**International Journal of Computer Engineering and Technology (IJCET)**
Volume 15, Issue 4, July-Aug 2024, pp. 29-38, Article ID: IJCET_15_04_003
Available online at https://iaeme.com/Home/issue/IJCET?Volume=15&Issue=3
ISSN Print: 0976-6367 and ISSN Online: 0976-6375
Impact Factor (2024): 18.59 (Based on Google Scholar Citation)



# DIGITIZED DOCUMENT CLASSIFICATION USING MACHINE AND DEEP LEARNING A SURVEY

**A. Vijaya Mahendra Varman*, AN Sigappi**
Department of Computer Science and Engineering,
Annamalai University, Annamalainagar-561203, India

## ABSTRACT

*The digital transformation has led to a widespread application of artificial intelligence (AI) technology to overcome human-induced errors in a range of systems utilized in our daily lives. The World Wide Web's rapid expansion has made it impossible for humans to classify information, which has sparked the development of methods like data mining, natural language processing, and machine learning for the automatic classification of textual documents. Due to the abundance of information available from many sources, classification jobs have become increasingly important. One important way to handle and process a large number of digital documents is through automated text classification. This essay offers an understanding of the steps involved in text classification as well as different classifiers. Additionally, it seeks to evaluate and contrast different classifiers that are currently accessible based on a few parameters, including performance and time complexity.*

**Keywords: -** Data Mining, Natural Language Processing, Classifier, Text classification, Machine Learning and Deep Learning.

**Cite this Article:** A.Vijaya Mahendra Varman and AN Sigappi, Digitized Document Classification Using Machine and Deep Learning A Survey, *International Journal of Computer Engineering and Technology (IJCET)*, 15(4), 2024, pp. 29-38.
https://iaeme.com/MasterAdmin/Journal_uploads/IJCET/VOLUME_15_ISSUE_4/IJCET_15_04_003.pdf

## 1. INTRODUCTION

A vast amount of quickly accessible information can be found in millions of electronic papers that are already available. These documents are getting more and more readily available on a daily basis. Documents must be sorted into distinct categories, which often calls for human user input. Using a machine learning system, the aim is to automatically classify papers into groups according to their content. The paper aims to introduce several methods, enumerate their benefits and drawbacks, and recommend appropriate strategies to address the issue.

ML uses three learning methods: supervised learning, untrained learning, and supervised learning. ML methods involve feature extraction and ask domain experts to select appropriate features for a given problem. Deep learning (DL) technology solves feature selection problems. DL is a subset of ML, DL can automatically extract the most important features from the raw input data. [4]. Imaging technology is used to capture abnormalities in the human body. The captured images are essential for the diagnosis, prognosis and treatment planning of serious diseases. Understanding the medical condition is done by trained medical professionals. However, the unavailability of human experts and associated fatigue and estimation methods reduce the effectiveness of image recognition performed by trained medical professionals. CNNs, or convolutional neural networks, are useful tools for comprehending images[5].DL has two characteristics: [1] learning of different features of the data at multiple levels of abstraction by multiple processing layers; and [2] learning of feature presentations on each layer, either supervised or unsupervised. The capabilities of advanced DL Architectures in the medical fields of MRI [6], Radiology [7], Cardiology [8], and Neurology [9] have been highlighted in a number of recent review papers.

## 2. RELATED WORK

In the field of document analysis and recognition, document image classification has a long history, and as the world becomes more digitally connected, its significance only increases. Extracting textual data from document images is necessary to accomplish efficient indexing, retrieval, organization, and analysis. Document image classification is required for these tasks.

Harley et al. [8] used CNN-extracted features instead of hand-crafted features due to the popularity and robustness of CNN. To learn document structures, Roy et al. [9] proposed an ensemble model. In their model, DCNN models trained overall or specific regions of the documents, and the output of the models was ensembled with a Support Vector Machine (SVM). Csurka [10] extracted shallow features from document images with Sunlight Histogram (RL) and Fisher Vector (FV) and then used classical machine learning algorithms to classify these features. CNNs trained on ImageNet were applied with pre-processing, data augmentation, transfer learning, and domain adaptation to classify document images [11]

S¸ahin et al. [12] utilized only text to predict the classes of the document images. For this purpose, at first keywords for each class were extracted then words in OCRed Turkish documents were matched with class keywords. At the end, classes were determined according to the number of matches. Yaman et al. [13] used fine-tuned deep CNN architectures namely VGG-16, AlexNet, and GoogLeNet to classify document images by using visual features. VGG-16 yielded results as the most successful method.

Noce et al. [14] combined text and imagebased approaches to improve classification accuracy. In their study, the structural embedding of textual topics obtained with OCR were represented with colored boxes and fed into the CNN for classification tasks. In Audebert et al.'s study [15], document images were classified by fusing textual and visual features based on an end-toend learnable multimodal deep network model. While textual features were represented with FastText generated document embeddings, visual features were learned by using MobileNetv2, a state-of-the-art CNN architecture

## 3. DOCUMENT CLASSIFICATION

Document classification is considered an important task that triggers various document image processing pipelines, such as text recognition, information extraction, and document retrieval. By first classifying documents into various predetermined groups, the document processing system gains overall efficiency, time is saved, and a diverse range of task can be accomplished. In the field of literature, there are three main types of methods for classifying document images.

The initial group of methods relies on the structure and arrangement of the document images. This approach requires a significant amount of time to identify the key elements of the documents and apply them for categorization. The second type of method concentrates on developing either local or global descriptions of the images. The objective of these descriptors is to classify documents. This process is time-consuming, involving the extraction of both local and global features. In the third category of methods, CNN [14] is employed to automatically extract features from document images, which are then classified. Utilizing CNN, the most widely used neural network model for image classification, simplifies the task of distinguishing images without the need for the extraction of manually crafted elements [6]. The presence of high interclass similarities in document images, along with variations in shape, poses significant challenges in the development of an effective document classification identification algorithm.

### 3.1. Datasets

The process of document classification commences with the collection of diverse document types, encompassing various formats such as HTML, .pdf, .doc, web content, among others.

### 3.2. Pre-processing and segmentation

The initial phase of classification involves several modifications to the image prior to its processing. The RGB image is initially transformed into a grayscale version, which is then converted and stored in binary format. Fundamentally, after the aforementioned transformations, the original image is converted into new pixels, which are calculated and divided using binary equations. This primary objective of this transformation is to enhance the visibility of significant features within the image, thereby facilitating subsequent processing steps.

### 3.3. Feature Point Extraction

The method of transforming raw data derived from images into numerical features for classification purposes is referred to as feature extraction. In contrast to the direct application of machine learning on the unprocessed data, the data obtained through feature extraction yields enhanced results. The process of manual feature extraction necessitates the identification and detailed description of pertinent characteristics relevant to a specific issue. A comprehensive grasp of the context or domain under consideration can facilitate informed decisions regarding the potential usefulness of various features.

Upon segmenting the hand area, the contour of the hand was subsequently extracted. Following the identification of the region of interest, the features were extracted. In the course of these studies, an array of feature extraction methods were utilized to generate patterns or descriptions of each group, which were subsequently fed into the classifier for the purpose of training. A detailed examination of the most commonly employed feature extraction methods in this context is presented below.

## 4. NATURAL LANGUAGE PROCESSING

The feature extraction and reduction phases of the text classification process are executed through the application of Natural Language Processing techniques. Within these phases, linguistic features can be extracted from texts and subsequently utilized as components of their feature vectors[17]. For instance, certain elements of the text, including direct speech, various declinations, the length of sentences, and the distribution of different components within sentences (such as noun phrases, prepositional phrases, or verb phrases), can all be identified and utilized as feature vectors. These elements can also complement or enhance the use of word frequency vectors. [18].

## 5. CLASSIFICATION

Classification constitutes the method by which individual entities, including patterns, pixels, or images, are organized into categories according to their similarity to the defined characteristics of the group. Presently, there is a significant increase in interest among researchers regarding the automation of document classification within predetermined categories. The three primary methods employed for this purpose are supervised, unsupervised, and semi-supervised classification techniques. In the preceding decade, remarkable and swift advancements have been observed in this domain, encompassing various machine learning methodologies. These include, but are not limited to, Bayesian classification, Decision Tree, K-Nearest Neighbors (K-NN), Support Vector Machines (SVMs), Convolutional Neural Networks, and ResNet-50.

### 5.1. K-Nearest Neighbour

K-Nearest Neighbors (K-NN) constitutes a sophisticated supervised machine learning algorithm that systematically stores all available data cases and categorizes new instances based on a similarity criterion, which could be distance functions among others. K-NN operates under the principle that data points (documents) that are proximate in space are likely to belong to the same class. The algorithm assimilates the training dataset and then predicts the response for a new instance by considering a predetermined number of its nearest neighbors. This is achieved by employing a specific similarity metric, such as Euclidean distance. The formula for calculating the Euclidean distance between two data points is utilized.

$$Dist(X,Y)=\sqrt{(X_i-Y_i)}^2$$

A significant drawback of the similarity metric employed in k-Nearest Neighbors (k-NN) lies in its utilization of all available features in calculating distances, which adversely affects its performance. Across numerous document data sets, only a subset of the total vocabulary is likely to be beneficial for categorizing documents. An anticipated strategy to address this issue could involve assigning weights to various features (or words within the document data, for instance)[19]. The Weight Adjusted k-Nearest Neighbor (WAKNN) classification algorithm is predicated on the k-Nearest Neighbors (k-NN) classification paradigm, and is designed to augment the efficacy of text classification systems[20].

### 5.2. Support Vector Machine

Initially, Support Vector Machines (SVM) were primarily designed for constructing the most effective binary (two-class) classifier. Subsequently, the method was broadened to encompass both regression and clustering challenges. The foundational principle of SVM is to identify a hyperplane, either linear or non-linear, that maximizes the margin. To maximize the margin is synonymous with:

$$\begin{aligned} &\underset{w,b,\zeta_i}{\text{minimize}} && \frac{1}{2}w^T w + C\left(\textstyle\sum_{i=1}^{N}\zeta_i\right) \\ &\text{subject to} && y_i(w^T x_i - b) + \zeta_i - 1 \geq 0, \quad 1 \leq i \leq N \\ & && \zeta_i \geq 0, \quad 1 \leq i \leq N \end{aligned}$$

Support Vector Machines (SVM) represent a subset of kernel-based methodologies. They transform feature vectors into a higher-dimensional space through the employment of a kernel function. Subsequently, they construct an optimal linear discriminant function within this expanded space, or an optimal hyperplane that aligns with the characteristics of the training data. In the instance of SVM, the kernel function itself is not explicitly delineated.

Rather, it is necessary to establish a distance measure between any two points within the expanded space. The primary attributes of Support Vector Machines (SVMs) include the employment of kernel functions, the non-existence of local minima, the sparse nature of their solutions, and the optimalization of capacity through margin control. In addition to these benefits, SVMs also present certain practical limitations. One significant issue that remains unresolved is the determination of the parameters for kernel functions. For Gaussian kernels, this involves selecting the width parameter [sigma], while for the epsilon-insensitive loss function, the choice of [epsilon] value is critical.

## 5.3. Naive Bayes

The Naive Bayes classifier serves as a probabilistic classification algorithm grounded on the principles of Bayes theorem, whilst operating under the assumptions of both strong and naïve independence. It is recognized as a foundational technique within text classification, finding extensive applications in domains such as email spam filtering, personal email organization, document categorization, identification of sexually explicit content, language classification, and sentiment analysis. Experimental observations have indicated that this algorithm exhibits robust performance across both numerical and textual data sets. While it frequently lags behind other methodologies like Decision Trees, Random Forests, Maximum Entropy, and Support Vector Machines, among others, the Naive Bayes classifier is notably efficient. This is attributed to its reduced computational demands across both CPU and memory resources, as well as its minimal requirement for training data. However, the model's performance suffers when faced with real-world data characterized by highly correlated features, thereby challenging the assumption of conditional independence and diminishing its capabilities.

## 5.4. Convolutional neural networks

Convolutional Neural Networks (CNNs) are computational models that are significantly inspired by the structure and functionality of the human visual cortex. These models operate by comparing images in a piecewise manner, employing a filter map that moves over distinct patches within the image. These individual patches, or 'features,' are identified by their proximity and similarity across different images. CNNs exhibit a superior capability in visual perception and classification tasks relative to other types of neural networks. The fundamental architecture of CNNs is widely recognized, characterized by a standard structure that includes several convolutional and dense layers. Generally, each CNN consists of three layers. The foundational architecture begins with a series of two convolutional layers, each equipped with 32 filters and a 3 by 3 window size. Subsequently, a maxpool layer is employed, followed by a dropout layer, establishing a 3-layer deep architecture. Following the initial group, we encounter an additional configuration comprising two consecutive convolutional layers, each equipped with 64 filters, alongside a max pooling layer and a dropout layer. Subsequently, we proceed with the establishment of two more convolutional layers, each again set to utilize 64 filters, followed by another max pooling layer.

The architecture culminates in a fully connected hidden layer, utilizing 512 neurons and employing the Rectified Linear Unit (ReLU) activation function, culminating in an output layer, this time employing the softmax activation function. The input to the first convolutional layer is an image of size (100, 100). The output from this stage consists of 36 neurons, corresponding to the 36 categories of ISL (International Sign Language) signage. This architectural framework is illustrated in Figure 4.1.

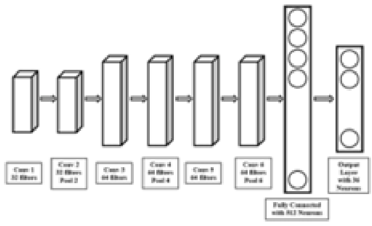

**Figure. 4.1**.CNN architecture

## 5.5. ResNet50

The ResNet50 architecture, also recognized as the Residual Network, is employed in the context of neural network models to identify optimal shortcuts. It stands as the preeminent choice among convolutional neural networks (CNNs) according to He et al. (2016). Fundamentally, the ResNet architecture comprises numerous residual modules, which constitute its key elements. As the complexity and scope of the network increase, the process of training presents challenges. Typically, input feature maps proceed through a series of operations including a convolutional filter, a non-linear activation function, and a pooling operation. Furthermore, the backpropagation algorithm is applied in its design. As network size expands, achieving convergence becomes increasingly arduous.

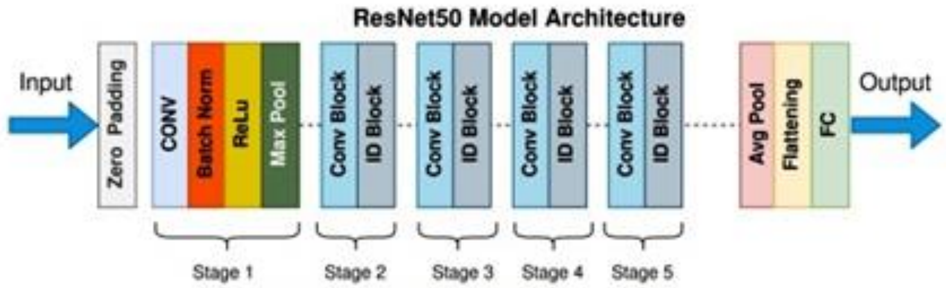


*Figure 4.2 ResNet50 Architecture* [21]

## 5.6. Comparative Study

Table 3 serves as a comprehensive analysis, juxtaposing diverse methodologies associated with document classification and deep learning research. The aim is to evaluate the effectiveness of these methodologies as delineated in the existing literature. The accuracy of classification for various attributes is systematically compared across the methods.

**Table 1.** Study on Document classification

| Authors | Approach | dataset | Language of Dataset | Accuracy |
|---|---|---|---|---|
| Abdulmunim etal. [20] | Using a decision tree classifier and the bagging technique can enhance classification results. | Images of official Arabic printed documents were gathered from several official websites. | Document images inthe Arabic language. | 97.35% |
| al-Khurabi et al.[21] | Artificial Neural Networks (ANN) were employed in the suggested system. | 120 different sorts of collected document photos. | Document images inthe Arabic language. | 86% |
| Arindam Das et al.[22] | The suggested setup made use of VGG16. | RVL – CDIP dataset | Document images in the English language. | 92.21% |
| Afzal et al. [23] | The suggested CNN model made use of network (AlexNet) pertained weights. | Tobacco-3428 Legislation | Document images in the English language. | 77.6% |
| Kölsch et al. [24] | The suggested CNN model made use of Extreme Learning Machines (ELMs), which offer real-time training, and took weights from a network (AlexNet). | Tobacco-3428 Legislation dataset | Document images in the English language. | 83.24% |
| Taghreed Alghamdi et al.[25] | The proposed method is based on a supervised deep learning technique using convolutional neural networks. The CNN model applies the dropout layer after every VGG block and uses pertained weights from a network (VGG16). | Arabic document images were collected from different sources. | Document images inthe Arabic language. | 92% |

## 6. CONCLUSION AND FUTURE SCOPE

This survey offers insights into the diverse deep learning techniques employed in document classification. It is evident that deep learning networks are effectively utilized in the analysis of large volumes of data, facilitating the discovery, application, and prediction of knowledge. Furthermore, deep learning serves as a potent catalyst for generating actionable outcomes. The application of deep learning techniques not only overcomes the limitations inherent in conventional machine learning methods but also ensures the generation of accurate and efficient solutions.

## REFERENCES


[1] A. Cruz-roa, J. C. Caicedo, and F. A. González, “Artificial Intelligence in Medicine Visual pattern mining in histology image collections using bag of features,” Artif. Intell. Med., vol. 52, no. 2, pp. 91–106, 2011, doi: 10.1016/j.artmed.2011.04.010.

[2] M. Puttagunta and S. Ravi, “Medical image analysis based on deep learning approach,” Multimed. Tools Appl., vol. 80, no. 16, pp. 24365–24398, 2021, doi: 10.1007/s11042-021-10707-4

[3] D. Jaswal, S. V, and K. P. Soman, “Image Classification Using Convolutional Neural Networks,” Int. J. Sci. Eng. Res., vol. 5, no. 6, pp. 1661–1668, 2014, doi: 10.14299/ijser.2014.06.002

[4] W. Liu, Z. Wang, X. Liu, Y. Liu, and F. E. Alsaadi, “A Survey of Deep Neural Network Architectures and Their Applications,” Neurocomputing, 2016, doi: 10.1016/j.neucom.2016.12.038.

[5] D. R. Sarvamangala and R. V. Kulkarni, “Convolutional neural networks in medical image understanding: a survey,” Evol. Intell., no. 0123456789, 2021, doi: 10.1007/s12065-020-00540-3.

[6] S. Bauer, R. Wiest, L. Nolte, and M. Reyes, “A survey of MRI-based medical image analysis for brain tumor studies,” vol. 97, 2013, doi: 10.1088/0031-9155/58/13/R97.

[7] M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir, “Deep Learning in Radiology : An Overview of the Concepts and a Survey of the State of the Art With Focus on MRI,” 2018, doi: 10.1002/jmri.26534.

[8] Harley AW, Ufkes A, Derpanis KG (2015) Evaluation of deep convolutional nets for document image classification and retrieval. In: 2015 13th international conference on document analysis and recognition (ICDAR), IEEE, pp 991–995

[9] Roy S, Das A, Bhattacharya U (2016) Generalized stacking of layerwise-trained deep convolutional neural networks for document image classification. In: 2016 23rd international conference on pattern recognition (ICPR), IEEE, pp 1273–1278

[10] Csurka G (2017) Document image classification, with a specific view on applications of patent images. In: Current challenges in patent information retrieval, Springer, pp 325–350

[11] Tensmeyer C, Martinez T (2017) Analysis of convolutional neural networks for document image classification. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR), vol 1. IEEE, pp 388–393

[12] S¸ahin S et al (2020) Dijital Dok¨umanların Anahtar Kelime Tabanlı Dogrulanması. In: Proceedings of the 6. Ulusal Y¨ ˘ uksek Bas¸arımlı Hesaplama Konferansı (in Turkish), pp 1–6

[13] Yaman D, Eyiokur FI, Ekenel HK (2017) Comparison of convolutional neural network models for document image classification. In: 2017 25th signal processing and communications applications conference (SIU), IEEE, pp 1–4

[14] Noce L et al (2016) Embedded textual content for document image classification with convolutional neural networks. In: Proceedings of the 2016 ACM symposium on document engineering, pp 165– 173

[15] Audebert N et al (2019) Multimodal deep networks for text and image-based document classification. In: Joint european conference on machine learning and knowledge discovery in databases, Springer, pp 427–443

[16] Fetterly, D., Manasse, M. & Najork, M. (2005). Detecting phrase-level duplication on the world wide web. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (pp. pp. 170-177). : ACM Press, Salvador, Brazil

[17] Hunnisett, D. S. & Teahan, W.J. (2004). Context-based methods for text categorisation. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (pp. pp. 578-579). : ACM Press, Sheffield, United Kingdom

[18] Stamatatos, E., Kokkinakis, G. & Fakotakis, N. (2000). Automatic text categorization in terms of genre and author. Computational Linguistics, 26, pp. 471-495.

[19] Muhammed Miah, "Improved k-NN Algorithm for Text Classification", Department of Computer Science and Engineering University of Texas at Arlington, TX, USA.

[20] Abdulmunim, M. E., & Abass, H. K. (2019). Classification and Retrieving Printed Arabic Document Images Based on Bagged Decision Tree Classifier. AL-MANSOUR JOURNAL, (32).

[21] Al-Khurabi, Abd Allah Ali& Mansur, Muhammad Abd Allah. 2004. Arabic document image classification using neural networks. Mansoura Engineering Journal‹Vol. 29, no. 1, pp.1-8.

[22] Arindam Das et al."Document image classification with intra – domain transfer learning and stacked generalization of deep convolutional neural networks "24th international conference in pattern recognition(ICPR)2018.

[23] Afzal, M. Z., Capobianco, S., Malik, M. I., Marinai, S., Breuel, T. M., Dengel, A., & Liwicki, M. (2015, August). Deepdocclassifier: Document classification with a deep convolutional neural network. In 2015 13th international conference on document analysis and recognition (ICDAR) (pp. 1111-1115). IEEE.

[24] Kölsch, A., Afzal, M. Z., Ebbecke, M., & Liwicki, M. (2017, November). Real-time document image classification using deep CNN and extreme learning machines. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (Vol. 1, pp. 1318-1323). IEEE.

[25] Taghreed Alghamdi et al "Arabic Document classificationby deep learning".In 2021 international journel of advanced computer science and applications (IJACSA)(vol.12. No. 10.2021)

[26] Fang Lu Qingyuan Bai, "A Refined Weighted KNearest Neighbours Algorithm for Text Categorization", IEEE 2010.

[27] Analysis of Convolutional Neural Networks for Document Image Classification-Chris Tensmeyer and Tony Martinez-14th IAPR International Conference on Document Analysis and Recognition 2017

[28] Rupali Bhaisare,T. RajuRao 2013 "Review On Text Mining With Pattern Discovery".

[29] Fang Lu Qingyuan Bai, "A Refined Weighted KNearest Neighbours Algorithm for Text Categorization", IEEE 2010.

[30] Vidhya. K.A G.Aghila, “A Survey of Naïve Bayes Machine Learning approach in Text Document Classification”, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, 2010.

[31] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation." CoRR, vol. abs/1311.2524, 2013

[32] S. Sarraf, “Binary Image Segmentation Using Classification Methods: Support Vector Machines, Artificial Neural Networks and K th Nearest Neighbours,” Int. J. Comput., vol. 24, no. 1, pp. 56–79, 2017.

**Citation:** A. Vijaya Mahendra Varman and AN Sigappi, Digitized Document Classification Using Machine and Deep Learning A Survey, International Journal of Computer Engineering and Technology (IJCET), 15(4), 2024, pp. 29-38

**Abstract Link:** https://iaeme.com/Home/article_id/IJCET_15_04_003

**Article Link:**
https://iaeme.com/MasterAdmin/Journal_uploads/IJCET/VOLUME_15_ISSUE_4/IJCET_15_04_003.pdf



**editor@iaeme.com**