/s21206763

[9] Qureshi, R., Uzair, M., Khurshid, K., Yan, H. (2019). Hyperspectral document image processing: Applications, challenges and future prospects. Pattern Recognition, 90: 12-22. https://doi.org/10.1016/j.patcog.2019.01.026

[10] Al-Anzi, F.S., AbuZeina, D. (2017). Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. Journal of King Saud University-Computer and Information Sciences, 29(2): 189-195. https://doi.org/10.1016/j.jksuci.2016.04.001

[11] Liu, Z., Lin, Y.T., Cao, Y., Hu, H., Wei, Y.X., Zhang, Z., Lin, S., Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 9992-10002. https://doi.org/10.1109/ICCV48922.2021.00986

[12] Rabbimov, I.M., Kobilov, S.S. (2020). Multi-class text classification of Uzbek news articles using machine learning. Journal of Physics: Conference Series, 1546(1): 012097. https://doi.org/10.1088/1742-6596/1546/1/012097

[13] Tian, C., Zheng, M., Zuo, W., Zhang, B., Zhang, Y., Zhang, D. (2023). Multi-stage image denoising with the wavelet transform. Pattern Recognition, 134: 109050. https://doi.org/10.1016/j.patcog.2022.109050

[14] Dhar, P., Abedin, M.Z. (2021). Bengali news headline categorization using optimized machine learning pipeline. I. J. Information Engineering and Electronic Business, 13(1): 15-24. https://doi.org/10.5815/ijieeb.2021.01.02

[15] Guo, W., Xu, G., Liu, B., Wang, Y. (2022). Hyperspectral image classification using CNN-enhanced multi-level Haar wavelet features fusion network. IEEE Geoscience and Remote Sensing Letters, 19: 1-5. https://doi.org/10.1109/LGRS.2022.3167535

[16] Liu, B., Liu, Y., Zhang, W., Tian, Y., Kong, W. (2023). Spectral Swin Transformer network for hyperspectral image classification. Remote Sensing, 15(15): 3721. https://doi.org/10.3390/rs15153721

[17] Mahmoud, W.A., Stephan, J.J., Razzak, A.A.W. (2020). Facial expression recognition using fast Walidlet hybrid transform. Journal Port Science Research, 3(1): 59-69. https://doi.org/10.36371/port.2020.3.4

[18] Shahin, M.M.H., Ahmmed, T., Piyal, S.H., Shopon, M. (2020). Classification of Bangla news articles using bidirectional long short term memory. In 2020 IEEE Region 10 Symposium (TENSYMP): Dhaka, Bangladesh, pp. 1547-1551. https://doi.org/10.1109/TENSYMP50017.2020.9230737

[19] Williams, T., Li, R. (2016). Advanced image classification using wavelets and convolutional neural networks. In 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA): Anaheim, CA, USA, pp. 233-239. https://doi.org/10.1109/ICMLA.2016.0046

[20] Yang, Y., Nie, J., Kan, Z., Yang, S., Zhao, H., Li, J. (2021). Cotton stubble detection based on wavelet decomposition and texture features. Plant Methods, 17: 113. https://doi.org/10.1186/s13007-021-00809-3

[21] Sutha, S., Leavline, E.J., Singh, D.A.A.G. (2013). A comprehensive study on wavelet based shrinkage methods for denoising natural images. WSEAS Transactions on Signal Processing, 9(4): 203-215.

[22] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929

[23] Xu, Y., Wang, X., Zhang, H., Lin, H. (2024). SE-Swin: An improved Swin-Transformer network of self-ensemble feature extraction framework for image retrieval. IET Image Processing, 18(1): 13-21. https://doi.org/10.1049/ipr2.12929

[24] Xia, H.H., Gao, H., Shao, H., Gao, K., Liu, W. (2023). Multi-focus microscopy image fusion based on Swin Transformer architecture. Applied Sciences, 13(23): 12798. https://doi.org/10.3390/app132312798

[25] Salman, A.H., Waleed, A., Al-Jawher, M. (2023). Image document classification prediction based on SVM and gradient-boosting algorithms. Journal Port Science Research, 6(4): 348-356. https://doi.org/10.36371/port.2023.4.5