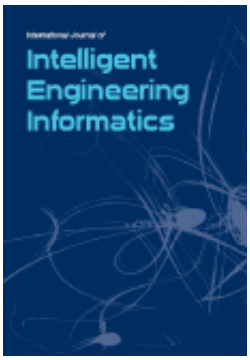
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Intelligent Engineering Informatics

ISSN online: 1758-8723 - ISSN print: 1758-8715
https://www.inderscience.com/ijiei

Fine-tuned convolutional neural networks for feature extraction and classification of scanned document images using semi-automatic labelling approach

Krishna Kumar, Nakkala Srinivas Mudiraj, Meenakshi Mittal, Satwinder Singh

DOI: 10.1504/IJIEI.2024.10062930

Article History:

Received:	10 August 2023
Last revised:	11 December 2023
Accepted:	16 December 2023
Published online:	02 April 2024

Fine-tuned convolutional neural networks for feature extraction and classification of scanned document images using semi-automatic labelling approach

Krishna Kumar, Nakkala Srinivas Mudiraj, Meenakshi Mittal and Satwinder Singh*

Department of Computer Science and Technology,
Central University of Punjab,
Bathinda, India
Email: krishnamathsforyou@gmail.com
Email: srinivas.mudhiraj2111@gmail.com
Email: meenakshi@cup.edu.in
Email: satwinder.singh@cup.edu.in
*Corresponding author

Abstract: Organising documents into relevant categories through image classification is crucial for management and safeguarding of valuable information. Many studies have done work on it with manual intervention, but still there is a scope of improvement. After finding gaps in existing studies, this research fine-tuned a hyper-parameter of pre-trained model based on various convolutional neural networks (CNNs), specifically the EfficientNetB3 and DenseNet201 models, for feature extraction and classification. These models are fine-tuned with the subset of the Ryerson Vision Lab Complex Document Information Processing (RVL_CDIP) dataset. The dataset comprises 16,000 image-scanned documents categorised into 16 classes with semi-automatic approach of labelling. The modified models are fine-tuned by adding a few more layers. The modified models outperformed in terms of accuracy, precision, recall and F1-Score for EfficientNetB3 and DenseNet201. These results highlight a significant improvement when comparing the proposed CNN models with baseline models through the utilisation of semi-automatic labelling and fine-tuning.

Keywords: convolutional neural networks; CNNs; document image classification; deep learning; hyperparameter tuning; image based classification; semi-automatic labelling; text-based classification; transfer learning.

Reference to this paper should be made as follows: Kumar, K., Mudiraj, N.S., Mittal, M. and Singh, S. (2024) 'Fine-tuned convolutional neural networks for feature extraction and classification of scanned document images using semi-automatic labelling approach', *Int. J. Intelligent Engineering Informatics*, Vol. 12, No. 1, pp.103–134.

Biographical notes: Krishna Kumar was a research scholar in the Department of Computer Science and Technology. He has completed his MTech in 2023. He has completed his BTech in 2021. His major area of interest is data analysis and visualisation for image-based data.

Nakkala Srinivas Mudiraj is a Junior Research Fellow in the Department of Computer Science and Technology, Central University of Punjab, Bathinda. He is working under the ICMR sponsored project. He has done his MTech in 2020 from Central University of Punjab, Bathinda. His major area of research is Computer Vision and Image Processing. He has a total three year of research experience.

Meenakshi Mittal is an Assistant Professor in the Department of Computer Science and Technology since 2011. She has done her MTech from Punjab Engineering College, Chandigarh. She has over 13 years of vast experience in teaching and research. Her major area of research is analysis with deep learning and machine learning techniques. She had published article also in the area of DDoS using deep learning approaches.

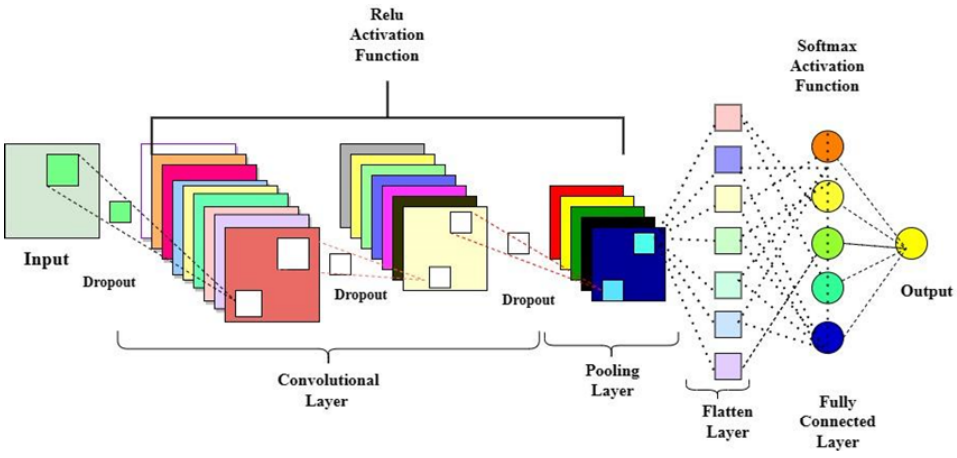
Satwinder Singh serves as the Dean, School of Engineering and Technology at the Central University of Punjab. He earned his MTech and PhD degrees from Guru Nanak Dev University, Amritsar. With an extensive background spanning two decades, he has accumulated substantial experience in both teaching and research. He is affiliated with several professional societies, including IET, IE(I), ISTE, and CSI. His research interests encompass various aspects such as data analysis, data visualisation, data science, and software engineering. Since 2006, he has been actively involved in guiding Master's and Doctoral students.

1 Introduction

Document image classification automatically categorises digital images of documents based on their content or visual characteristics. The classification of document images holds significance as it involves categorising documents into appropriate groups, which aids in their organisation. This procedure simplifies the task of managing, analysing, and safeguarding the valuable information present in the documents. By employing diverse machine learning and deep learning methods, document image classification algorithms categorise documents into distinct types, including invoices, resumes, contracts, and others. This enables expedited and precise document processing, resulting in improved efficiency.

Various document image processing systems have been developed that use the machine and deep learning models to extract information from documents, focusing solely on visual resources (Sarkhel and Nandi, 2019; Sharma and Kumar, 2020; Luo et al., 2021; Kim et al., 2022; Li et al., 2022), solely on textual resources (Seuret et al., 2017; Rasjid and Setiawan, 2017; Qazi and Goudar, 2018; Sahare and Dhok, 2018; Baygin, 2019; Bakkali et al., 2021), or a combination of both (Afzal et al., 2017; Kanchi et al., 2022; Bakkali et al., 2020). However, a significant challenge in document image processing is the wide range of layout formats utilised in real-world scenarios (Aly and Nguyen-an, 2018; Xu et al., 2018; Riba et al., 2019; Bhowmik et al., 2021; Pfitzmann et al., 2022). To perform a thorough analysis of document layouts, it is crucial to have the ability to comprehend texts presented in a diverse range of formats. Accomplishing this requires a skilful blending of computer vision and natural language processing (NLP) techniques.

Figure 1 The general structure of convolution neural network (see online version for colours)



Lately, there has been an increasing fascination with methods that acquire knowledge directly from data, and we have chosen to follow the same approach (Singh and Singla, 2017; Ali Reshi and Singh, 2018). Out of the various feature learning methods available, the ones based on convolutional neural networks (CNNs) (Afzal et al., 2017; Kaur and Singh, 2016; Das et al., 2018; Audebert et al., 2020; Siddiqui et al., 2021; Sharma and Singh, 2020; Omurca et al., 2022; Bakkali et al., 2023; Jeyanthi et al., 2023) have garnered the most attention. CNNs have proven highly effective in achieving state-of-the-art performance by utilising convolutional layers for feature learning (Alkhonin et al., 2020; Khan et al., 2021; Selvakumar and Thangaraju, 2023) in case of image analysis. CNNs are designed to automatically learn and extract relevant features from input data, primarily images, for subsequent tasks like object recognition, classification, and segmentation. The architecture of a CNN typically comprises a series of convolutional and pooling layers, culminating in a fully connected (FC) layer responsible for classification. Figure 1, shows the general CNN architecture. The convolutional layer plays a crucial role in feature extraction. Feature extraction can be accomplished with the following steps:

1 *Convolutional layers*: CNN uses a convolution layer that applies a sequence of convolution operations, utilising convolutional filters or kernels with a specified stride value and commonly employing the rectified linear unit (ReLU) activation function. Multiple filters are employed to extract various features from the input image. The feature maps produced by these layers highlight relevant local patterns within the image.

2 *Hierarchical feature extraction*: CNNs have multiple convolutional layers arranged hierarchically. As you move deeper into the network, these layers learn increasingly complex and abstract features.

3 *Pooling layers*: Subsequently, the pooling layer reduces the size of the image representation through pooling operations, such as maximum or average pooling. This step enhances the model's generality by capturing essential information while reducing the overall dimensionality. This also helps to capture invariances and reduces computational complexity.

The output from the convolutional and pooling layers is then flattened, transforming into a one-dimensional tensor. This tensor is subsequently passed as input to the FC layer. The FC layer calculates the probability scores for each label in the training dataset and generates the final output using the Softmax activation function.

There are multiple pre-trained models built on CNNs such as ResNet, DenseNet, and EfficientNet. Each of these models has distinct architectural variations and benefits. These models have shown remarkable effectiveness in tasks like document image classification and similar endeavours. The applications of the image classification can be found in Anter et al. (2013), Aziz et al. (2013), Jothi et al. (2013), Emary et al. (2014a, 2014b), Suganthi and Sathiaseelan (2022) and Chelliah et al. (2023a, 2023b).

1.1 Motivation and contribution

In today's world, everything is in online mode such as job applications, admission applications, document verification forms, client feedback forms, event registration forms, etc. In all of this, a client must require the fast and furious method to determine the type of document to facilitate the task of classification. In many cases, there are problems that for any specific documents, applicants submit random documents and forms. It becomes hectic for employers to handle such documents manually. So there is a need to automate this process. For this purpose, the current study will propose a deep learning model which will classify the different types of documents and save them in their respective folder. This research makes several significant contributions in the light of the above motivation:

- The IndoML Datathon 2022 dataset has been pre-processed to ensure data consistency, which includes resizing, rotating, and cropping.
- Labelled the IndoML Datathon 2022 dataset with a semi-automatic approach.
- The two models, EfficientNetB3 and DenseNet201, are modified by adding some layers and are fine-tuned using the pre-processed dataset for the classification of document image dataset.
- The document image dataset is classified into 16 classes or types of documents with modified models.
- The models are rigorously evaluated and compared with pre-existing techniques on the basis of different measuring parameters (precision, recall, accuracy and F1 score), and detailed classification reports for each class of the dataset have been provided.

The paper is structured as follows: Section 2 reviews the related work on document image classification. Section 3 presents our methodology, including the complete training procedure and the selection of hyperparameters. In Section 4, we report the experiments and results of the fine-tuned model. Finally, in the last section, we provide a summary and conclusions of our study, as well as discuss potential future strategies for improving the accuracy of document image classification.

2 Related work

Classification is a process to classify the objects into two or more classes. Various studies has classified the text (Singh and Singla, 2017; Ali Reshi and Singh, 2018; Audebert et al., 2020) for classification of text into two or more classes. The current study of image-based document classification applies image-based classification techniques (Anter et al., 2013; Aziz et al., 2013; Jothi et al., 2013; Emary et al., 2014a, 2014b; Thiruvenkatasuresh and Venkatachalam, 2019; Ferrando et al., 2020; Siddiqui et al., 2021). researchers have employed text-based, image-based and fusion (using both image and text) techniques for the image-based document classification. The following literature review is classified as per these techniques.

2.1 Text-based classification techniques

In the early days, documents were classified based on text extraction. One of the major techniques for text extraction is the optical character recognition (OCR). Audebert et al. (2020) have used the OCR for image-based document classification. The authors of this study have included the multimodal technique, recognising the text structure with OCR and image features with modern CNN techniques. For the text-based analysis apart from the OCR, many studies have used the machine learning approach that includes the k-NN (Rasjid and Setiawan, 2017), naive Bayes (Rasjid and Setiawan, 2017; Khan and Mollah, 2020), random forest, SVM, and AdaBoost (Khan and Mollah, 2020; Sharma and Singh, 2020). As reported in these studies the best-performing algorithm is a random forest, with an accuracy of 87.59%. In other cases like SVM, AdaBoost and naive Bayes, the maximum accuracy is achieved up to 86%. Out of these studies, Khan and Mollah (2020) have created their own dataset for addressing the problem of component-level object classification in complex scene and document images for the text/non-text-based classification in documents. In a separate investigation conducted by Tran et al. (2018), a learning approach centred on extracting text and non-text elements was employed for document classification. The study focuses on the analysis of white space in the maximum horizontal homogeneous region. This study includes classification and segmentation stages using mathematical morphology and machine learning approaches. The study has recorded a good F1 score of 82.61% for full text recognition on UW-III (A1) dataset. For the text-based classification, the researcher has also employed the NLP approach for the text-based classification of image documents. Combining NLP (for text) with a neural network (for image) approach has made a significant contribution compared to the machine learning approaches for document classification, as reported by Rabut et al. (2019). They proposed the classification task using Word2Vec and FastText word embedding methods to generate custom-built word embedding vectors. The study has incorporated part-of-speech (POS) tag vectors to provide additional semantic information about the words in the corpus. The experimental results showed that their model outperformed the existing classification models over 20 newsgroup dataset (collected by the authors), such as naïve Bayes, linear support vector machine, and logistic regression.

Further, in continuation to the text-based document classification technique, Riba et al. (2019), proposed a graph-based approach for table detection in document images, utilising the location, context, and content type instead of raw textual data. They employ graph neural networks (GNNs) to capture tables' local repetitive structural information in invoice documents. The proposed model of the study is trained in a supervised manner

with table data and has achieved promising results on two invoice datasets and addresses. The results of the study demonstrate robust performance with an F1-score of 78.4% on the CON-ANONYM dataset. Future research directions include exploring the generalisability of the architecture to other unconstrained tabular layouts.

2.2 Image-based technique

Most of the studies for image document classification have primarily used the dataset of RVL-CDIP for the training of their models (Afzal et al., 2017; Kolsch et al., 2018; Hassanpour and Malek, 2019; Ferrando et al., 2020; Siddiqui et al., 2021). This dataset includes the 400,000 un-labelled images of 16 classes. Another dataset which is being used by the studies is IIT-CDIP (Harley et al., 2015); this dataset is a subset of the Legacy Tobacco Document Library, known as tobacco dataset of 3,482 images of nine classes. Many of the studies who have used the RVL-CDIP dataset have used the tobacco dataset for third-party validation (Hassanpour and Malek, 2019; Ferrando et al., 2020; Siddiqui et al., 2021). Deep CNNs were used by Harley et al. (2015) for their study through extensive experiments on the IIT-CDIP dataset and achieved an accuracy of 89.3%. This research also contributes a subset of the IIT-CDIP labelled dataset collection, providing valuable resources for further investigations in document analysis. Another study by Kolsch et al. (2018), used the two-stage approach combining deep neural networks for feature extraction and extreme learning machines (ELMs) for classification. As claimed by the study, their method significantly improves accuracy over the Tobacco-3482 dataset and results in a 25% relative error reduction compared to previous CNN-based approaches. This approach makes deep learning-based document classification suitable for large-scale real-time applications. In Afzal et al. (2017), deep learning-based pre-trained models such as GoogleNet, VGG, and ResNet have extensively investigated document image classification over 400,000 records of RVL-CDIP dataset and validation with the Tabbaco dataset. The study has achieved an impressive accuracy of 90.97% on the RVL-CDIP dataset using VGG-16, which corresponds to an error reduction of 11.5%. The study has highlighted the importance of dataset size and network architecture for these impressive results. In continuation to the role of network architecture, another out-of-box study by Hassanpour and Malek (2019) has proposed document-based image classification by SqueezeNet networks. The study has demonstrated strong performance in image classification tasks comparable to state-of-the-art CNNs. This research evaluates the suitability of SqueezeNet for document classification and found that SqueezeNet achieves an accuracy of approximately 75% on the Tobacco-3482 dataset. One of the other major studies with the RVL-CDIP dataset was proposed by Ferrando et al. (2020). This study has proposed lightweight EfficientNet models over heavier CNNs for document classification tasks and has shown improvement in the results. The study also introduced an ensemble pipeline which achieved a new state-of-the-art accuracy of 89.47%. The study by Siddiqui et al. (2021) has compared the self-supervised and pre-trained models. They trained the ResNet-50 image encoder with two self-supervision methods (SimCLR and Barlow Twins). The results showed that self-supervised embeddings outperformed ImageNet pre-trained embeddings, achieving an accuracy of 86.75% (compared to 71.43%) on RVL-CDIP and 88.52% (compared to 74.16%) on the Tobacco-3482 dataset. These findings highlight the potential of self-supervised representations for document image classification, especially in scenarios with limited labeled data. This study has also optimised the model performance

through hyperparameter tuning and document-specific augmentations. This study of self-supervised and hyper-parameter tuning motivates for the investigation of more fine or optimised results.

2.3 Fusion based (both image and text) technique

Apart from the text and image-based technique for image-based document classification, few studies have fused both techniques for document classification. Researchers have also used the other dataset apart from the benchmark dataset such as RVL-CDIP and achieved good performance. Fused studies allow simultaneous learning of discriminant features from image and text based modalities (Wang et al., 2017; Engin et al., 2019; Jaume et al., 2019; Vu and Nguyen, 2020; Audebert et al., 2020; Bakkali et al., 2021). Many of the studies have used the CNN based approach to develop a fusion model for the classification of image-based documents. One of the studies by Wang et al, (2017), addresses the challenge of exploiting web meta-data for visual recognition. The proposed approach combines CNNs for modelling web text and images, utilising a multimodal fusion technique at both the decision and feature levels. The framework achieves a significant improvement in large-scale image classification on the Pascal VOC-2007 and VOC-2012 datasets, with the highest accuracy of 82.1%. Further, Jaume et al. (2019) have developed a FUNSD dataset. The dataset provides comprehensive annotations for text detection, OCR, spatial layout analysis, and entity labelling/linking tasks. It is the first publicly available dataset specifically designed for form understanding. The authors also present baselines and evaluation metrics tailored for the FUNSD dataset, establishing a foundation for advancements in document understanding. The dataset was used for multi-modality feature extraction (Jaume et al., 2019; Vu and Nguyen, 2020). Their work paves the way for developing end-to-end deep learning pipelines that address the challenges of form understanding. The evaluation results demonstrate that the vision model achieves a precision of 79.8%, while the faster R-CNN model achieves the best recall of 84.8% and the F1-score of 76% on the FUNSD dataset. Later with the help of the FUNSD dataset, Vu and Nguyen (2020) have developed a fusion model with a CNN-based approach of the U-Net model. The study used a dataset for the key-value detection task. The key-value detection network takes a two-channel input, comprising a text mask and a greyscale document image. The network architecture is based on a U-Net model with varying numbers of filters (16, 32, 64, 128, 256) across its layers. The authors employ a combination of dice loss and categorical cross-entropy loss as the loss function, and the final loss value is calculated using weights of four dice loss and 0.5 for cross-entropy loss. The model's performance is evaluated using mean Intersection over union (IoU) scores, which are calculated per class and then averaged. Their experiments demonstrate that by using the document image as input, significantly improves the results, achieving a mean IoU of 0.69 compared to 0.55 when using the text mask alone. The same dataset (FUNSD) is used in another study Appalaraju et al. (2021) to introduce DocFormer, a multi-modal transformer architecture for visual document understanding (VDU). It is pre-trained in an unsupervised manner using carefully designed tasks that promote multi-modal interaction. It combines text, vision, and spatial features through a novel multi-modal self-attention layer. The proposed transformer shares spatial embeddings across modalities, facilitating correlations between text and visual tokens. The authors evaluate the proposed transformer on four diverse datasets, achieving

state-of-the-art results on all of them, even surpassing larger models by up to 4 times in terms of parameters, with one of the parameters F1 score being 84.5%.

As reported in studies mentioned above, the experiments conducted on the Tobacco-3482 and RVL-CDIP datasets validated the effectiveness of the proposed approaches. All these have highlighted the standardisation of the datasets. Similarly, a study by Audebert et al. (2020) has highlighted the significance of hybrid image/text approaches in document classification tasks. By experimenting with the standard RVL-CDIP and Tobacco datasets, researchers have addressed the image-based document classification problem by utilising both image and text-based features. The study has proposed a multimodal neural network that incorporates word embeddings computed from OCR-extracted text and image features. The approach demonstrated improved accuracy, with an 87.8% accuracy on the Tobacco-3482 dataset and a 90.6% accuracy on the RVL-CDIP dataset, even without cleaning text information. By leveraging MobileNetv2 (a CNN-based approach) as the visual feature extractor and employing Tesseract OCR for text extraction, the authors developed an end-to-end learnable multimodal deep network that jointly learns and fuses both text and image features for classification.

Image-based document classification has a good utility in the legal community. It helps them to make their document classification task easy and automatic. One of the study (Luz de Araujo et al., 2022) has implemented this objective and achieved a F1-score of 65.64%. The study utilises a fusion module to combine visual and textual features extracted separately from neural networks trained on image and text data. The study shows that the multimodal approaches outperform both textual and visual models. This study had created their own dataset named SVic+ for the verification of the results. It is a novel dataset of Brazilian lawsuits with visual and textual information on legal documents.

The literature survey found that document image classification is an active research area with various applications. However, there is still room for improvement in accuracy and efficiency. Further, to find research gaps and motivation from above studies, the comparative study is presented in Table 1, with the focus on the dataset, model used, labelling approach number of classes and parameter tuning.

According to the aforementioned research, the majority of existing literature assesses datasets that are either pre-labelled or manually labelled. Moreover, these studies generally overlook the task of labelling unlabelled datasets. Additionally, a notable gap in the literature is the lack of emphasis on hyperparameter tuning to enhance model efficiency or model optimisation. As reflected in Table 1. a few of the studies used the hyperparameter tuning but very limited parameters were tuned, mostly Epochs and Learning rate were hypertunned. In contrast, the present study employs a semi-automatic labelling approach and enhances classification through hyperparameter tuning (Table 2) of modified CNN models(EfficentNetB3 and Densenet201). These models are fine-tuned with training data and are evaluated on the test data. Modified models used the Ryerson Vision Lab Complex Document Information Processing (RVL_CDIP) dataset for training and testing. Study has achieved state-of-the-art performance on this dataset and plans to release our code and trained model for further research. Study provides a promising direction for future research in document image classification.

Table 1 Characteristic comparisons of the proposed study with few existing studies

Reference	*Model used*	*Dataset used*	*Data labelling approach*	*No of classes*	*Hyper parameter tuning*
Bakkali et al. (2023)	VLCDoC	RVL-CDIP	Labelled	16	Epoch, learning rate and optimiser
Kanchi et al. (2022)	BERT & EfficentNetB0	RVL-CDIP and Tobacco-3482	Labelled	16 9	Epoch, delta and learning rate
Ferrando et al. (2020)	EfficientNet, BERT	RVL-CDIP and Tobacco-3482	Labelled	16 9	Epoch, learning rate and optimiser
Kolsch et al. (2018)	ELMs	Tobacco-3482	Labelled	9	No
Afzal et al. (2017)	AlexNet, GoogleNet ResNet-50	RVL-CDIP	Labelled	16	No
Audebert et al. (2020)	MobileNetV2	RVL-CDIP	Labelled	16	Epoch and learning rate
Siddiqui et al. (2021)	RESNet-50	RVL-CDIP	Labelled	16	Epoch, learning rate and scaling factor
Harley et al. (2015)	Ensemble of CNN	RVL-CDIP	Labelled	16	No
Current study	Modified EfficentNetB3 and modified Densenet201	RVL-CDIP	Semi-automatic labelling	16	Yes refer Table 2

3 Proposed methodology

The proposed methodology comprises five stages: data collection, data preparation, data labeling, data splitting, model building and implementation. The block diagram in Figure 2 illustrates the sequential steps involved in the process.

Figure 2 Block diagram of process flow (see online version for colours)

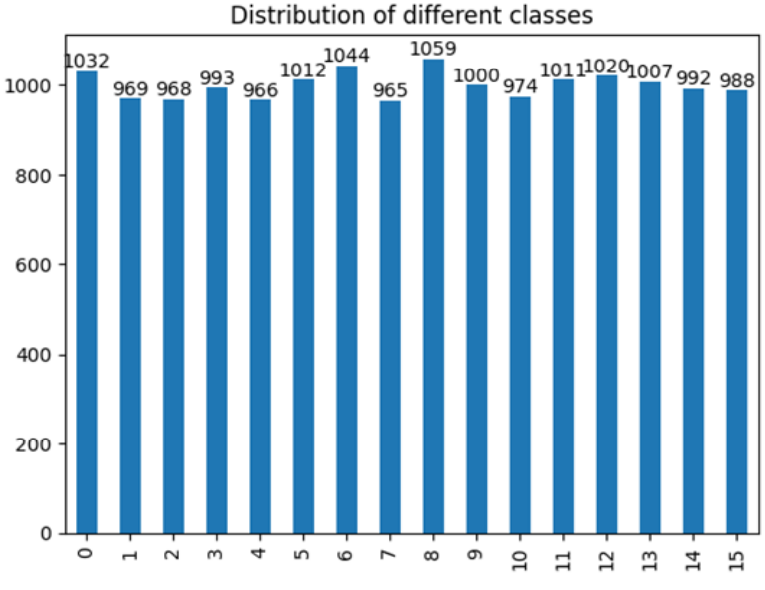
Figure 2 illustrates the standard five-stage data evaluation process. In the initial stage, data collection involves determining parameters such as the number of records and target classes. In the subsequent stage, data preparation addresses issues like missing values, outliers, and errors, as well as data transformation to a format suitable for modelling. The third stage involves data labelling using a semi-automatic method outlined in Section 3.3. Following data labelling, a train-test split is applied to enable model training and testing.

The fifth stage focuses on model building, as detailed in Section 3.5. Then, the model is trained with the training dataset, and the model's performance is assessed using the testing dataset. The detailed description is given as below.

3.1 Data collection

Data is the backbone of any deep learning algorithm, and the model's accuracy largely depends on the quality and type of data collected. The data collection process begins with identifying and collecting the required data type from appropriate sources, such as persons, places, or organisations. In this research, we used the IndoML Datathon 2022 Dataset which is the subset of the RVL_CDIP dataset. The IndoML dataset consists of 16,000 scanned document images of 16 classes and there is one CSV file which contains the label of each image as id, label. The distribution of the dataset according to different classes and sample images of the data are shown in Figure 3 and Figure 4, respectively.

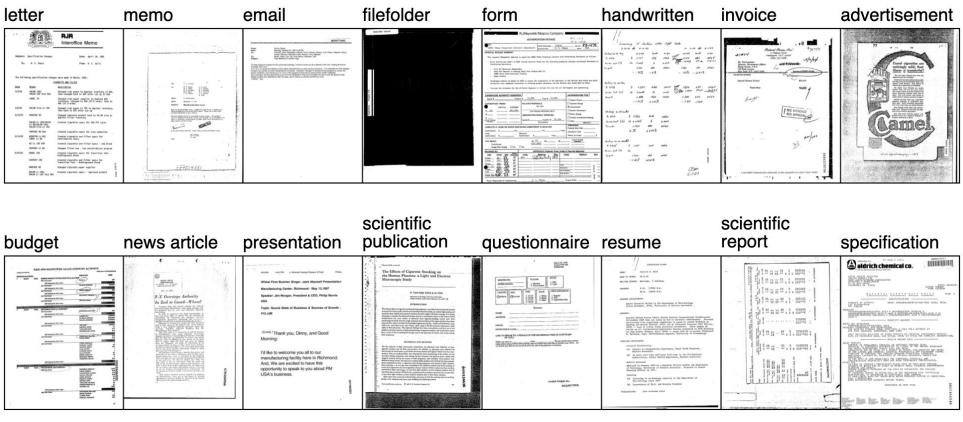
Figure 3 Distribution of dataset according to different classes (see online version for colours)



3.2 Data preparation

Once the data is collected, it undergoes preprocessing to clean it and eliminate noise. Noise in a dataset may arise due to deformation, distortion, poor quality, pinholes, irregular orientation, and image skewness, which may have occurred during the data collection or acquisition phase. Then perform resizing or cropping of images; in this dataset, input images also come in different sizes and resolutions, so they were resized to 224 × 224 × 3 pixels to ensure the size of each image is consistent during processing in the model.

Figure 4 Sample images of IndoML Datathon 2022 dataset



3.3 Data labelling

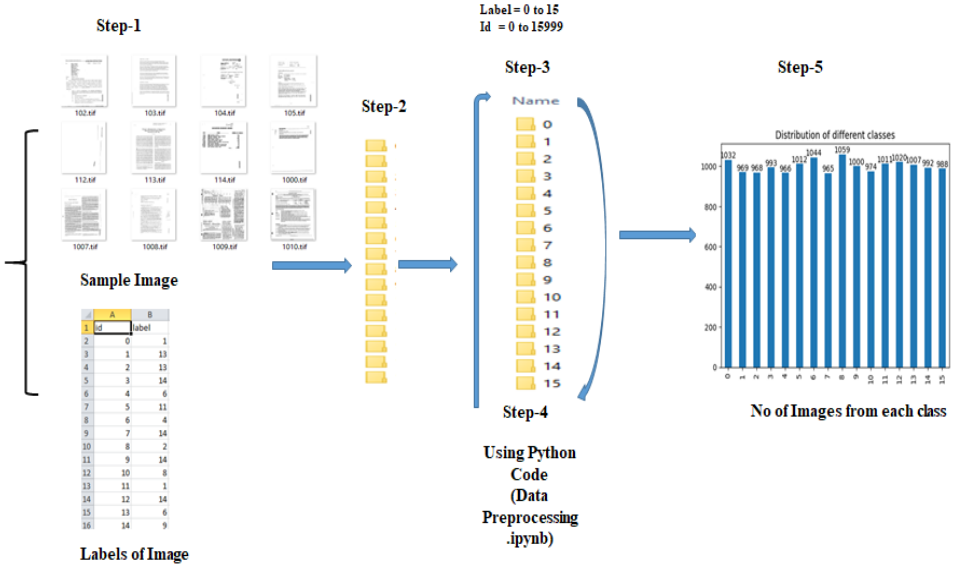
The IndoML Datathon 2022 dataset has been labelled using a semi-automatic approach. The process involved reading a CSV file containing information about the images, such as their id and labels. With the help of Python libraries like Pandas, the dataset was organised by creating 16 folders. Each image was then copied from a source directory to its respective folder based on its label with python script. This semi-automatic approach streamlines the labelling process and ensures the dataset is properly categorised for analysis and utilisation. Here is an algorithmic representation of the process described in Algorithm 1 and Figure 5 gives the visualisation explanation of data labelling process.

Algorithm 1 Semi-automatic approach for data labelling

Input: Image folder containing grayscale images and CSV File having Image id & Label
Output: Labelled images in respective 16 folders
START
def semi_automatic_labelling(image_folder, csv_file):
#Step 1: Read the CSV file containing image information into a DataFrame
    df = read_csv(csv_file)
#Step 2: Create 16 folders to organise the dataset
    for label_folder in range(16):
        folder_path = join(image_folder, label_folder)
        create_folder(folder_path)
    End for loop label_folders
#Step 3: For each row in the DataFrame
Step 3a: Extract the image ID and label from the respective columns of CSV File
    for index, row in df.iterrows():
        image_id = row['ImageID']
        label = row['Label']
End for loop index, row
#Step 3b: Create a folder with the label as the name if it doesn't exist already (0 to 15)

```
    if 0 <= label <= 15 and str(label) not in label_folders:
        folder_path = join(image_folder, str(label))
        create_folder(folder_path)
#Step 3c: Copy the image from the source directory to the corresponding label folder
    if 0 <= image_id <= 15999:
        copy_image(source_image_path[image_id], destination_image_path[image_id])
End for loop DataFrame
END
```

Figure 5 Illustration of data labelling process (see online version for colours)



3.4 Data splitting

Once the data pre-processing is complete, the next step is to split the dataset. The dataset contains 16,000 scanned document image samples, with 16 classes in 16 folders according to their id and labels. This dataset is split into train, test, and validation data in an 80:10:10 ratio.

3.5 Model building and implementation

CNNs are famous for document image classification due to their ability to extract features from image data. Instead of building a CNN from scratch, using a pre-trained CNN architecture can be advantageous. Pre-trained models such as VGG, ResNet, DenseNet, and Inception have performed well on computer vision tasks, including image classification. The block diagram of fine-tuned EfficientNetB3 and DenseNet201 models are shown in Figures 6 and 7, respectively. However, fine-tuning is necessary to improve their performance on a specific dataset. This study fine-tuned two models (EfficientNetB3 and DenseNet201) by adding four extra layers: one batch normalisation

layer, one dropout layer, two dense layers. Among the two dense layers, the last layer uses a softmax activation function which works as an output layer as shown in Figures 8 and 9 respectively. The input layer of the model takes 224 × 224 images as input, and the first convolutional layer uses 64 filters or 'local receptive fields'. The convolutional layers have a fixed kernel size of 3 × 3 with the same padding, which remains constant when the output size is the same as the input size or when the stride size is 1. The ReLU activation function ensures no negative value passes to the next layer. Figures 10 and 11 provide a summary of the modified models (EfficientNetB3 and DenseNet201). The fine-tuned parameter values set in our modified models during training to attain the desired performance are described in Table 2.

Table 2 Hyperparameter values used for training the modified models (EfficientNetB3 and DenseNet201)

Name of hyperparameter	*Values used*	*Best value*	*Description*
Batch size	16, 32, 40, 64	32	The number of images processed together during training to update the model's parameters and improve image classification accuracy
Epochs	5, 10, 15, 20, 25, 30	23 (EfficientNetB3), 22 (DenseNet 201)	A full iteration through the entire set of training images during training to update the model's parameters and improve the accuracy of image classification
Patience	1,2,3	3	Number of epochs to wait to adjust learning rate if monitored value does not improve
Stop patience	3	3	Number of epochs to wait before stopping training if monitored value does not improve
Threshold	0.2, 0.4, 0.5, 0.7, 0.9	0.9	The minimum improvement threshold considered significant for early stopping based on validation loss
Loss scaling factor	0.1, 0.15, 0.2	0.2	The factor by which the learning rate is reduced when a metric has stopped improving during training
Learning rate	0.1,0.001, 0.0001, 0.00001	0.001	The rate at which the model's parameters are adjusted during training
Backpropagation optimiser	Adam	Adam	The method used to update the model's parameters based on performance feedback
Dropout	0.4, 0.45, 0.5, 0.6	0.5	The fraction of input units randomly ignored during training to prevent overfitting.

Figure 6 Block diagram of proposed EfficientNetB3 model (see online version for colours)

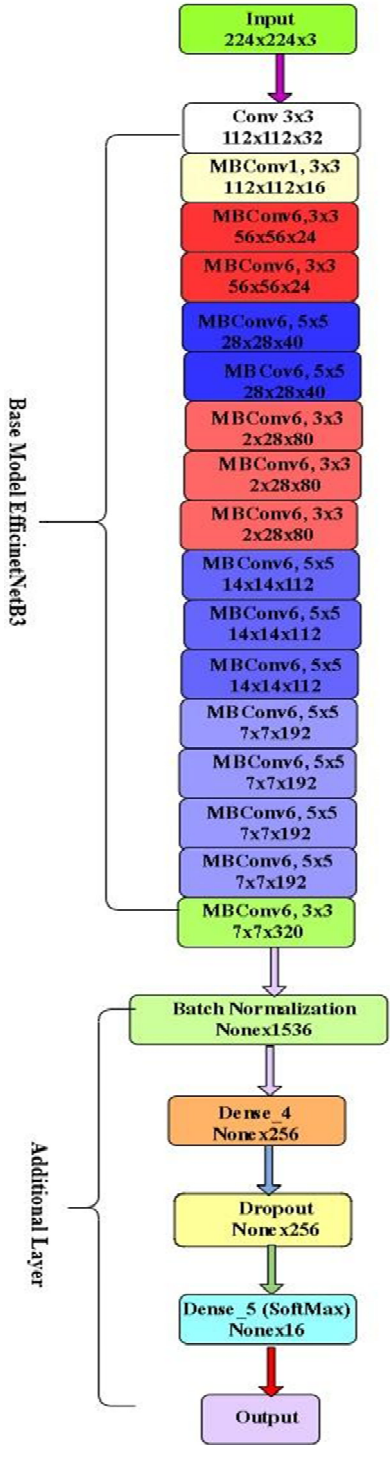
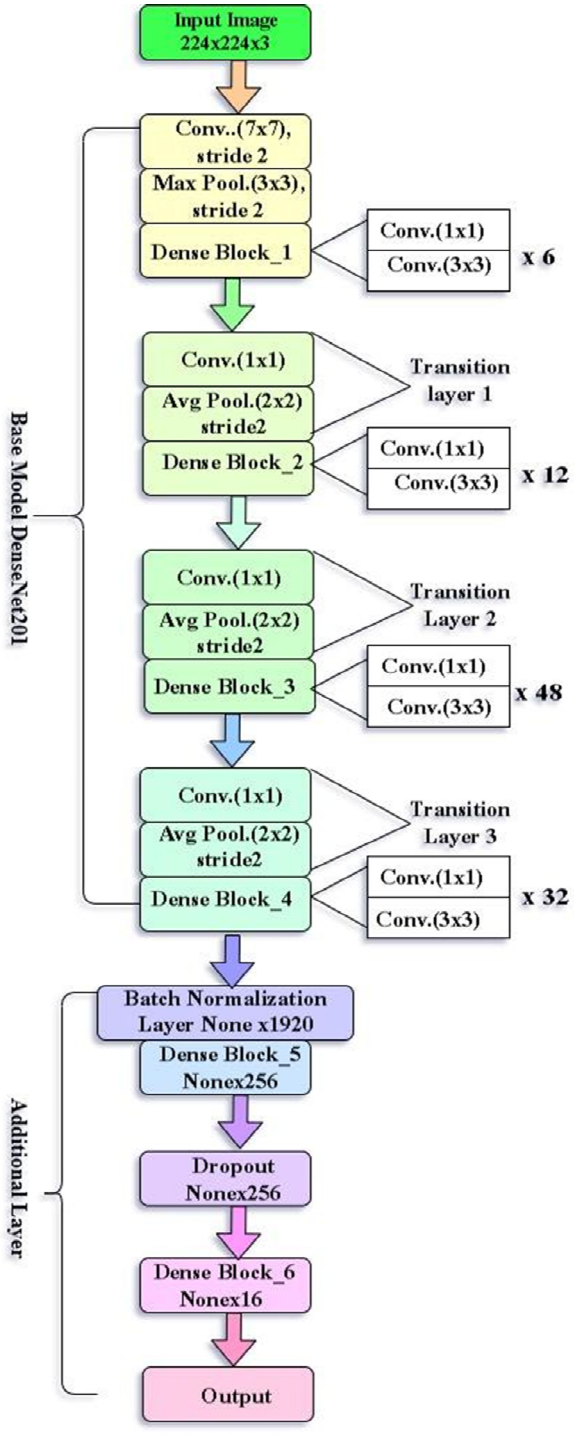


Figure 7 Block diagram of proposed Densenet201 model (see online version for colours)



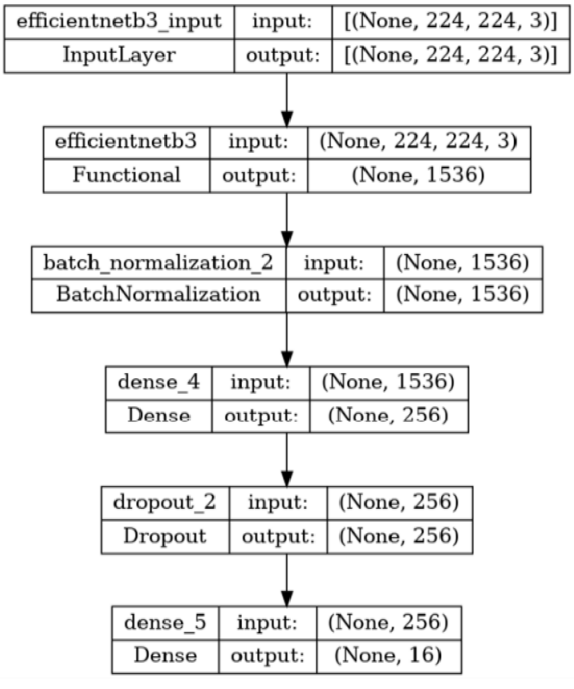
3.5.1 EfficientNetB3 model architecture

The modified EfficientNetB3 model architecture is shown in Figure 6 and Figure 8 and the summary provided is shown in Figure 10. It utilises the pre-trained EfficientNetB3 as the base model to extract relevant features from input images. The model incorporates a batch normalisation layer for activation normalisation and a dense layer for feature mapping. A dropout layer is added to prevent overfitting. The final dense layer generates predicted probabilities for each class using Softmax activation. The model has a total of 11,187,263 parameters, with 11,096,888 trainable parameters. This architecture combines transfer learning with additional trainable layers, making it well-suited for document image classification tasks.

3.5.2 DenseNet201 model architecture:

The modified DenseNet201 model architecture is shown in Figure 7 and Figure 9 and the summary provided in Figure 11. It employs the pre-trained DenseNet201 as the base model for extracting relevant features from input images. A batch normalisation layer is incorporated to normalise the activations, and a dense layer is utilised for feature mapping. To prevent overfitting, a dropout layer is included. The final dense layer employs Softmax activation to generate predicted probabilities for each class. The model comprises a total of 18,825,552 parameters, with 18,592,656 of them being trainable. The remaining 232,896 parameters are non-trainable. This architecture effectively combines transfer learning with trainable layers, making it suitable for classifying document images.

Figure 8 Architecture of proposed EfficientNetB3 model



A simplified flow diagram for transfer learning on document image datasets using well-defined pre-trained models is shown in Figure 12. The step-by-step working of both models are described in Algorithm 2 and Figure 12.

Figure 9 Architecture of proposed DenseNet201 model

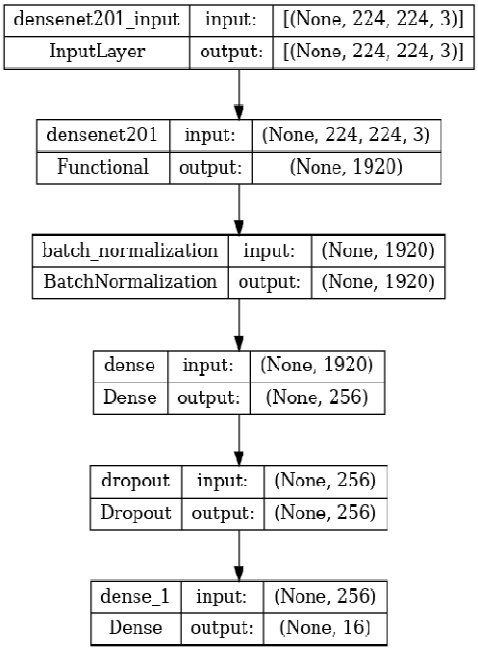


Figure 10 Proposed EfficientNetB3 modes summary

Model: sequential_2

Later(type)	Output Shape	Param #
efficentnetb3 (Functional)	(None, 1536)	10783535
batch_normalisation_2 (BatchNoramlisation)	None, 1536	6144
dense_4 (Dense)	(None, 256)	393472
dropout_2 (Dropout)	(None, 256)	0
dense_5 (Dense)	(None,16)	4112
Total params:	11,187.263	
Trainable params:	11,096,888	
Non-trainable params:	90,375	

Figure 11 Proposed Densenet201 model summary

Model: sequential

Later(type)	Output Shape	Param #
Densenet201 (Functional)	(None, 1920)	18321984
batch_normalisation (BatchNormalisation)	None, 1920	76780
Dense (Dense)	(None, 256)	49176
dropout (Dropout)	(None, 256)	0
dense_1 (Dense)	(None, 16)	4112
Total params:	18,825,552	
Trainable params:	18,592,656	
Non-trainable params:	232,896	

Algorithm 2 Working of the modified models

Input: Image dataset along with the CSV File

Output: Classification of the image based document into 16 classes

1 *Import the necessary libraries: numpy, TensorFlow, and Keras.*

2 *Preprocess the data (e.g., normalise, resize, etc.)*

3 *Perform labelling of dataset according to id and label as per Algorithm1.*

4 *Perform data splitting on the labelled data as train ,test and validation.*

5 *Load the training, validation and testing data.*

6 *Define the architecture of the model:*

- a *Create a Sequential model.*
- b *Add the base model (EfficientnetB3 and DenseNet201) convolutional base as a functional layer with the following configuration: include_top=False, weights='imagenet', pooling='avg'.*
- c *Add a BatchNormalization layer.*
- d *Add a Dense layer with 256 neurons and a ReLU activation function.*
- e *Add a Dropout layer with a dropout rate of 0.5.*
- f *Add a Dense layer with 16 neurons and a softmax activation function.*

7 *Compile the model:*

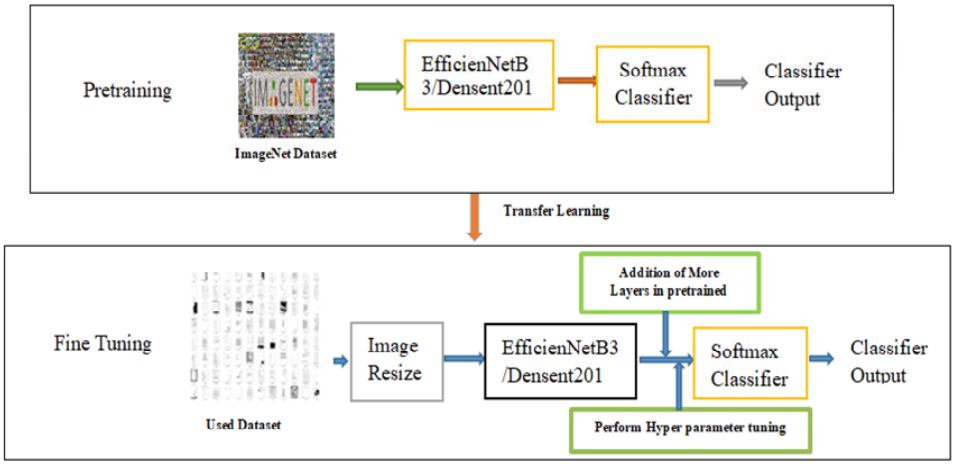
- a *Specify the optimiser (e.g., Adam) and the learning rate.*
- b *Specify the loss function (e.g., categorical_crossentropy for multi-class classification).*
- c *Specify the evaluation metric (e.g., accuracy).*

8 *Train the model:*

- a *Specify the number of epochs and the batch size.*
- b *Pass the training data and labels to the model's fit () function.*

9 *Evaluate the model:*
   a *Pass the validation data and labels to the model's evaluate () function.*
   b *Retrieve and store the loss and accuracy values.*
10 *Make predictions:*
   a *Pass the testing data to the model's predict () function to obtain predictions for each sample.*
11 *Display or utilise the obtained results (e.g., accuracy, predictions, etc.).*

Figure 12 Transfer learning on document image datasets using well-defined pre-trained models (see online version for colours)



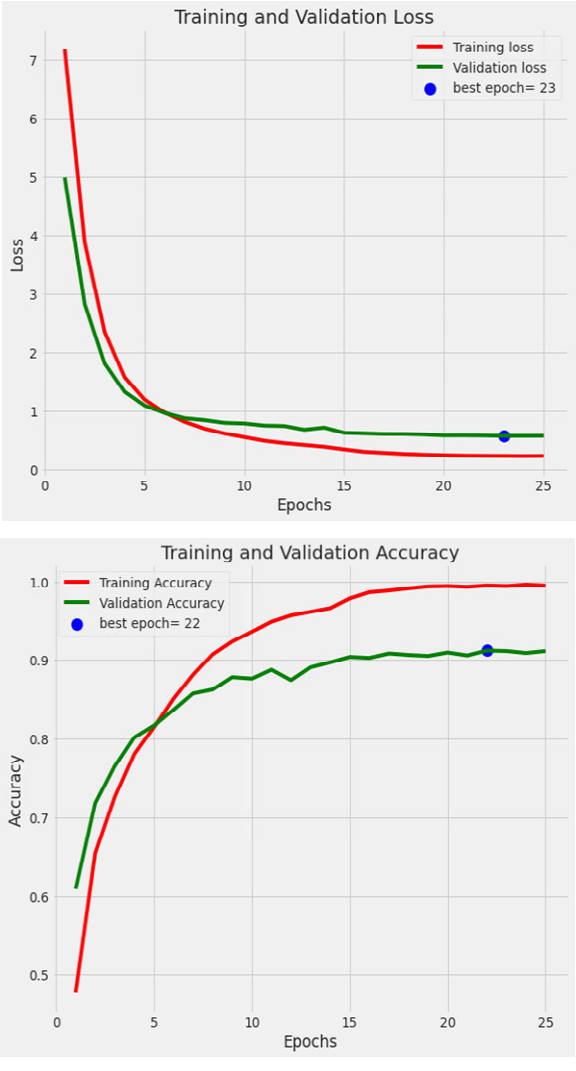
4 Results and discussion

The results of the modified EfficientNetB3 and DenseNet201 models are presented in the subsequent sections:

4.1 Loss and Accuracy curves of modified EfficientNetB3 model

The modified and fine-tuned EfficientNetB3 model was evaluated using cross-entropy loss and accuracy curves, as shown in Figure 13. Through continuous monitoring of training and validation losses, the model achieved its best performance at epoch 23, where the validation loss was minimised. This indicates the model's effectiveness in reducing the difference between predicted and actual labels. The model attained its highest validation accuracy at epoch 22, showcasing its ability to classify document images accurately. As the training and validation accuracy curves consistently increased, the model demonstrated its capability to improve its predictions over successive training epochs. The convergence of the loss curves suggests that the model effectively learned from the training data, while the increasing accuracy curves highlight its proficiency in classifying document images correctly.

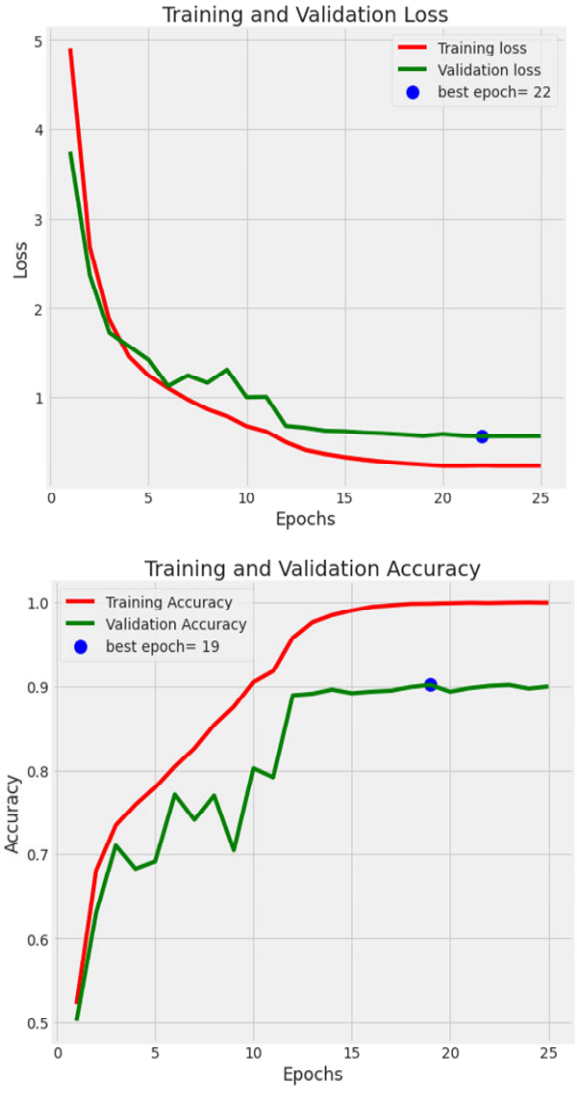
Figure 13 Loss and accuracy of EfficientNetB3 (see online version for colours)



4.2 Loss and accuracy curves of modified DenseNet201 model

The cross-entropy loss and accuracy curves presented in Figure 14 demonstrate the performance evaluation of the modified and fine-tuned DenseNet201 model. The model was trained and validated over 25 epochs, with epoch 22 identified as the best epoch based on the minimised validation loss. This indicates the model's proficiency in reducing the discrepancy between predicted and actual labels. Additionally, the model achieved its highest validation accuracy at epoch 19, showcasing its ability to classify document images accurately. The convergence of the loss and increasing accuracy curves further signify the model's effective learning from the training data and its capability to make precise predictions.

Figure 14 Loss and accuracy of DenseNet201 (see online version for colours)



The EfficientNetB3 and DenseNet201 models were trained for 25 epochs, and during training, metrics such as training accuracy, training loss, validation accuracy, and validation loss were monitored. The results indicated that the EfficientNetB3 model outperformed the DenseNet201 model. The EfficientNetB3 achieved a training accuracy of 100% and a testing accuracy of 90.97%, while the DenseNet201 achieved a training accuracy of 99.85% and a testing accuracy of 90.58%. These findings are summarised in Table 3. Table has aslo included the readings of accuracy of Base models (EfficentNetB3 and Base DenseNet201). In each case the proposed models has better accuracy than the base models. Inclusion of the new layers along with the hypertunning of both models has shown the good response for classification. The time complexity per epoch for the

proposed model are presented in Table 4, providing insights into the computational efficiency of the model.

Table 3 Accuracy of modified EfficientNetB3 and DenseNet201 models in comparison to baseline models

Model	*Accuracy type*	*Percentage*
Base EfficientNetB3	Training accuracy	99.53
	Validation accuracy	78.92
	Testing accuracy	77.71
Base DenseNet201	Training accuracy	99.95
	Validation accuracy	80.14
	Testing accuracy	77.85
Modified model (EfficientNetB3)	Training accuracy	100
	Validation accuracy	90.32
	Testing accuracy	90.97
Modified model (DenseNet201)	Training accuracy	99.85
	Validation accuracy	91.22
	Testing accuracy	90.58

4.3 Evaluation of performance metrics for the modified EfficientNetB3 and DenseNet201 models

Furthermore, the performance of both models has been visualised using a confusion matrix for the test dataset, as depicted in Figure 15. The confusion matrix helps to analyse the accuracy of the models in classifying different document categories. These results demonstrated that the modified EfficientNetB3 and DenseNet201 models achieved state-of-the-art performance in document image classification on the given dataset.

On the other hand, the confusion matrix in Figure 15 provides a detailed breakdown of the model's predictions and the actual labels across classes. It highlights the model's strengths in correctly classifying instances, such as class 6, 8 and 15, where the diagonal elements show high numbers of correctly classified instances. Even class 11 stands out with some difficulty in accurate classification. By considering both figures, we gain a holistic understanding of the model's performance, identifying classes where it excels and areas that require further refinement.

As depicted in Table 5, the classification report comprehensively evaluates the modified EfficientNetB3 and DenseNet201 models' performance across different classes in document image classification. It showcases essential metrics such as precision, recall, and F1-score, which assess the model's ability to classify instances accurately. Notably, class 6 scientific publication and class 15 Memo have a precision of 97%, class eight file folder has a recall of 100%, and F1-Score 98% exhibiting high precision, recall, and F1-scores, indicating good performance in correctly predicting positive instances using EfficientNetB3 model. In contrast, the class 8 file folder having a precision of 99%, recall of 99%, and F1-score of 99% exhibits high precision, recall, and F1-scores, indicating strong performance in correctly predicting positive instances using the DenseNet201

model. However, class 11 displays lower scores, suggesting room for improvement in its classification using the modified EfficientNetB3 and DenseNet201 models.

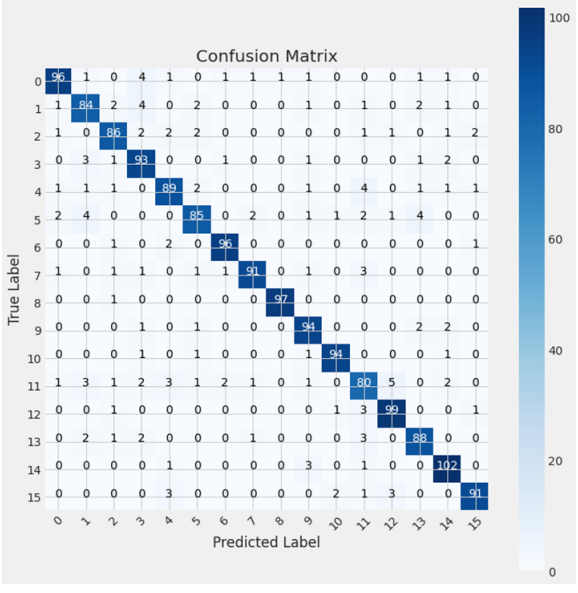
Table 4 Time complexity of modified EfficientNetB3 and DenseNet201 models

Epoch number	*Elapsed time in second for EfficientNetB3 model*	*Elapsed time in second for DenseNet201 model*
1	139.88	191.09
2	92.96	96.41
3	92.99	97.63
4	92.65	97.28
5	92.63	96.62
6	92.42	97.16
7	92.46	96.40
8	92.66	96.90
9	92.22	97.18
10	92.89	95.53
11	92.79	96.48
12	92.95	96.38
13	92.80	95.78
14	93.27	98.17
15	93.26	97.71
16	93.71	97.02
17	93.50	96.56
18	93.31	96.72
19	92.49	97.14
20	92.97	96.80
21	93.16	97.69
22	92.88	96.79
23	93.00	95.60
24	92.97	94.63
25	92.94	95.95
Total training time	0.0 hours, 39.0 minutes, 35.61 seconds	0.0 hours, 41.0 minutes, 59.40 seconds

The weighted average precision, recall, and F1-score of 0.91 demonstrate the models' solid overall performance in classifying document images. This evaluation provides insights into the models' strengths and areas for improvement, allowing for targeted refinements to enhance its performance across all classes.

Table 5 shows the comparative performance of each class of document with the baseline models. For each of the 16 classes the proposed modified model has shown the improvement in average weighted value of performance parameters. In every case, it is found to be 91%. While looking individually at each class, only the resume class or document type has not surpassed the precision value results of the modified EfficientNetB3 model. In case of baseline model, precision value is 91% and modified model precision is 88%.

Figure 15 Confusion matrices for document classification on the test dataset using (a) DenseNet201 and (b) EfficientNetB3 (see online version for colours)



To do the comparison of performance parameter of baseline models with the modified models, study has set the following hypothesis:

H0 Not significantly increase in the performance of modified model as compare to the baseline model, i.e., $\mu 1 > \mu\ 2$.

H1 Significance increase in the performance of modified model as compare to the baseline model, i.e., $\mu 1 < \mu 2$.

Table 5 Performance parameter comparison with baseline and proposed model the document type's classification on the RVL-CDIP dataset obtained with the proposed EfficentNetB3 and proposed DenseNet201

Class	*Document type*	*Base EfficientNetB3*			*Base DenseNet201*			*Modified EfficientNetB3*			*Modified DenseNet201*		
		Precision (%)	*Recall (%)*	*F1-score (%)*	*Precision (%)*	*Recall (%)*	*F1-score (%)*	*Precision (%)*	*Recall (%)*	*F1-score (%)*	*Precision (%)*	*Recall (%)*	*F1-score (%)*
0	Letter	76	68	72	77	67	71	90	90	90	93	89	91
1	Form	62	63	63	58	65	61	91	88	80	86	86	86
2	E-mail	69	67	68	72	69	70	90	85	87	90	88	89
3	Handwritten	71	77	74	77	77	77	88	91	89	85	91	88
4	Advertisement	69	76	72	64	75	69	87	91	89	88	87	88
5	Scientific report	71	62	67	69	62	66	91	80	85	89	83	86
6	Scientific publication	85	87	86	86	88	87	97	95	96	95	96	96
7	Specification	78	79	79	80	81	80	89	94	91	95	91	93
8	File folder	95	97	96	95	97	96	95	100	98	99	99	99
9	News article	85	90	88	81	90	85	95	95	95	90	94	92
10	Budget	81	87	84	84	89	86	95	96	95	96	96	96
11	Invoice	53	54	53	65	57	61	81	97	79	81	78	80
12	Presentation	86	87	87	87	86	87	93	95	94	91	94	93
13	Questionnaire	85	80	82	77	79	78	91	93	92	89	91	90
14	Resume	91	93	92	86	91	88	88	93	90	90	95	93
15	Memo	85	76	80	89	74	81	97	92	94	95	91	93
Weighted avg.		78	78	78	78	78	78	91	91	91	91	91	91
p-value*								0.8×10^{-4}	1.28×10^{-4}	4.33×10^{-4}	0.75×10^{-4}	4.81×10^{-4}	0.71×10^{-4}
t-statistics*								–4.52	–4.3	–3.94	–4.58	–3.91	–4.6

Note: p-value and t-statistics is calculated with 95% of confidence interval.

Table 5 provides a comparative analysis of the evaluation parameters – precision, recall, and F1 score – before and after modifications for each of the two base models. The last row of the table demonstrates a significant improvement in the performance of the modified model compared to the baseline model, as confirmed by the t-test. Consequently, the null hypothesis ($\mu 1 > \mu 2$) is rejected at the 95% confidence interval, and the alternative hypothesis is accepted.

5 Critical analysis and comparative performance

5.1 Comparison with existing works

According to the comparison in Table 6, the proposed models, DenseNet201 and EfficientNetB3, demonstrated superior performance compared to the existing models for image classification on the RVL-CDIP dataset. The existing models achieved accuracy scores ranging from 88.6% to 90.4%. However, modified EfficientNetB3 and DenseNet201 models achieved an accuracy of 90.97% and 90.59% respectively, surpassing all the existing techniques. These results indicate that the modified models exhibit enhanced classification capabilities and are more effective at accurately identifying the various classes within the dataset than the existing models. As a result, the modified models offer notable advancements and improved performance in image classification tasks on the RVL-CDIP dataset. Due to scarcity of resources current study has worked on small dataset as compared to other studies but able to surpass the performance as compared to other studies as shown in Table 5.

Table 6 Performance metrics comparison of Models on RVL-CDIP dataset

References	*No. of classes/ dataset*	*Used techniques*	*Accuracy (%)*	*Precision*	*Recall*	*F1-score*
Afzal et al. (2017)	16 RVL-CDIP	AlexNet	88.6	X	X	X
		GoogleNet	89.02	X	X	X
		ResNet-50	90.40	X	X	X
Kanchi et al. (2022)	16 RVL-CDIP	BERT& EfficientNet	90.3	X	X	X
Audebert et al. (2020)	16 RVL-CDIP	MobileNetV2	89.1	X	X	X
Siddiqui et al. (2021)	16 RVL-CDIP	RESNet-50	89.09	X	X	X
Ferrando et al. (2020)	16 RVL-CDIP	EfficientNet, BERT	89.47	X	X	X
Harley et al. (2015)	16 RVL-CDIP	Ensemble of CNN	89.8	X	X	X
Modified DenseNet201	*16 RVL-CDIP*	*DenseNet201*	*90.59*	*91*	*91*	*91*
Modified EfficientNetB3	*16 RVL-CDIP*	*EfficienNetB3*	*90.97*	*91*	*91*	*91*

It is crucial to note that the enhanced models exhibit marginal improvement, potentially attributed to the semi-automatic labelling approach. In contrast, all other cases presented

in Table 6 involve manual or pre-labelled labelling. With enhanced hardware resources, training can be optimised, leading to further enhancements in performance parameters.

5.2 Implication of research

The research conducted on document image classification using CNNs has significant implications for the field of document management and classification systems. By leveraging the power of CNNs and exploring different architectural variations like ResNet, DenseNet, and EfficientNet, remarkable progress has been made in document image classification. This study demonstrates that hypertuned CNN-based approaches can greatly enhance the overall accuracy and efficiency of document image processing and classification. Furthermore, the research demonstrates the adaptability of CNNs in handling diverse document layouts, which is a major challenge in the field. By successfully applying CNNs to various formats, these proposed models exhibit the capability to comprehend and extract information from different layout structures. This opens up avenues for practical applications in areas such as document management, information retrieval from image documents, and document classification in legal proceedings and business houses. The integration of computer vision and NLP techniques is crucial in achieving comprehensive document understanding in the current study. The research also provides insights for future directions, including further optimisation of CNN architectures, exploration of hybrid models, and handling multi-modal documents. Overall, the research highlights the potential of CNN-based approaches in revolutionising document image processing and lays the foundation for advancements in the field. Many organisations in the field of human resource management, legal firms, financial institutions, health care providers are adopting this technology. They are also looking for the enhancement in this field

5.3 Threats to validity

The research findings on document image processing using CNNs may have some limitations that affect their reliability. One such limitation is that the results may not apply well to different datasets and real-world situations because the research focused on a specific dataset with its own characteristics. Another concern is dataset bias, where the accuracy of CNN models depends heavily on the quality and representativeness of the training data. Careful consideration of hyperparameter settings is necessary because CNN models are sensitive to these choices and may overfit the data, meaning they memorise rather than learn general patterns. Additionally, CNN models can be challenging to interpret, making it difficult to understand and trust their decisions. Acknowledging these limitations helps researchers better evaluate and improve CNN-based approaches for document image processing.

5.4 Novelty

The novelty of this research lies in several aspects. Firstly, it introduces a comprehensive methodology for document image classification using CNNs, examining two different pre-trained models. This evaluation enhances our understanding of their performance in this specific domain. Secondly, advanced techniques like transfer learning and hyperparameter tuning are employed to improve accuracy and generalisation. This

combination of methods adds value by enhancing classification outcomes. Thirdly, the study emphasises the effectiveness of the EfficientNetB3 and DenseNet201 models in accurately classifying document images, highlighting their suitability for this task. Lastly, a detailed evaluation utilising metrics such as precision, recall, and F1-score provides a thorough understanding of model performance across various document classes. This analysis identifies strengths and areas for improvement, guiding future research directions. The model can be applied during the inspection of online applications to classify various document types like matriculation certificate, graduation, post-graduation, etc.

6 Conclusions and future work

In this research article, the IndoML Datathon 2022 dataset underwent a preprocessing phase and was semi-automatically labelled. To classify document images into 16 distinct document categories, the models, EfficientNetB3 and DenseNet201, were subjected to modifications involving the addition of supplementary layers. Subsequently, these models were fine-tuned utilising the IndoML Datathon 2022 dataset. A comprehensive evaluation process was conducted, comparing their performance with pre-existing techniques. The results presented in Table 5 demonstrated that the modified EfficientNetB3 and DenseNet201 models outperformed the existing models. Also, the results of both the models (EfficientNetB3 and DenseNet201) are comparable as they achieved an overall accuracy of 90.97% and 90.59% respectively. The results indicated good performance in correctly predicting most of the classes while remaining classes require further improvement. Further, the study has also recorded the other performance parameters for evaluation like precision, recall and F1 score. In each case, the proposed model outperforms the standard models. One of the major application of the study is with courts and education institutes where there are number of documents are being sent by the public. This study will help the organisation to sort the different types of documents without the manual interaction. This will also speed up the process. Future endeavours involve the enlargement of the dataset, augmenting the class count, enhancing model precision, and implementing real-time deployment within document management systems.

Acknowledgements

Dr. Satwinder Singh and Nakala Srinivas Mudhiraj would like to thanks Indian Council of Medical Research (ICMR, New Delhi) for providing computational facility for computational work of the study under the project 2021-6329.

Dataset availability statement

This dataset is freely accessible to the public at [https://www.kaggle.com/competitions/datathonindoml-2022/data] and is made available specifically for research and development purposes in the field of document image classification. Researchers and

practitioners can utilise this dataset to advance their work and contribute to the progress of document image classification techniques.

References

Afzal, M.Z., Kolsch, A., Ahmed, S. and Liwicki, M. (2017) 'Cutting the error by half: investigation of very deep CNN and advanced training strategies for document image classification', *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, Vol. 1, pp.883–888.

Ali Reshi, J. and Singh, S. (2018) 'Investigating the role of code smells in preventive maintenance', *Journal of Information Technology Management*, Vol. 10, No. 4, pp.41–63.

Alkhonin, A., Almutairi, A., Alburaidi, A. and Saudagar, A.K.J. (2020) 'Recognition of flowers using convolutional neural networks', *International Journal of Intelligent Engineering Informatics*, Vol. 8, No. 3, pp.186–197.

Aly, T. and Nguyen-an, K. (2018) 'Document layout analysis : a maximum homogeneous region approach', *1st International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, Ho Chi Minh City, Vietnam, pp.1–5.

Anter, A.M., Azar, A.T., Hassanien, A.E., El-Bendary, N. and Elsoud, M.A. (2013) 'Automatic computer aided segmentation for liver and hepatic lesions using hybrid segmentations techniques', *2013 Federated Conference on Computer Science and Information Systems, FedCSIS 2013*, pp.193–198.

Appalaraju, S., Jasani, B., Kota, B.U., Xie, Y. and Manmatha, R. (2021) 'DocFormer: end-to-end transformer for document understanding', *Proceedings of the IEEE International Conference on Computer Vision*, pp.973–983.

Audebert, N., Herold, C., Slimani, K. and Vidal, C. (2020) 'Multimodal deep networks for text and image-based document classification', *Communications in Computer and Information Science, CCIS*, Vol. 1167, pp.427–443.

Aziz, A.S.A., Azar, A.T., Salama, M.A., Hassanien, A.E. and Hanafy, S.E.O. (2013) 'Genetic algorithm with different feature selection techniques for anomaly detectors generation', *2013 Federated Conference on Computer Science and Information Systems, FedCSIS 2013*, pp.769–774.

Bakkali, S., Ming, Z., Coustaty, M. and Rusinol, M. (2020) 'Visual and textual deep feature fusion for document image classification', *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June, pp.2394–2403.

Bakkali, S., Ming, Z., Coustaty, M. and Rusiñol, M. (2021) 'EAML: ensemble self-attention-based mutual learning network for document image classification', *International Journal on Document Analysis and Recognition*, Vol. 24, No. 3, pp.251–268, Springer, Berlin, Heidelberg.

Bakkali, S., Ming, Z., Coustaty, M., Rusiñol, M. and Terrades, O.R. (2023) 'VLCDoC: vision-language contrastive pre-training model for cross-modal document classification', *Pattern Recognition*, July, Vol. 139, p.109419.

Baygin, M. (2019) 'Classification of text documents based on naive Bayes using N-gram features', *2018 International Conference on Artificial Intelligence and Data Processing, IDAP 2018*, IEEE pp.1–5.

Bhowmik, S., Kundu, S. and Sarkar, R. (2021) 'BINYAS : a complex document layout analysis system', *Multimedia Tools and Applications*, Vol. 80, No. 6, pp.8471–8504.

Chelliah, B.J., Harshitha, K. and Pandey, S. (2023a) 'Adaptive and effective spatio-temporal modelling for offensive video classification using deep neural network', *International Journal of Intelligent Engineering Informatics. Inderscience Publishers (IEL)*, Vol. 11, No. 1, pp.19-34.

Chelliah, B.J., Malik, M.M., Kumar, A., Singh, N. and Regin, R. (2023b) 'Similarity-based optimised and adaptive adversarial attack on image classification using neural network', *International Journal of Intelligent Engineering Informatics*, Vol. 11, No. 1, pp.71–95, Inderscience Publishers (IEL).

Das, A., Roy, S., Bhattacharya, U. and Parui, S.K. (2018) 'Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks', *Proceedings – International Conference on Pattern Recognition*, IEEE, August, pp.3180–3185.

Emary, E., Zawbaa, H.M., Hassanien, A.E., Schaefer, G. and Azar, A.T. (2014a) 'Retinal blood vessel segmentation using bee colony optimisation and pattern search', *Proceedings of the International Joint Conference on Neural Networks*, Vol. 1, No. 3, pp.1001–1006.

Emary, E., Zawbaa, H.M., Hassanien, A.E., Schaefer, G. and Azar, A.T. (2014b) 'Retinal vessel segmentation based on possibilistic fuzzy c-means clustering optimised with cuckoo search', *Proceedings of the International Joint Conference on Neural Networks*, pp.1792–1796.

Engin, D., Emekligil, E., Akpınar, M.Y. and Oral, B. (2019) 'Multimodal deep neural networks for banking document classification', *The Ninth International Conference on Advances in Information Mining and Management*, No. C, pp.21–25.

Ferrando, J., Domínguez, J.L., Torres, J., García, R., García, D., Garrido, D., Cortada, J. and Valero, M. (2020) 'Improving accuracy and speeding up document image classification through parallel systems', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 12138, LNCS, pp.387–400.

Harley, A.W., Ufkes, A. and Derpanis, K.G. (2015) 'Evaluation of deep convolutional nets for document image classification and retrieval', *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, November, pp.991–995.

Hassanpour, M. and Malek, H. (2019) 'Document image classification using SqueezeNet convolutional neural network', *5th Iranian Conference on Signal Processing and Intelligent Systems, ICSPIS 2019*, IEEE, December, pp.18–19.

Jaume, G., Kemal Ekenel, H. and Thiran, J-P. (2019) 'FUNSD: a dataset for form understanding in noisy scanned documents', in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, IEEE, Vol. 2, pp.1–6.

Jeyanthi, S., Venkatakrishnaiah, R. and Raju, K.V.B. (2023) 'Utilising recurrent neural network technique for predicting strand settlement on brittle sand and geocell', *International Journal of Intelligent Engineering Informatics*, Vol. 11, No. 2, pp.122–137, Inderscience Publishers (IEL).

Jothi, G., Inbarani, H.H. and Azar, A.T. (2013) 'Hybrid tolerance rough set: PSO based supervised feature selection for digital mammogram images', *International Journal of Fuzzy System Applications*, Vol. 3, No. 4, pp.15–30.

Kanchi, S., Pagani, A., Mokayed, H., Liwicki, M., Stricker, D. and Afzal, M.Z. (2022) 'EmmDocClassifier: efficient multimodal document image classifier for scarce data', *Applied Sciences (Switzerland)*, Vol. 12, No. 3, pp.1457–1474.

Kaur, J. and Singh, S. (2016) 'Neural network based refactoring area identification in software system with object oriented metrics', *Indian Journal of Science and Technology*, Vol. 9, No. 10, pp.1–8.

Khan, S., Thirunavukkarasu, K., Hammad, R., Bali, V. and Qader, M.R. (2021) 'Convolutional neural network based SARS-CoV-2 patients detection model using CT images', *International Journal of Intelligent Engineering Informatics*, Vol. 9, No. 2, pp.211–228.

Khan, T. and Mollah, A.F. (2020) 'Text non-text classification based on area occupancy of equidistant pixels', *Procedia Computer Science*, Vol. 167, pp.1889–1900.

Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D. and Park, S. (2022) 'OCR-free document understanding transformer', in *European Conference on Computer Vision*, pp.498–517, Springer Nature, Switzerland, Cham.

Kolsch, A., Afzal, M.Z., Ebbecke, M. and Liwicki, M. (2018) 'Real-time document image classification using deep CNN and extreme learning machines', *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, Vol. 1, pp.1318–1323.

Li, X., Zheng, Y., Hu, Y., Cao, H., Wu, Y., Jiang, D., Liu, Y. and Ren, B. (2022) 'Relational representation learning in visually-rich documents', in *Proceedings of the 30th ACM International Conference on Multimedia*, pp.4614–4624.

Luo, C., Tang, G., Zheng, Q., Yao, C., Jin, L., Li, C., Xue, Y. and Si, L. (2021) *Bi-VLDoc : Bidirectional Vision-Language Modeling for Visually-Rich Document Understanding*, arXiv preprint, arXiv:2206.13155.

Luz de Araujo, P.H., de Almeida, A.P.G.S., Ataides Braz, F., Correia da Silva, N., de Barros Vidal, F. and de Campos, T.E. (2022) 'Sequence-aware multimodal page classification of Brazilian legal documents', *International Journal on Document Analysis and Recognition*, Vol. 26, No. 1, pp.33–49.

Omurca, S.İ., Ekinci, E., Sevim, S., Edinç, E.B., Eken, S. and Sayar, A. (2022) 'A document image classification system fusing deep and machine learning models', *Applied Intelligence*, Vol. 53, No. 12, pp.15295–15310.

Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A.S. and Staar, P. (2022) *DocLayNet : A Large Human-Annotated Dataset for Document-Layout Analysis* [online] https://github.com/DS4SD/DocLayNet (accessed 1 July 2023).

Qazi, A. and Goudar, R.H. (2018) 'An ontology-based term weighting technique for web document categorization', *Procedia Computer Science*, Vol. 133, pp.75–81, Elsevier BV.

Rabut, B.A., Fajardo, A.C. and Medina, R.P. (2019) 'Multi-class document classification using improved word embeddings', *ACM International Conference Proceeding Series*, October, pp.42–46.

Rasjid, Z.E. and Setiawan, R. (2017) 'Performance comparison and optimization of text document classification using k-NN and naïve Bayes classification techniques', *Procedia Computer Science*, Vol. 116, pp.107–112, Elsevier B.V.

Riba, P., Dutta, A., Goldmann, L., Fornes, A., Ramos, O. and Llados, J. (2019) 'Table detection in invoice documents by graph neural networks', *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, pp.122–127.

Sahare, P. and Dhok, S.B. (2018) 'Multilingual character segmentation and recognition schemes for Indian document images', *IEEE Access*, Vol. 6, pp.10603–10617.

Sarkhel, R. and Nandi, A. (2019) 'Visual segmentation for information extraction from heterogeneous visually rich documents', *Proceedings of the ACM SIGMOD International Conference on Management of Data*, June, pp.247–262.

Selvakumar, A.A. and Thangaraju, P. (2023) 'Efficient de-noising brain MRI images using various filtering techniques', *International Journal of Intelligent Engineering Informatics*', Vol. 11, No. 2, pp.176–190, Inderscience Publishers (IEL).

Seuret, M., Alberti, M., Liwicki, M. and Ingold, R. (2017) 'PCA-initialized deep neural networks applied to document image analysis', *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, Vol. 1, pp.877–882.

Sharma, S. and Kumar, V. (2020) 'Low-level features based 2D face recognition using machine learning', *International Journal of Intelligent Engineering Informatics*, Vol. 8, No. 4, pp.305–330.

Sharma, S. and Singh, S. (2020) 'Texture-based automated classification of Ransomware', *Journal of The Institution of Engineers (India): Series B*, Vol. 102, No. 1, pp.131–142.

Siddiqui, S.A., Dengel, A. and Ahmed, S. (2021) 'Self-supervised representation learning for document image classification', *IEEE Access*, IEEE, Vol. 9, pp.164358–164367.

Singh, S. and Singla, R. (2017) 'Classification of defective modules using object-oriented metrics', *International Journal of Intelligent Systems Technologies and Applications*, Vol. 16, No. 1, pp.1–13.

Suganthi, M. and Sathiaseelan, J.G.R. (2022) 'A novel feature extraction method for identifying quality seed selection', *International Journal of Intelligent Engineering Informatics*, Vol. 10, No. 5, pp.359–378, Inderscience Publishers (IEL).

Thiruvenkatasuresh, M.P. and Venkatachalam, V. (2019) 'Analysis and evaluation of classification and segmentation of brain tumour images', *International Journal of Biomedical Engineering and Technology*, Vol. 30, No. 2, pp.153–178.

Tran, T.A., Nguyen-An, K. and Quang Vo, N. (2018) 'Document layout analysis: a maximum homogeneous region approach', *1st International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, Ho Chi Minh City, Vietnam, pp.1–5.

Vu, H.M. and Nguyen, D.T.N. (2020) *Revising Funds Dataset for Key-Value Detection in Document Images*, arXiv preprint arXiv:2010.05322.

Wang, D., Mao, K. and Ng, G.W. (2017) 'Convolutional neural networks and multimodal fusion for text aided image classification', *20th International Conference on Information Fusion, Fusion 2017*, Xi'an, China, pp.1–7.

Xu, Y., Yin, F., Zhang, Z. and Liu, C.L. (2018) 'Multi-task layout analysis for historical handwritten documents using fully convolutional networks', *IJCAI International Joint Conference on Artificial Intelligence*, July, pp.1057–1063.