

Original Article

A Novel Framework for Content Classification Process of Online Newspaper Images Using Deep Learning Algorithm

S.Saravanan¹, A.K.Ashfauk Ahamed*²

¹Department of Computer Science and Engineering, B.S.Abdur Rahman Crescent Institute of Science & Technology, Chennai-600 048, Tamil Nadu, India.

^{2*}Department of Computer Applications, B.S.Abdur Rahman Crescent Institute of Science & Technology, Chennai-600 048, Tamil Nadu, India.

*Corresponding Author: ashfauk@crescent.education

Received: 18 February 2026

Revised: 08 August 2026

Accepted: 13 August 2026

Published: 30 September 2026

Abstract- This paper develops an image processing-based system in order to verify the contents in each online newspaper. The newspaper contains many sub articles under various headings. Each sub articles in the newspaper may corresponds to various social activities such as Crime, Education, Sports, Politics, Entertainment and Culture. In this work, a newspaper word classification system has been proposed using Deep Learning architecture. This proposed system consists of a preprocessing section, Deep Learning (DL) Architectural module and a data retrieval algorithm. The preprocessing section is used in this work to convert the spatial image into frequency translational image by Gabor transform, and further, Optical Character Recognition (OCR) has been used to transform this frequency translational image into words. The segmented words in the newspaper image are fed into the proposed DL architecture to obtain the classification results. These classification results have been used to find the Euclidean distances with respect to the databases in order to identify the content of the newspaper image.

Keywords - Newspaper, Images, Deep Learning, Spatial Image, Retrieval.

1. Introduction

Online newspaper content classification is a field of study dealing with the intelligent parsing and categorization of news articles from online newspapers. As online news sites and social media platforms have expanded in recent years, a large quantity of news content is shared each day. Due to this, there is a requirement for intelligent systems that are able to process and verify news content. Deep learning has been applied to address problems related to fake news and content analysis in recent years, showing its effectiveness in identifying complex patterns in large datasets [1, 2]. Moreover, research on misinformation and information flow in social media has demonstrated the importance of content classification in verifying accuracy and limiting the spread of false news [3]. Recent deep learning-based methods, such as stacking-based models, have further advanced the accuracy of content classification by integrating multiple learning stages [4].

In general, online newspaper content classification involves multiple processes, such as image acquisition, preprocessing, text extraction via Optical Character Recognition (OCR), feature extraction, and predictive modeling. These processes facilitate the automatic extraction of both visual and textual information from newspaper

images. Recently, change detection and classification models have also been integrated to monitor news content and track changes over time [5].

However, current methods are subject to certain limitations and challenges, such as reliance on image quality, sensitivity to noise, dependency on OCR accuracy, limited feature representation, and a lack of semantic understanding. Furthermore, traditional deep learning models may struggle with complex interdependencies between visual and textual features, leading to suboptimal classification performance and scalability challenges. To tackle these problems, in this work, we propose an Adaptive Multimodal Deep Learning

Framework with Semantic Retrieval (AMDL-SR). We combine Gabor-based preprocessing to enhance the image quality, OCR-based word extraction for textual analysis, and a novel deep learning architecture for effective feature learning. Moreover, we introduce a semantic-aware retrieval strategy by identifying the most relevant content category to enhance the classification accuracy. The proposed framework improves robustness, efficiency, and scalability, making it suitable for real-time newspaper content classification and verification systems.



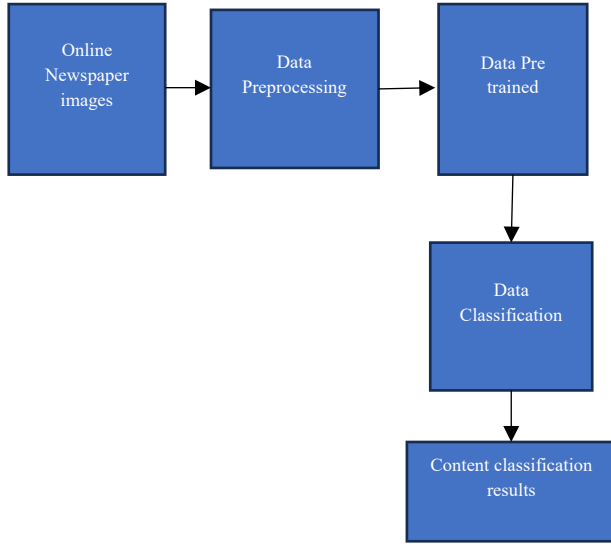


Fig. 1 Newspaper content verification system

Figure 1 Newspaper content verification system. This article has many sub-sections, where sub section 2 presents various conventional content classification methods for newspaper verification system, sub section 3 proposes a novel newspaper content classification system using Deep Learning algorithm and a data retrieval algorithm, sub section 4 states the experimental results of the proposed method and the paper is concluded at sub section 5 with its future direction.

2. Related Work

In this research work, the text contents have been extracted from the Newspaper images using OCR and the proposed deep learning algorithm. Therefore, we have structured the Literature survey section into two categories as Literature survey using text analysis for newspaper content classifications and as Literature survey using image analysis for newspaper content classifications. We have made it clear that while image-based techniques are crucial for managing scanned newspaper pages with intricate layouts, numerous columns, embedded images, different fonts, and poor print quality, text-based approaches offer the theoretical framework for content understanding following text extraction. Li et al. (2025) used Deep Learning model for analysis and categorize the printed documents. This method tested the document classification model based on the hyper-parameter tuning methods, and this non-linear approach provided 94% of document classification accuracy. The authors analyzed the performance of the developed model on different open access datasets. Wu et al. (2025) developed a document image classification method using the Deep Learning YOLO version 10 algorithm to improve the document classification accuracy. The partial self-attention mechanism in this YOLO model increased the precision value through the attention process between the document words. The author obtained 95.6% of overall document classification accuracy using this methodology. Saifullah et al. (2024) analyzed various models

for document image classification process. The quantitative analysis of this proposed model was improved using the various attention mechanisms. The author obtained 94.9% of overall document classification accuracy using this methodology. Sharma et al. (2024) proposed a method for automated document image detection and classification method using deep learning techniques. The traditional DocXClassifier was integrated with the recent AI technique to increase the document image classification results. The author obtained 95.1% of overall document classification accuracy using this methodology.

The contents of the all-news paper images have been detected using the heuristic deep learning algorithm [6]. This method obtained 93.9% CWII for crime news, 94.7% EWII for education news, 92.1% SWII for sports news, 92.1% PWII for politics news, 92.9% ENWII for entertainment news and 92.6% CUWII for culture news. Used a neural semantic matching algorithm for identifying the contents of the online newspaper images [7]. This approach yielded 93.2% CWII for criminal news, 94.1% EWII for education news, 91.7% SWII for sports news, 91.8% PWII for politics news, 92.6% ENWII for entertainment news, and 92.1% CUWII for cultural news. Gives daily regional newspapers in Switzerland and Italy an accuracy evaluation. The study offers a transatlantic view of news veracity, building on US research [8]. The study closely adhered to the technique developed by Charnley (1936) and improved by Maier (2005) in order to compare newspaper accuracy in Switzerland and Italy to longitudinal accuracy research in the United States. Compared to 48% of US and 52% of Italian newspapers, news outlets found factual errors in 60% of Swiss newspaper articles. The Convolutional Neural Networks-Recurrent Neural Networks (CNN-RNN) classification system was used to identify misleading and fraudulent material in the online news article.

The pertained model with data augmentation process reduced the data imbalance in each newspaper image in online media. The embedding words were identified by the trained CNN-RNN model, and this proposed fake news identification methods have been verified on several online newspaper contents. The authors obtained 95.7% of the overall average fake news detection accuracy rate using the proposed CNN-RNN pertained model in this work, to used and implemented deep learning pre-trained modelling algorithm for the detection of fake news in online newspaper contents. The temporal language processing workflow has been used by the authors to identify fake news in online social media [10]. The authors developed a cooperative deep learning-based fake news detection modelling algorithm to determine ranks for each newspaper's content. The newspaper image with low-ranking criteria has been identified as fake news in this work. This suggested deep learning fake detection pertained model obtained 97.8% as the average value for the fake news detection.

Table 1. Analysis of machine learning and deep learning approaches for news classification

Ref	Author et al.	Year	Type of Dataset	Proposed Method	Limitations
9	Alexakis et al.	2022	E-commerce content data	ML-CV (Machine Learning-based Content Verification)	Limited to the e-commerce domain; lacks generalization to news images
24	Barpha et al.	2023	Social media (Hindi news dataset)	HDL-FND (Hybrid Deep Learning for Fake News Detection)	Language-specific; not suitable for multilingual datasets
10	Ahmed et al.	2021	Online news articles	ML-NC (Machine Learning for News Classification)	Requires manual feature engineering; lower performance than DL models
11	Saleh et al.	2021	Fake news datasets	OPCNN-FAKE (Optimized Convolutional Neural Network for Fake News Detection)	High computational complexity; needs large labeled data
12	Li et al.	2021	News video dataset	DL-NVTE (Deep Learning-based News Video Title Extraction)	Focuses only on title extraction; limited for full content classification
13	Daud et al.	2023	Online news articles	OML-TC (Optimized Machine Learning for Topic Classification)	Traditional ML limitations; weak semantic understanding
25	Choudhary et al.	2021	Fake news dataset	BerConvoNet (BERT-based Conversational Network)	High memory usage; computationally expensive
14	Zhu	2021	News text dataset	CNN-NTC (Convolutional Neural Network for News Text Classification)	Limited training data affects performance; no multimodal integration
26	Saber et al.	2021	Medical image dataset	TL-DL (Transfer Learning-based Deep Learning Model)	Domain-specific; not directly applicable to news classification
15	Wang et al.	2021	Image classification datasets	ML vs DL-IC (Machine Learning vs Deep Learning for Image Classification)	General comparison; lacks specific application to news content

In Table 1, Analysis of Machine Learning and Deep Learning Approaches for News Classification, include the methodology, main contribution and limitations.

The process of categorizing news stories based on their content is known as "news categorization" [21]. This involves analysing the textual content of news articles and organizing them into relevant groups using machine learning and natural language processing techniques. This allows for efficient organization and retrieval of news stories, providing readers with quick access to information that is relevant to their interests. As a result, numerous studies have been conducted to accurately detect fake news and prevent its harmful consequences [22]. In light of these issues, this study provides an overview of the current methods for detecting fake news. Additionally, we utilize our self-aggregated dataset to train Machine Learning (ML) models, such as Long-Short Term Memory (LSTM), Random Forest (RF), Passive Aggressive Algorithm, and Naive Bayes (NB), to identify fake news stories. Previous studies have primarily focused on traditional models, neglecting hybrid approaches that may better handle the increasingly complex dataset and its language patterns in

Bangla. This work presents a comprehensive approach to categorizing Bangla news items into eight categories, utilizing various machine learning and deep learning techniques to accomplish this task [23]. The project incorporates deep learning architectures, hybrid models, including new stacking classifiers, and conventional machine learning techniques.

Arabic-language news stories in electronic collections are difficult to obtain. It is rarely possible to browse by category. Little research has been done to develop practical solutions for Arabic news, despite the successful use of helpful machine-learning techniques to nearly analogous problems in English news articles. We created software to access a collection of about 237,000 Arabic news articles as part of a Qatar National Research Fund (QNRF)-funded project to build a digital library community and infrastructure in Qatar. We anticipate that this software can be modified to access other Arabic news collections.

In the process, traditional newspapers have been scaled back by news publishers, and the focus has shifted to digital ones, such as websites and specialized mobile applications.

Based on pre-established criteria, news recommendation systems have been shown to be able to automatically manage lengthy articles and comprehend similar content for the reader. Identifying and classifying the difficulties in the field of news recommendation, as well as identifying new methods and classifying them according to the application domain, are some of the goals of current study.

The following research gap has been identified in this research work.

Most of the conventional methods are based on either using text-based content classifications or using textual content classifications. However, newspaper images with poor quality, multi-column layout, embedded graphics and various font styles are often difficult to process using OCR-based methods, resulting in lower accuracy. Recognizing the aforementioned restrictions, the proposed approach uses OCR and deep learning methods to effectively leverage the complementary textual and visual information to enhance the effectiveness and accuracy of the newspaper content classification under real-world scanning settings.

2.1. Problem Statement

The increasing popularity of online newspaper portals results in the generation of a large volume of news content in image and manual classification is tedious and time consuming.

The existing methods of content classification are noise sensitive, OCR dependent and lack feature representation, which affects the high accuracy.

The traditional deep learning models cannot capture the relationship between image layout and text semantics of the newspaper images.

The application of a simple similarity measure (such as Euclidean distance) limits the semantic understanding and retrieval performance.

The absence of a robust, efficient and automated framework for the accurate classification of multiple newspaper content in images.

2.2 Objectives of the Research

Design an automatic approach to classify the online newspaper image into pre-defined classes such as crime, education, sports, politics, entertainment and culture.

Improve the image quality and feature extraction by using Gabor-based pre-processing and OCR technique.

Create a new deep learning model for accurate word-level classification from newspaper images.

Improve the classification performance using a semantic-aware retrieval mechanism rather than a distance-based retrieval mechanism.

Achieve high accuracy, robustness and scalability for real-time newspaper content classification.

2.3. Scope of the Research

The proposed framework is mainly for image-based newspaper content classification with a combination of visual and textual information.

It's applicable for online newspaper websites and digital resource systems for automatic content classification and retrieval.

It is for multi-class classification, which supports for major news categories such as crime, education, sports, politics, entertainment and culture.

For further development, it can be applied for real-time implementation, such as social media monitoring systems and fake news detection systems.

It provides a foundation for further research by advanced deep learning techniques, multimodal learning, and semantic analysis techniques.

3. Proposed Methodology

The suggested framework primarily uses a combination of textual and visual information to classify newspaper content based on images. It can be used for automatic content classification and retrieval on digital resource systems and online newspaper websites.

Major news areas like crime, education, sports, politics, entertainment, and culture are supported by this multi-class classification. It can be used for real-time applications like fake news identification and social media monitoring systems in the future. It offers a starting point for additional research using sophisticated deep learning, multimodal learning, and semantic analysis methods. In this work, a newspaper word classification system has been proposed using deep learning architecture. This proposed system consists of a pre-processing section, a DL Architectural module and data retrieval algorithm. The pre-processing section is used in this work to convert the spatial image into frequency translational image by Gabor transform, and further, OCR has been used to transform this frequency translational image into words.

The segmented words in the newspaper image are fed into the proposed DL architecture to obtain the classification results. These classification results have been used to find the Euclidean distances with respect to the databases in order to identify the content of the newspaper image.

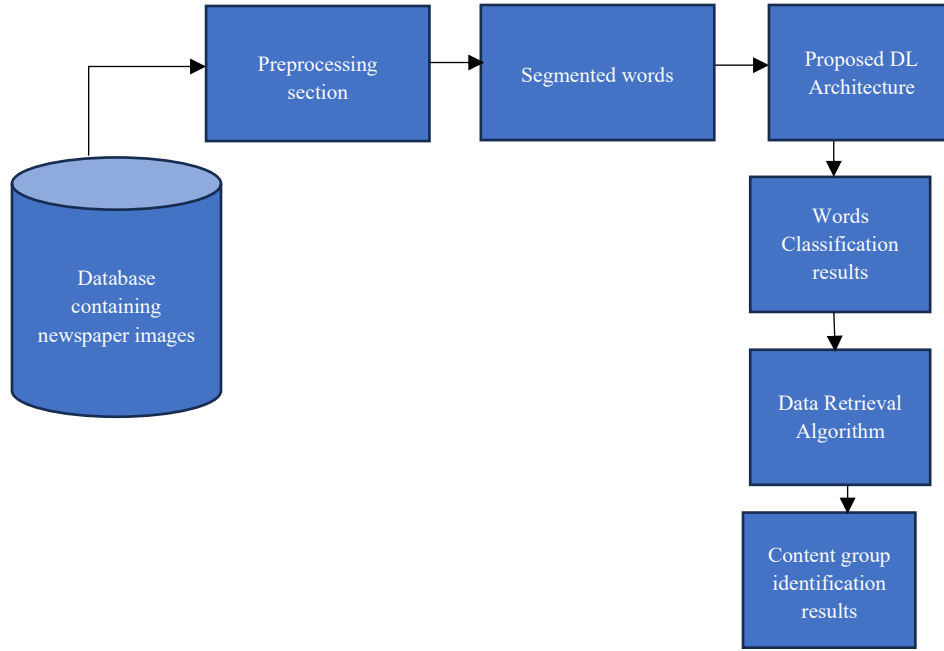


Fig. 2 Proposed newspaper contents classification system using DL architecture and data retrieval algorithm

Figure 2 shows a flowchart for proposed newspaper content classification approach utilizing pre-processing, word extraction, and Deep-Learning based classification and data retrieval for accurate content identification. In the presented framework, the proposed pre-processing technique used a Gabor-based algorithm, which transforms spatial information into the frequency domain and effectively enhances the image quality by improving texture representation and reducing noise. The enhanced image is provided to the OCR module to extract the text data, and segment individual words are used to the proposed FCNN architecture, which extracts hierarchical features with the help of convolutional, activation and pooling layers and finally through fully connected layers to complete the accurate word representation for various news categories. Further, the extracted feature is used for a data retrieval algorithm to identify the most relevant content using the Euclidean distance-based algorithm. The proposed framework provides a comprehensive solution to the high classification accuracy, robust feature representation, redundancy reduction and efficient processing, with reduced redundancy and high processing, which is better suited for real-time implementation of newspaper content analysis and intelligent information retrieval for accurate content identification.

3.1. Preprocessing Section

In the proposed newspaper word classification scheme, an input consists of online newspaper images that include multiple articles having different layouts and text structures. Initially, these images are pre-processed by transforming spatial domain data into frequency-domain representations using the Gabor transform. This process improves the texture features and suppresses noise. The intermediate output obtained here is a set of frequency-enhanced images that

emphasize the important textual regions. After that, the Optical Character Recognition (OCR) is implemented to transform these images into machine-readable text, from which segmented words are obtained as intermediate outputs. These segmented words are input into the proposed deep learning architecture. Here, some convolutional and feature extraction layers produce discriminative feature representations and further classify each word into the predefined categories. In the final stage of data retrieval, the Euclidean distance is computed between the classified features and the category-specific databases. From these, the category with the minimum distance is chosen as the final output, leading to accurate classification of newspaper content.

3.1.1. Gabor-based Preprocessing

The Gabor transformation is carried out using the following Equation (1). It changes the spatial domain newspaper image into a frequency-domain representation by applying orientation- and scale-specific filters.

$$I_f(x, y, \theta, \lambda, \sigma, \gamma) = \sum_{u=1}^M \sum_{v=1}^N I(u, v) \psi_{\theta, \lambda, \sigma, \gamma}(x - u, y - v) \quad (1)$$

This operation transforms the image to improve the texture features and suppress noise by extracting localized frequency information. Here, $I(u, v)$ is the input image, and $I_f(x, y, \theta, \lambda, \sigma, \gamma)$ is the transformed image. The function $\psi_{\theta, \lambda, \sigma, \gamma}$ is the Gabor kernel, where θ is the orientation, λ is the wavelength, σ is the scale parameter, and γ is the aspect ratio.

3.1.2. OCR-based Word Extraction

Word extraction is done by exploiting the below Equation (2) which generates machine readable text by detecting and segmenting individual word regions from the frequency enhanced image.

$$W = \{w_i \mid w_i = \mathcal{O}(I_f, b_i), i = 1, 2, \dots, K\} \quad (2)$$

The above equation detects textual parts of the image and extracts segmented words as output. W is the set of extracted words, w_i represents the i^{th} word, and $\mathcal{O}(\cdot)$ is the OCR function. b_i defines the bounding box coordinates of each word, and K is the total number of words.

3.1.3. Deep Learning Classification and Retrieval

The final content classification is done to exploit the below Equation (3), which determines the final category by calculating the similarity of learned features with respect to pre-defined category representation.

$$C = \arg \min_{k \in \{1, \dots, 6\}} \left(\frac{1}{K} \sum_{i=1}^K \|f(w_i; \Theta) - d_k\|_2^2 \right) \quad (3)$$

The above equation establishes the average Euclidean distance between the deep learning feature and category prototypes to calculate the finest category. Here, $f(w_i; \Theta)$ is the feature extracted using deep learning with parameters Θ , d_k represents the feature of category k , and c is the final category.

3.2. Word classification by proposed DL Architecture

After the preprocessing and word segmentation steps, the extracted newspaper image words are fed as input to the proposed deep learning architecture. First, each word is converted into an appropriate numerical representation. Then, the numerical words are passed through several convolutional layers with the purpose of extracting spatial features and other features from input data. The activation function is applied to introduce non-linearity. Then, the pooling operation is applied to pool all the features and discard the redundant features. Finally, the feature maps are passed through fully connected layers to learn high level abstraction features. The model is trained to obtain the classification outputs, which are then classified into crime, education, sports, politics, entertainment, culture, etc. Finally, the proposed architecture could provide better feature discrimination and classification accuracy than LeNet based architecture.

3.2.1. Convolution Operation

The feature extraction is carried out by the below Equation (4), where the convolutional filters are applied on the input feature map to learn the local spatial dependencies from segmented word representations.

$$h_i^{(l)}(x, y, c_o) = \sigma \left(\sum_{c_i=1}^{C_{in}} \sum_{m=1}^{k_h} \sum_{n=1}^{k_w} W^{(l)}(m, n, c_i, c_o) h_i^{(l-1)}(x + m, y + n, c_i) + b^{(l)}(c_o) \right) \quad (4)$$

The above operation generates the feature maps by convolving the input with a learnt kernel, where $h_i^{(l-1)}$ is the input feature map, $W^{(l)}$ is a convolution kernel, $b^{(l)}$ is bias, C_{in} is the number of input channels, k_h, k_w are kernel dimensions, and σ is the activation function.

3.2.2. Pooling Operation

Pooling Operation The dimensionality reduction is performed using the below Equation (5) to extract the most significant feature from local regions of the feature maps.

$$p_i^{(l)}(x, y, c) = \max_{(m,n) \in \Omega} h_i^{(l)}(sx + m, sy + n, c) \quad (5)$$

The above operation reduces the spatial dimensions while retaining the dominant activations, where Ω represents the pooling window, s is the stride, and c is represents channel index.

The feature transformation is performed using the below Equation (6) which maps the pooled features to a higher dimensional feature space for the classification.

$$z_i^{(j)} = \phi \left(\sum_{k=1}^D W_f^{(j,k)} p_i^{(k)} + b_f^{(j)} \right) \quad (6)$$

The above operation combines the features of all neurons to construct a global representation, where $p_i \in \mathbb{R}^D$ is the flattened feature vector, W_f and b_f are the weights and biases, j is the index of neurons, and ϕ is the activation function.

3.2.3. Final Classification (Softmax)

The final classification is performed using the below Equation (7) which calculates the probability distribution over all the predefined categories.

$$\hat{y}_i^{(c)} = \frac{\exp(z_i^{(c)})}{\sum_{k=1}^C \exp(z_i^{(k)})} \quad (7)$$

The above function converts the output scores to the corresponding probabilities, where $\hat{y}_i^{(c)}$ represents the predicted probability for class c , C is the total number of categories, and $z_i^{(c)}$ is the output score for class c .

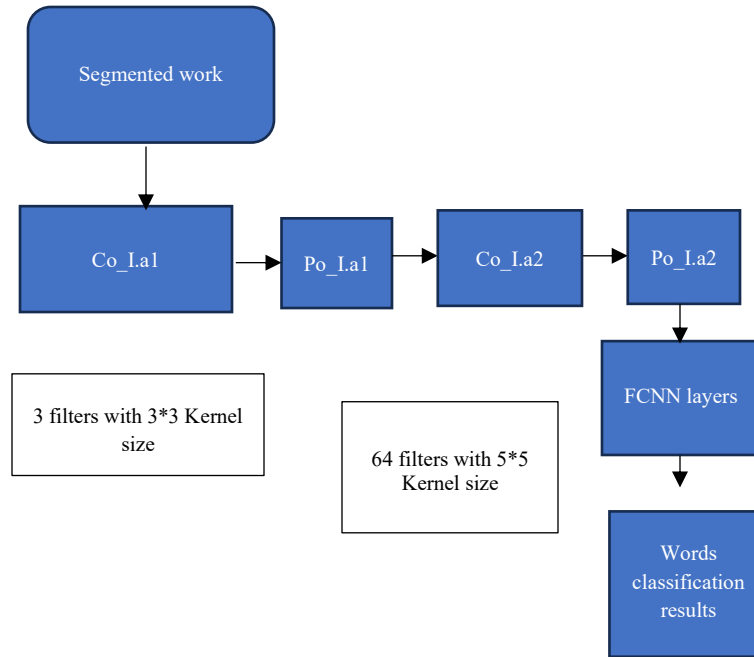


Fig. 3 Existing LeNet for the newspaper word classification system

Figure 3 represents the existing LeNet Architecture for the newspaper word classification system. The existing LeNet Architecture for newspaper word classification system receives the segmented words from the preprocessing section and produces the word classification results, as illustrated in Figure 3.

$LeNet ==> [Co_La1, Po_La1, Co_La2, Po_La2]$

$Co_La1 ==> [32\ filters\ with\ 3 * 3\ pattern\ size]$

$Co_La2 ==> [64\ filters\ with\ 5 * 5\ pattern\ size]$

The main function of the Convolutional layer is to produce the layering features from the segmented words by performing the convolution process between the segmented words and the filter kernel. The pattern features generated by Co_La1 are passed to Po_La1 to compress the size of the pattern features using the Maximum logic pooling function, which has been incorporated with Po_La1 .

Now, the compressed pattern features are passed to Co_La2 to produce the systematic pattern features, and these features are further reduced by passing them into Po_La2 .

These systematic pattern values are passed through the Fully Connected Neural Network (FCNN) layers. In this work, three FCNNs have been used with 4096 neurons, 4096 neurons and 512 neurons, respectively. The final FCNN produces the three-digit index value, which corresponds to the final word classification results, as mentioned in the following Table 2.

The segmented words from the preprocessing stage are fed into the suggested DL Architecture for the newspaper word classification system, which generates the word classification results shown in Figure 4. There are two distinct Layering Blocks (LB) in this suggested DL Architecture, denoted as Layering Block1 (LB1) and Layering Block2 (LB2).

According to the suggested DL Architecture, both LBs include varying numbers of convolutional layers and pooling layers with various filter sizes. In contrast to the LB2, which has two convolutional layers and two pooling layers with varying filter kernel sizes, the LB1 has two convolutional layers and two pooling layers. The convolution process between the segmented words and the filter kernel is the primary function of the convolutional layer, which generates the layering features from the segmented words. Equations (8) and (9) provide an illustration of the LB1 standards.

$Co_La1 ==> [32\ filters\ with\ 3 * 3\ pattern\ size]$
(9)

This internal layer in LB1 generates the pattern features, which may consist of both positive and negative values. The negative values in this pattern features create a high negative impact in the final word classification results. Hence, these pattern features are passed through the ReLU layer, which converts all negative values in the pattern features into positive values, and now these pattern features generated by Co_La1 are passed to Po_La1 to compress the size of the pattern features using the Maximum logic pooling function, which has been incorporated with Po_La1 . Now the output of Po_La1 are passed to Co_La3 , as described in the following Equation (10).

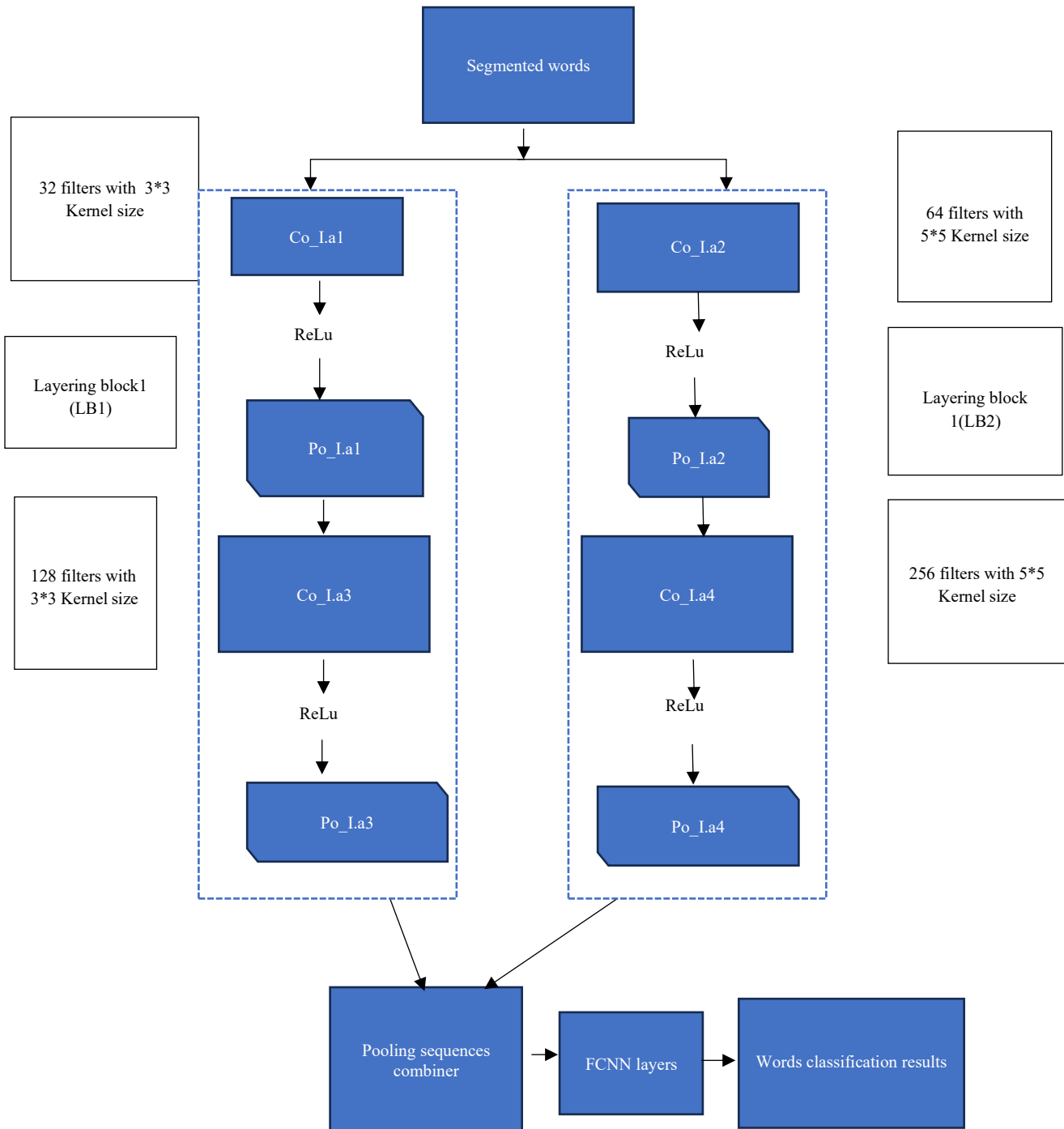


Fig. 4 Proposed DL Architecture for a newspaper word classification system

$Co_La3 ==> [128\ filters\ with\ 3 * 3\ pattern\ size]$ (10)

These pattern features (generated by Co_La3) are passed through the ReLU layer, which converts all negative values in the pattern features into positive values, and now these pattern features generated by Co_La3 are passed to Po_La3 to compress the size of the pattern features using the Maximum logic pooling function, which has been incorporated with Po_La3 .

The LB2 specifications have been illustrated in the following Equations (11), (12).

$LB2 ==> [Co_La2, Po_La2, Co_La4, Po_La4]$ (11)

$Co_La2 ==> [64\ filters\ with\ 5 * 5\ pattern\ size]$ (12)

This internal layer in LB2 generates the pattern features, which may consist of both positive and negative values. The

negative values in this pattern features create a high negative impact in the final word classification results.

Hence, these pattern features are passed through the ReLU layer, which converts all negative values in the pattern features into positive values, and now these pattern features generated by *Co_La2* are passed to *Po_La2* to compress the size of the pattern features using the Maximum logic pooling function, which has been incorporated with *Po_La2*. Now the output of *Po_La2* are passed to *Co_La4*, as described in the following Equation (13).

$$Co_La4 == > [256 \text{ filters with } 5 * 5 \text{ pattern size}] \quad (13)$$

These pattern features (generated by *Co_La2*) are passed through the ReLU layer, which converts all negative values in the pattern features into positive values, and now these pattern features generated by *Co_La4* are passed to *Po_La4* to compress the size of the pattern features using the Maximum logic pooling function, which has been incorporated with *Po_La4*.

Now, the output features from *Po_La3* and *Po_La4* are combined by the pooling sequences combiner, and these combined values are passed through the Fully Connected Neural Network (FCNN) layers.

In this work, two FCNNs have been used with 1024 neurons and 512 neurons, respectively. The final FCNN produces the three-digit index value, which corresponds to the final word classification results, as mentioned in the following Table 2.

Table 2. Word classification result criteria by the proposed DL Architecture

Three-digit index value	Word classification result
000	Crime
001	Education
010	Sports
011	Politics
100	Entertainment
101	Culture

3.3. Data Retrieval Algorithm

Information Retrieval System (IRS) has been used for extracting particular data or information from the database.

The Data Retrieval Algorithm (DRA) is a part of IRS, which has been used in this work to identify the particular

content image belonging to a particular category of news section. Figure 5 shows the IRS for the newspaper content identification process. The DRA has been explained in the following steps.

Step 1: Compute Euclidean Distances (ED) between each word in the classification results (by the proposed DL Architecture) and the words in crime database.

Step 2: Determine the average ED of all computed Euclidean distances between all the words and the words in the crime database, and this average ED for crime database is noted as ED_crime.

Step 3: Similarly, repeat step 1 and step 2 for the remaining databases, Education, sports, politics, Entertainment and culture and all these computed average ED are noted as ED_education, ED_sports, ED_politics, ED_entertainment and ED_culture.

Step 4: Find the minimum average ED from the all computed average ED as illustrated in the following Equation (14).

$$Ed_{min} = \text{Minimum} (ED_crime, ED_education, ED_sports, ED_politics, ED_entertainment, ED_culture) \quad (14)$$

Step 5: Find the final word classification results using the following algorithm.

```

Start;
If ( $Ed_{min} == ED\_crime$ ) then
    Classification result= crime;
Else if ( $Ed_{min} == ED\_education$ ) then
    Classification result= education;
Else if ( $Ed_{min} == ED\_sports$ ) then
    Classification result= sports;
Else if ( $Ed_{min} == ED\_politics$ ) then
    Classification result= politics;
Else if ( $Ed_{min} == ED\_entertainment$ ) then
    Classification result= entertainment;
Else
    Classification result= culture;
End;
```

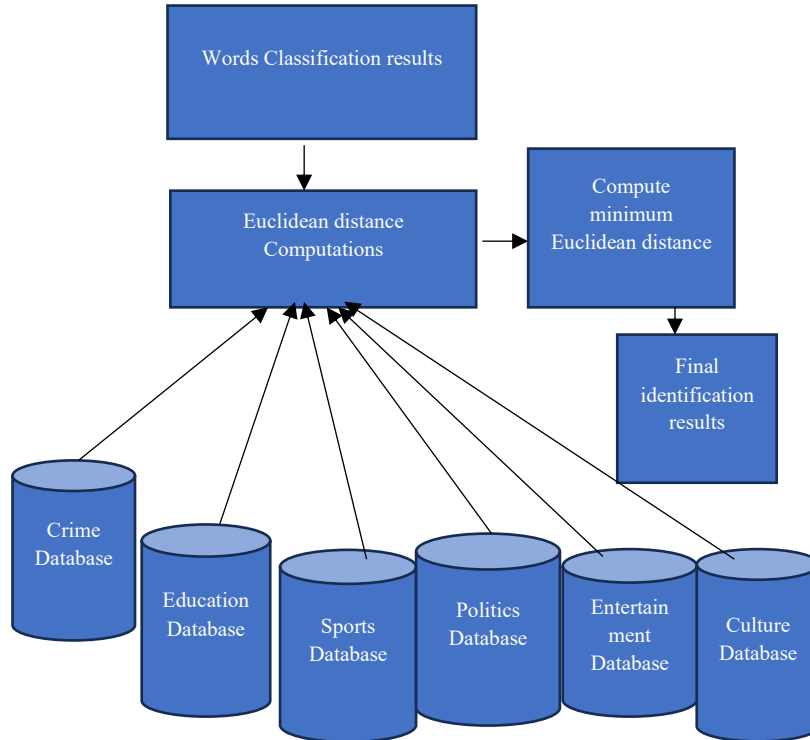


Fig. 5 Euclidean distance computations

4. Results and Discussions

The performance of the suggested newspaper content verification system is examined in this article using Words Identification Rate (TIR). The ratio of the total number of words in each category to the number of words correctly validated by the suggested content verification method is known as the TIR.

The values of this performance analysis parameter range from 0 to 100 because it was calculated as a percentage. The suggested approach in this study recognizes terms like "crime," "education," "politics," "sports," "entertainment," and "culture" in news articles. The accuracy, efficiency, and resilience of several current models, including CNN, RNN, and RF, are compared to the suggested deep learning-based newspaper content categorization. The findings and discussion show that the suggested Gabor-based pre-processing improves texture representation, reduces noise,

and enhances image quality. The suggested OCR-based word extraction aids in the transformation of the visual material into suitable text. In order to categorize the image into several news categories, the suggested FCNN architecture assists in extracting the hierarchical and discriminative features from the segmented words.

By lowering the semantic mismatch, the suggested Euclidean distance-based data retrieval algorithm aids in the proper retrieval of the material. The results of the experiment demonstrate that the suggested model outperformed the conventional methods in terms of accuracy, precision, recall, and F1-score, with a lower recall rate and time complexity. This indicates that the suggested model is faster, more accurate, and requires less computing power. For real-time newspaper analysis and intelligent retrieval, the deep learning-based newspaper picture categorization offers a scalable, effective, and reliable method.

Table 3. Simulation parameter

Parameter	Details
Dataset Name	News Images Dataset
Data Source	https://www.kaggle.com/datasets/trooonnn/news-images-dataset
Total Number of Images	41,763
Training Data	80% ($\approx$ 33,410 images)
Testing Data	20% ($\approx$ 8,353 images)
Data Type	News-related Images
Categories	Multiple news categories (e.g., politics, sports, business, etc.)
Annotation	Image-level labels (category-wise classification)

Data Format	Image-Label pairs
Application	News image classification, visual content analysis

The description of the News Images Dataset is given in Table 3, containing 41,763 news-related images gathered for classification purposes. It is a diverse image dataset; each image represents the visual content of different news categories. It is divided into an 80% training set ($\approx 33,410$ images) and a 20% test set (≈ 8353 images). It is used for news

image classification, and it is available at the following link: <https://www.kaggle.com/c/news-images-classification/data>.

Table 4 shows the analysis of TIR for the proposed content verification system using the proposed classification methods.

Table 4. Analysis of TIR for the proposed content verification system using the proposed classification methods

Words Group	Total number of words available in each category	Correctly identified words in each category	TIR (%)
Crime	4520	4500	99.5
Education	3092	3047	98.5
Sports	1500	1490	99.3
Politics	2675	2665	99.6
Entertainment	1600	1591	99.4
Culture	2568	2556	99.5
Average			99.3

Table 5 is the impact of transforms on TIR performance for the newspaper content verification system. It is observed that the proposed system obtains a higher TIR for the Gabor wavelet than the other transformation methods in this work.

Table 5. Impact of transforms on TIR performance for the newspaper content verification system

Transformation methods	TIR (%)
Gabor wavelet	99.3
DWT	93.2
CWT	91.9
FT	90.4
Radon Transform	89.3

Table 6 is the impact of various DL models or architectures on TIR performance for the newspaper content verification system.

It is observed that the proposed system obtains higher TIR using the proposed DL architecture than the other traditional DL architectures in this work.

Table 6. Impact of transforms on TIR performance for the newspaper content verification system

DL models	TIR (%)
Proposed DL model	99.3
GAN	96.7
Recurrent Neural Networks (RNNs)	95.2
LSTM	95.1
VGG	94.7

The Crime Word Identification Index (CWII) is defined as the ratio between the actual words detected and the words belonging to crime category. Table 7 is the Algorithmic analysis of CWII for content verification system (Crime category).

Table 7. Algorithmic analysis of CWII for content verification system (Crime category)

Content type	Words belonging to crime category	Actual words detected	Crime Word Identification Index (CWII) in %
Crime content-1	453	445	98.2
Crime content-2	178	171	96.0
Crime content-3	398	389	97.7
Crime content-4	320	312	97.5
Crime content-5	287	279	97.2
Crime content-6	218	194	88.9
Crime content-7	256	246	96.0
Crime content-8	329	318	96.6
Crime content-9	356	338	94.9
Crime content-10	167	159	95.2
Average CWII			95.8

The Education Word Identification Index (EWII) is defined as the ratio between the actual words detected and the words belonging to the education category. Table 8 is the Algorithmic analysis of EWII for content verification system (Education category).

The Sports Word Identification Index (CWII) is defined as the ratio between the actual words detected and the words belonging to the Sports category. Table 9 is the Algorithmic analysis of SWII for content verification system (Sports category).

Table 8. Algorithmic analysis of EWII for content verification system (Education category)

Content type	Words belonging to the education category	Actual words detected	Education Word Identification Index (EWII) in %
Education content-1	452	445	98.4
Education content-2	176	170	96.5
Education content-3	287	279	97.2
Education content-4	198	191	96.4
Education content-5	175	167	95.4
Education content-6	147	138	93.8
Education content-7	187	178	95.1
Education content-8	208	201	96.6
Education content-9	201	191	95.0
Education content-10	186	178	95.6
Average EWII			96

Table 9. Algorithmic analysis of SWII for content verification system (Sports category)

Content type	Words belonging to sports category	Actual words detected	Sports Word Identification Index (SWII) in %
Sports content-1	276	256	92.7
Sports content-2	156	145	92.9
Sports content-3	198	187	94.4
Sports content-4	134	127	94.7
Sports content-5	176	167	94.8
Sports content-6	208	198	95.1
Sports content-7	156	143	91.6
Sports content-8	181	176	97.2
Sports content-9	146	135	92.4
Sports content-10	198	176	88.8
Average SWII			93.4

The Politics Word Identification Index (PWII) is defined as the ratio between the actual words detected and the words belonging to Politics category.

Table 10 is the Algorithmic analysis of PWII for content verification system (Politics category).

Table 10. Algorithmic analysis of PWII for content verification system (Politics category)

Content type	Words belonging to Politics category	Actual words detected	Politics Word Identification Index (PWII) in %
Politics content-1	177	167	94.3
Politics content-2	198	182	91.9
Politics content-3	209	198	94.3
Politics content-4	457	428	93.6
Politics content-5	189	181	95.7
Politics content-6	146	140	95.8
Politics content-7	198	190	95.9
Politics content-8	267	261	97.7
Politics content-9	188	187	99.4
Politics content-10	156	140	89.7
Average PWII			94.8

The Entertainment Word Identification Index (ENWII) is defined as the ratio between the actual words detected and the words belonging to the entertainment category. Table 11 is the Algorithmic analysis of ENWII for content verification system (Entertainment category).

The Culture Word Identification Index (CUNWII) is defined as the ratio between the actual words detected and the words belonging to the culture category. Table 12 is the Algorithmic analysis of CUNWII for content verification system (Culture category).

Table 11. Algorithmic analysis of ENWII for content verification system (Entertainment category)

Content type	Words belonging to Entertainment category	Actual words detected	Entertainment Word Identification Index (ENWII) in %
Entertainment content-1	276	271	98.1
Entertainment content-2	156	150	96.1
Entertainment content-3	287	278	96.8
Entertainment content-4	145	139	95.8
Entertainment content-5	176	167	94.8
Entertainment content-6	134	129	96.2
Entertainment content-7	198	189	95.4
Entertainment content-8	124	119	95.9
Entertainment content-9	126	118	93.6
Entertainment content-10	165	156	94.5
Average ENWII			95.7

Table 12. Algorithmic analysis of CUWII for content verification system (Culture category)

Content type	Words in the field of entertainment	Actual words detected	Culture Word Identification Index (CUWII) in %
Culture content-1	165	161	97.5
Culture content-2	267	260	97.3
Culture content-3	167	156	93.4
Culture content-4	134	128	95.5
Culture content-5	198	190	95.9
Culture content-6	167	156	93.4
Culture content-7	187	171	91.4
Culture content-8	156	146	93.5
Culture content-9	209	189	90.4
Culture content-10	267	234	87.6
Average CUWII			93.5

Table 13. Comparative analysis of various parametric on content verification system with other works

Methods	Parametric evaluation in %					
	CWII	EWII	SWII	PWII	ENWII	CUWII
	Crime	Education	Sports	Politics	Entertainment	Culture
Proposed work	95.8	96	93.4	94.8	95.7	93.5
Yoram Timmerman et al. (2022)	94.1	95.1	92.9	93.9	93.7	93.1
Brautovic et al. (2020)	93.9	94.7	92.1	92.1	92.9	92.6
Chen et al. (2019)	93.2	94.1	91.7	91.8	92.6	92.1
Porlezza et al. (2012)	92.9	93.9	90.9	90.7	91.7	91.9

Table 13 is the Analytical computation of several parametric on pleased verification system with other works.

4.1. Dataset Description

The News Images Dataset consists of 41,763 news-related images collected from various online sources for image classification tasks. The dataset contains diverse visual content representing different news categories such as politics,

sports, business, technology, and entertainment under varying real-world conditions. The images capture a wide range of scenes, objects, and events, providing rich visual diversity for robust model training.

Unlike structured segmentation datasets, this dataset focuses on visual classification without pixel-level annotations, where each image is associated with a category

label. The dataset has been divided into 80% training data and 20% testing data to ensure effective model development and performance evaluation.

This dataset is widely used for news image classification and visual content analysis in machine learning applications.



Fig. 6 Sample image of news images dataset

Table 14. Performance of Comparison Model

Performance Metric	CNN (%)	RNN (%)	RF (%)	Proposed FCNN (%)
Accuracy	84.45	88.65	91.34	94.11
Precision	89.45	91.65	93.34	95.56
Recall	85.45	87.65	91.34	95.22
F1-Score	85.45	87.65	91.34	92.88
Recall Rate	20.45	15.65	10.34	05.60
Time Complexity	24.45	16.65	12.34	04.60
Computational Cost	85.45	88.65	90.34	94.10

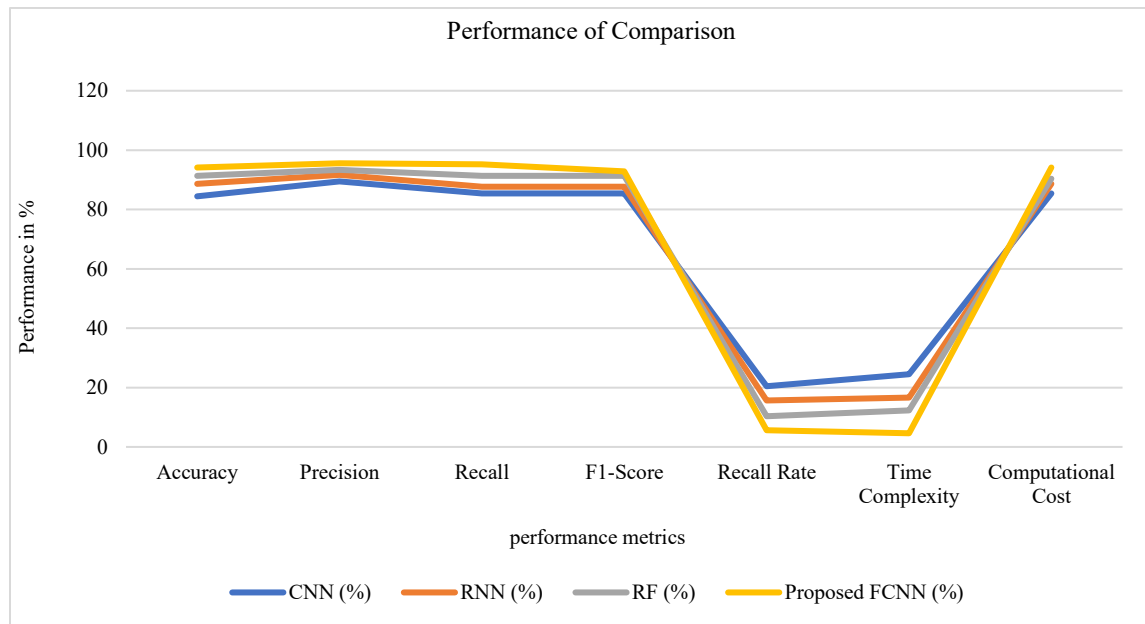


Fig. 7 Analysis in performance of comparison model

A thorough comparison of the four models-CNN, RNN, RF, and suggested FCNN for newspaper content categorization, utilizing a variety of performance evaluation indicators, is shown in Figure 6 and Table 14. The current CNN, RNN, and RF models obtained accuracy values of 84.45%, 88.65%, and 91.34%, respectively, indicating moderate performance in terms of precision, recall, and f1-score with gradual improvement in classification. However, these models showed increased recall values of 20.45%, 15.65%, and 10.34%, and time complexity values of 24.45%, 16.65%, and 12.34%, respectively, indicating poor performance in terms of false detection and time complexity. In contrast, the suggested FCNN model performed better in terms of achieved accuracy (94.11%), precision (95.56%), recall (95.22%), and f1-score (92.88%), while achieving lower recall (5.60%) and time complexity (4.60%), indicating high performance in terms of computational cost and time complexity. The suggested FCNN model performs better than the current model in terms of accuracy and efficiency, according to the aforementioned findings.

5. Conclusion

In this paper, Deep Learning algorithm-based newspaper content classification system has been proposed with a data retrieval method. The proposed automatic newspaper content classification system obtains 95.8% CWII for crime content, 96% EWII for Education content, 93.4% SWII for sports content, 94.8% PWII for politics content, 95.7% ENWII for

Entertainment content and 93.5% CUWII for culture content. In future, similar methodologies stated in this paper can be used for content classification in social media information. In summary, a robust FCNN model architecture with word-level classification, Gabor-based pre-processing, OCR-based word extraction, and an effective data retrieval algorithm utilizing Euclidean distance are all successfully integrated with the well-designed deep learning-based newspaper content classification framework.

Through enhanced feature representation, semantic comprehension, reduced noise, and improved multi-class retrieval, the suggested system guarantees high classification accuracy. Experimental evaluation of the suggested FCNN model reveals good accuracy of 94.11%, precision of 95.56%, recall of 95.22%, and total F1-score of 92.88%, confirming the strong discriminative ability. Additionally, the suggested model attained a lower recall rate of 5.60% and a low time complexity of 4.60%, confirming the reduced time complexity and false detections. Additionally, it is confirmed that the suggested approach outperforms the traditional CNN, RNN, and RF-based models in terms of high accuracy, stability, and efficiency. With an intelligent content organization, the suggested structure also guarantees the realization of scalable and real-time newspaper content analysis. To attain even greater performance and application, future work can concentrate on real-time data, lightweight models, and sophisticated semantic comprehension.

References

- [1] Muhammad Zubair Asghar et al., "Exploring Deep Neural Networks for Rumor Detection," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 4, pp. 4315-4333, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [2] Nicollas R. de Oliveira, Dianne S.V. Medeiros, and Diogo M.F. Mattos, "A Sensitive Stylistic Approach to Identify Fake News on Social Networking," *IEEE Signal Processing Letters*, vol. 27, pp. 1250-1254, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [3] Hridoy Sankar Dutta, and Tanmoy Chakraborty, "Blackmarket-Driven Collusion among Retweeters-Analysis, Detection, and Characterization," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1935-1944, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [4] Tao Jiang et al., "A Novel Stacking Approach for Accurate Detection of Fake News," *IEEE Access*, vol. 9, pp. 22626-22639, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [5] Yoram Timmerman, and Antoon Bronselaer, "Automated Monitoring of Online News Accuracy with Change Classification Models," *Information Processing and Management*, vol. 59, no. 6, pp. 1-22, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [6] Sandra Brautovic Maštrapa, and Sveučilište u Dubrovniku, "Accuracy in Online Media: Insufficient Journalistic Routines in Fact-checking and Corrections," *Media Studies*, vol. 11, no. 21, pp. 66-86, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [7] Yixin Nie, Haonan Chen, and Mohit Bansal, "Combining Fact Extraction and Verification with Neural Semantic Matching Networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6859-6866, 2019. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [8] Colin Porlezza, Scott R. Maier, and Stephan Russ-Mohl, "News Accuracy in Switzerland and Italy: A Transatlantic Comparison with the US Press," *Journalism Practice*, vol. 6, no. 4, pp. 530-546, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [9] Theodoros Alexakis et al., "A Machine Learning-based Method for Content Verification in the E-Commerce Domain," *Information*, vol. 13, no. 3, pp. 1-15, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [10] Jeelani Ahmed, and Muqem Ahmed, "Online News Classification using Machine Learning Techniques," *IJUM Engineering Journal*, vol. 22, no. 2, pp. 210-225, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [11] Hager Saleh, Abdullah Alharbi, and Saeed Hamood Alsamhi, "OPCNN-FAKE: Optimized Convolutional Neural Network for Fake News Detection," *IEEE Access*, vol. 9, pp. 129471-129489, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]

- [12] Shuyin Li, and Yang Liu, "News Video Title Extraction Algorithm based on Deep Learning," *IEEE Access*, vol. 9, pp. 12143-12157, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [13] Shahzada Daud et al., "Topic Classification of Online News Articles using Optimized Machine Learning Models," *Computers*, vol. 12, no. 1, pp. 1-16, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [14] Zhu, Yunlong. "Research on News Text Classification based on Deep Learning Convolutional Neural Network," *Wireless Communications and Mobile Computing*, vol. 2021, no. 1, pp. 1-6, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [15] Pin Wang, En Fan, and Peng Wang, "Comparative Analysis of Image Classification Algorithms based on Traditional Machine Learning and Deep Learning," *Pattern Recognition Letters*, vol. 141, pp. 61-67, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [16] Tamim Al Mahmud, Sazedra Sultana, and Antara Mondal, "A New Technique to Classification of Bengali News Grounded on ML and DL Models," *International Journal of Computer Applications*, vol. 185, no. 18, pp. 15-21, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [17] Dhiren Rohera et al., "A Taxonomy of Fake News Classification Techniques: Survey and Implementation Aspects," *IEEE Access*, vol. 10, pp. 30367-30394, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [18] Tanzir Hossain et al., "BanglaNewsClassifier: A Machine Learning Approach for News Classification in Bangla Newspapers using Hybrid Stacking Classifiers," *PLoS One*, vol. 20, no. 6, pp. 1-31, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [19] Chong Feng et al., "News Recommendation Systems-Accomplishments, Challenges and Future Directions," *IEEE Access*, vol. 8, pp. 16702-16725, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [20] Dong-Lin Li, Shih-Kai Lee, and Yin-Ting Liu, "Printed Document Layout Analysis and Optical Character Recognition System based on Deep Learning," *Scientific Reports*, vol. 15, no. 1, pp. 1-15, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [21] Zhenjie Wu, Weilan Wang, and Hongrui Li, "YOLOv10-CBRC: A High-Precision Document Image Layout Analysis Model," *Journal of King Saud University Computer and Information Sciences*, vol. 37, no. 6, pp. 1-21, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [22] Saifullah et al., "The Reality of High Performing Deep Learning Models: A Case Study on Document Image Classification," *IEEE Access*, vol. 12, pp. 103537-103564, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [23] Saifullah Saifullah et al., "DocXClassifier: Towards a Robust and Interpretable Deep Neural Network for Document Image Classification," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 27, no. 3, pp. 447-473, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [24] Vidhya Barpha, and Pramod S. Nair, "A Smart Model to Detect Hindi Fake News for Social Media Platform using Hybrid Deep Learning," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 5s, pp. 486-493, 2024. [[Publisher Link](#)]
- [25] Monika Choudhary et al., "BerConvoNet: A Deep Learning Framework for Fake News Classification," *Applied Soft Computing*, vol. 110, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [26] Abeer Saber et al., "A Novel Deep-Learning Model for Automatic Detection and Classification of Breast Cancer Using the Transfer-Learning Technique," *IEEE Access*, vol. 9, pp. 71194-71209, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]