Received: 19 January 2025 | Revised: 18 March 2025 | Accepted: 7 May 2025 | Published online: 8 September 2025

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2025, Vol. 00(00) 1–11
DOI: 10.47852/bonviewJCCE52025256

BON VIEW PUBLISHING

Enhanced Multimodal Webpage Classification Using Deep Learning for Efficient Information Retrieval

Manjunath Pujar[1], Monica Mundada[1], Sowmya B. J.[2], Supreeth Shivashankar[3,*], Ganesh Dalappagari Ramanjinappa[4] and Shambulingana Gouda[5]

[1] *Department of Computer Science and Engineering, M.S. Ramaiah Institute of Technology, India*

[2] *Department of Artificial Intelligence and Data Science, M.S. Ramaiah Institute of Technology, India*

[3] *School of Computer Science and Engineering, REVA University, India*

[4] *School of Computing and Information Technology, REVA University, India*

[5] *Electrical and Electronics Engineering, Rao Bahadaur Y. Mahabaleshwarappa Engineering College, India*

Abstract: Web data mining has become a crucial tool for efficiently retrieving valuable information, as users increasingly rely on the World Wide Web for data exchange. Traditional web classification methods often struggle with handling multimodal data, leading to challenges in accurately classifying diverse web contents. Online classification plays a key role in facilitating efficient retrieval of information from multimedia content. This study presents a novel multimodal approach for webpage classification by integrating deep learning techniques for audio-visual analysis. The personalized Long Short-Term Memory (LS) TM model, which is a specific version of Long Short-Term Memory (LSTM), has improved classification accuracy by combining deep audio and video features. Artificial Convolutional Neural Networks (A-CNNs) extract complex audio features, while transformer networks capture long-range dependencies from video data. The present study proposes a log-sigmoid activation function that provides a more flexible thresholding method in logistic regression, thus greatly improving the classification performance. The focus of this study, which is on single-modality classification, presents an innovative method of integrating deep learning-based multimodal fusion, thus setting a new standard for web classification. Experimental results show that the Logistic Sigmoid Long Short-Term Memory ((LS)^{2}TM) model has an accuracy of 88.09%, sensitivity of 89.14%, and specificity of 89.01%, outperforming state-of-the-art techniques such as LSTM, Deep Belief Network DBN, and A-CNN. The model also enhanced its precision (93.15%), recall (92.84%), and F-measure (93.25%), which are generally 5% higher than classical control methods. These findings highlight the potential of (LS)^{2}TM for improving web content mining through multimodal analysis. Future research should focus on the real-world validation and scalability of dynamic web environments.

Keywords: classifying web content, deep learning, efficient information retrieval, extracting key frames, Logistic Sigmoid Long Short-Term Memory model

1. Introduction

The Internet is presently the most extensive source of knowledge, allowing individuals to upload and retrieve boundless data almost immediately through search platforms, such as Google, Bing, Yahoo, and others. It showcases the unique traits of web data, such as being unlabeled, decentralized, diverse (mixed media), semi-organized, dynamic, and high-dimensional. These characteristics contribute to the rapid increase in Internet traffic, making information retrieval a challenging task. Extracting meaningful understanding and patterns from web link frameworks, page materials, and usage information through data analysis is referred to as web data discovery.

Web harvesting activities are classified as web framework extraction, web usage mining, and web content mining. Web content extraction is a developing subdiscipline that utilizes data analysis methods to derive knowledge from materials such as text, images, videos, and sound. Approaches to web content mining are classified into rule-based and machine-learning-based methods. Rule-based models struggle to generalize when predicting unseen content formats, whereas machine learning-based models are robust against unfamiliar content formats. However, the explosive growth of webpages and the vast volume of information on them make retrieving relevant results complex and tedious, leading to irrelevant search outputs and low productivity. Mining multimedia content is likely to be computation-intensive, as well as require a significant-sized training dataset. Finding useful content from this extensive multimedia database has become an urgent topic of concern.

This study is prompted by the demand for a more efficient multimodal webpage classification framework that will adequately process webpages with diverse formats of data including video, audio, and text in order to enhance the accuracy of information retrieval. Conventional classification methods cannot effectively process webpages with extensive multimedia content, causing inefficiency in search outputs and content categorization. This study seeks to close this gap through the implementation of deep-learning methods for effective classification.

Multimodal classification can be problematic considering the complex structure of multimedia content, computational demands, large training data requirements, and the needs to build an adaptive model to support dynamic data, which involves balancing accuracy against computational resources. The current methodology proposes

*Corresponding author: Supreeth Shivashankar, School of Computer Science and Engineering, REVA University, India. Email: supreeth.s@reva.edu.in multimodal classification to improve the classification of Webpages, by applying multimodal analysis, text, audio, and video analysis, for more comprehensive feature selection to improve classification performance through comparative context to capture richer contextual information of the site in the community. The proposed methodology also has ethical and safety measures to promote fair and transparent classification of web data, to limit biases or miss guided assumptions as well as policies to supports alignment with respect to anonymity of the participants. The clustering achieved better clustering and classification than existing literature focused on state of the art, thus shown a different and stronger solutions for analysis of web data content with the model.

Multimodal classification is a complex process that requires balancing accuracy with computational resources. This method intends to enhance the classification of web pages, with multimodal approach, text, audio, and video analysis. This global feature selection increases the expressive power of the classification model by incorporating richer context information. The approach also outlines ethical and safety considerations against fair and transparent classification, minimizes biases and ensures anonymity. Clustering was able to provide the best clustering and classification ever done by the past evidence, making it as strong solution to analyze web content data with the model. The approach includes ethical and safety considerations to maintain participant anonymity.

The proposed Logistic Sigmoid Long Short-Term Memory ((LS)²TM) model overcomes the limitations of existing approaches, such as Long Short-Term Memory (LSTM), Deep Belief Network (DBN), and Artificial Convolutional Neural Network (A-CNN). The (LS)²TM model fuses an A-CNN for audio feature extraction and transformer-based networks for greater temporal analysis was the superior architecture. Since traditional approaches for classifying web pages primarily rely on textual analysis and/or metadata analysis to classify traditional documents, their capacity to classify multimedia is hindered by their ability to successfully analyze these limiting conditions. This study proposed a new multimodal webpage genre classification framework, utilizing transformer-based networks for video analysis, and the A-CNN for audio feature extraction, operating in a single (LS)²TM data structure framework for the improvement of web content mining systems. Using a novel framework enabled improved temporal classification accuracy and greater potential for deeper semantic representations, moving away from categories that were previously stitched together from single-modality processing, relying on conventional deep-learning models. The proposed model combines a new log-sigmoid activation function, refines decision boundaries in logistic regression-based classifiers, and develops a new way to fuse multimodal data systematically, as well as collaboratively. This framework set about to develop a new decidedly optimized system to approach the key issues of web content mining, particularly the limitations regarding feature extraction, semantic representations, and refining decision boundaries across contexts.

1) Audio retrieval is performed by extracting block-level descriptors and I-vector representations, whereas video data is analyzed using aesthetic visual features (AVF) combined with attributes derived from deep learning models.
2) It implemented an upgraded clustering algorithm to improve similarity matching and identify related multimedia content more accurately.
3) The feature selection process is enhanced by implementing an efficient algorithm that minimizes computational time and effectively tackles scalability issues in large-scale data setting.
4) To enable effective web content mining, the system uses advanced machine learning (ML) and deep learning (DL) algorithms to evaluate audio data, video data, and their multimodal combinations, permitting full content understanding.

This paper is structured as follows. Section 2 provides a thorough overview of previous studies on web content extraction, multimodal analysis, and categorization techniques. Section 3 outlines the methodology for multimodal webpage classification, detailing the feature extraction techniques and the DL framework used in the study. Section 4 discusses the experimental setup and applied evaluation criteria. The results are presented in Section 5. Finally, Section 6 summarizes the findings and provides recommendations for future research.

2. Literature Review

In recent years, many methods have been introduced for classifying webpages, but traditional techniques still face significant challenges-especially when it comes to managing unstructured multimedia content, scaling to large datasets, and filtering out irrelevant or noisy features. In this section, we thoroughly review these existing approaches, point out their main shortcomings, and describe how our proposed framework addresses these gaps to deliver a more effective and reliable solution.

Markkandeyan and Indra Devi [1] proposed a machine learning model for website page classification which implements Principal Component Analysis (PCA). As noted previously, the technique achieved 97% accuracy on some course-related datasets. However, the method's dependence on examining features in a linear fashion poses a challenge. This might reduce its effectiveness in handling complex, nonlinear data structures which are typical in contemporary web pages containing diverse elements.

López-Sánchez et al. [2] proposed a web mining technique that incorporates deep neural networks (DNN) and transfer learning for image retrieval and webpage classification. However, the method's multimodal processing capabilities—multimedia data like text, audio, video and images—come short of fully utilizing the complexity characteristics of web content, making it unsuitable for advanced, resource-rich multimedia webpage classification.

Afzal et al. [3] proposed an approach for web video classification that applies category specific classifiers to predict content categories for each video. Their method incorporates visual features, along with content-based and context-aware classifiers. However, the framework relies heavily on visual information, neglecting other sources like audio, which can severely impact performance in multimodal settings.

Li et al. [4] introduced a webpage classification framework by exploiting the structure of a semantic network, such as Wikipedia and WordNet, combined with a Bayesian classifier. Although the method showed impressive scalability and robustness, it is mainly based on semantic extraction, which can not work for multimedia or adapt to all kinds of web data live on web. Furthermore, the method fails to incorporate multimodal features from various sources which may be useful to improve the accuracy of classification.

Hernández et al. [5] presented Classification Link Automatically (CALA), an automatic webpage classification technique based on URL pattern analysis. This approach achieved a high precision rate of 98% and recall of 90%. Nonetheless, its efficiency is restricted when applied to dynamic and multimodal web content because it ignores rich media elements such as images and videos, making it ineffective for multimedia-intensive webpages.

Ali et al. [6] introduced a fuzzy ontology-based approach using Support Vector Machines (SVM) for web content classification, in particular adult and medical web sites. Although the antipedophilic system performs well in identifying adult content, it cannot adapt to the rapid change in the content of the web site, in particular, with regard to complex web pages with rich content, like a multimedia one.

The proposed multimodal webpage genre classification framework leverages the synergy between audio, video and deep features for scalability, accuracy and robustness improvements. This holistic approach is a major development for web content mining, and provides an energetically sound basis for classification of the typical, multimedia-rich, current-day websites; we have now effectively bridged the gap between previous methods and the glamorous multimedia technology with regard to the classification of web pages.

3. Proposed System

The designed system is divided into three basic phases for enabling effective webpage categorization: categorization against audio content, video content, and a combination of audio and video content. Processing is initiated through the Web Classification URL Dataset consisting of website URLs with their associated categories. On reaching the preprocessing phase, all redundant entries are removed and only live URLs are picked up. Later, web content is accessed based on this upgraded dataset. For the automatic extraction of massive web page data and conversion of unstructured information to a structured one, Python library Beautiful Soup is utilized [7].

The proposed framework for interference webpage classification involves three steps: audio content analysis, video content analysis, and combined audio and video content analysis. CD-PAM is preferred for audio content analysis due to its higher accuracy. Log Radial-You Only Look Once Version 4 (LR-YOLOV4) optimizes grid levels for object detection. The Brownian Motion-based Blue Monkey–Complex Wavelet Recurrent Neural Network algorithm improves classification accuracy. Iterative Rank-based Hierarchical Optimization and Jaccard distance min–max K-means enhance clustering quality. The Logistic Sigmoid Long Short-Term Memory method exhibits balanced performance and computational economy. Using fitness against iteration [8], the system was tested against earlier studies including cockroach, lion, and dolphin swarm optimization methods. This method offers a consistent basis for good webpage classification. Classification algorithms including data privacy and bias mitigating methods were used in response to ethical questions. The model anonymizes the gathered data therefore prevents the storage and use of private data without authorization. The dataset was created from many categorical subgroups so that bias risk was reduced. Undercovering class imbalance problems using under sampling and fairness techniques helped to improve classifier fairness.

The proposed webpage classification algorithm uses an A-CNN and an ANOVA radial basis kernel to improve classification results and eliminate errors. It is compared with existing deep learning methods like convolutional neural network (CNN) and DBN based on accuracy, exactness, and sensitivity. The video is segmented into frames using Taxicab Distance-based Fidelity and LR-YOLOV4 technique. The feature extraction phase uses a Brownian Motion-based Blue Monkey–Complex Wavelet Recurrent Neural Network to optimize weight parameters and improve classification performance. The Complex Wavelet Radial Basis kernel is used to increase the accuracy and reduce the parameter complexity [9].

In the Third Phase (Audio and Video Content), after extracting the audio and video files, frequency conversion is performed using a Discrete Convolution Fast Fourier Transform to analyze the frequency of the audio files. The next step is pre-processing, where sampling denotes the procedure of transforming a continuous signal into a series of separate values, whereas quantization involves mapping those values from a potentially infinite range to a smaller, countable set of output values. Next, features such as the Mil Spectrogram, Spectral Contrast, Spectral Centroid, and Spectral Bandwidth are extracted.

The video file undergoes pre-processing to improve frame interpolation, keyframe extraction, and image contrast. The Distance-based Shot Reconstruction Degree technique extracts keyframes, whereas WTHE-PCE enhances the contrast. The Adaptive Self-Guided Filter-based Codebook algorithm enhances background and foreground separation. Objects were detected using Poisson Distribution-You Only Look Once V4 (PD-YOLOV4) to differentiate the URL link genres. The Poisson Distribution function reduces the grid levels, resulting in an accurate Bounding Box prediction and reduced loss function [10].

The suggested effort comprises three main phases for efficient webpage classification: audio content-based, video content-based, and combined audio and video content-based classifications. The system starts with the Web Classification URL Dataset, which contains the website URLs and their categories. URL preprocessing removes duplicates and identifies active links. Web content extraction was performed using the URL dataset. The Beautiful Soup package in Python automates the extraction of large volumes of data from websites and converts unstructured data into structured format. Next, from the optimized selected features, similar media are identified using Jaccard Distance Min Max–KMeans (JDMM-KMeans). To improve the clustering accuracy and reduce the clustering time, the Jaccard distance and min–max technique will be applied in the existing k-means algorithm.

The proposed algorithm uses $(LS)^2TM$ to classify clustered extracted features and their classes for webpage classification. The LSA function is used to improve categorization precision and reduces parameter complexity. The algorithm will be compared with existing deep learning techniques [11], such as CNNs and DBNs, focusing on accuracy, precision, and recall. To improve transparency and reproducibility, the availability of the trained model and its implementation details are clearly stated. If applicable, the trained model and code will be made publicly accessible through a dedicated repository, thus enabling further research and validation. Finally, the test URL was provided to the trained system to assess its performance. In this process, 80% of the dataset will be utilized for training, whereas the remaining 20% will be set aside for evaluation. The proposed structure is shown in Figure 1.

4. Experimental Setup

4.1. Dataset details

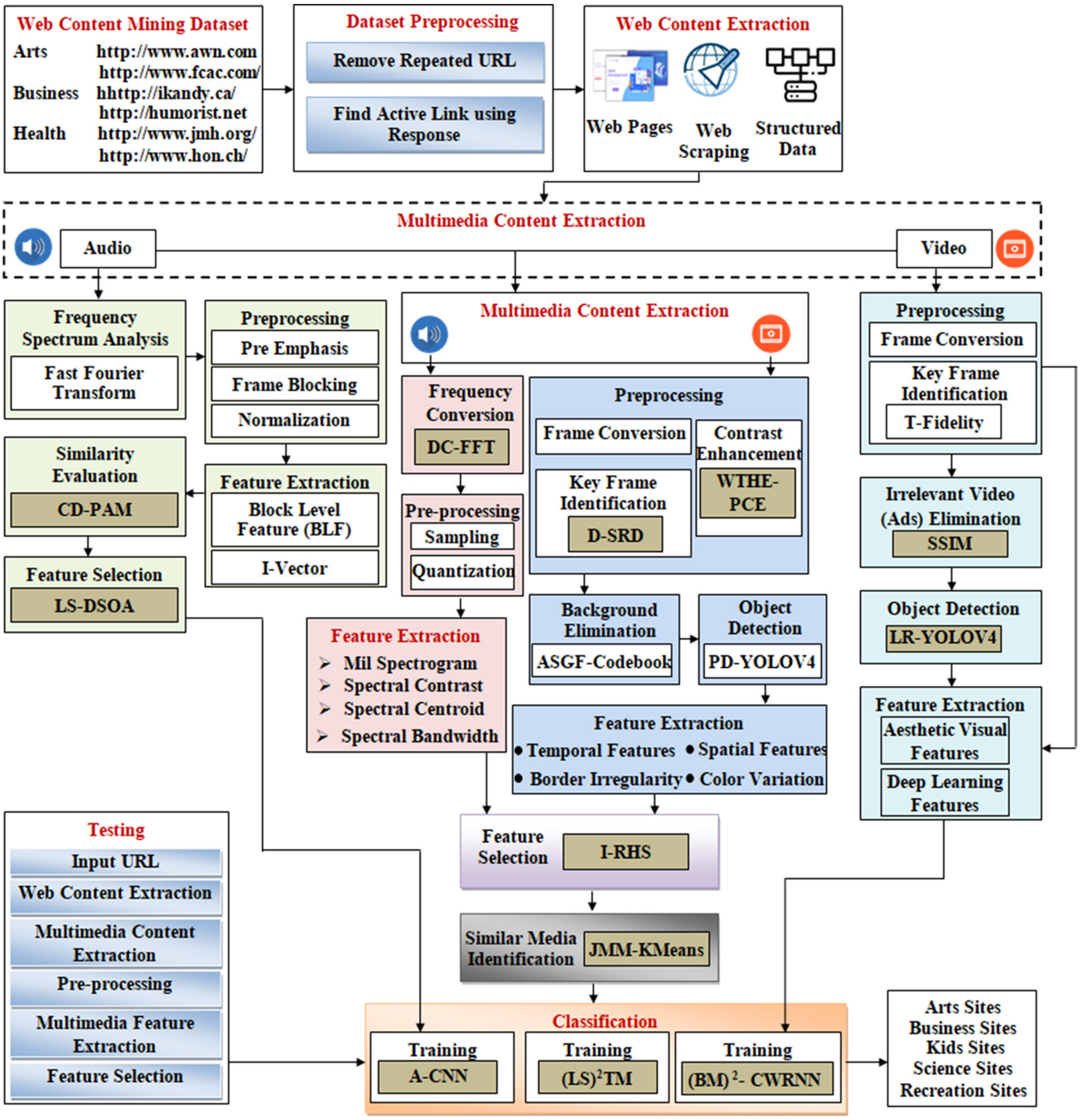
In our study, we used a publicly available dataset on the Internet. The dataset name is “Website classification using URL” from Kaggle [12].

The proposed work was implemented using Python, a widely adopted general-purpose programming language. Python prioritizes code readability and its simple syntax allows developers to convey ideas with fewer lines of code. This facilitates rapid development and efficient system integration. The use of a virtual environment tool creates an isolated Python space (within a directory) that is completely independent of the system-wide Python setup, thereby ensuring a clean and controlled development environment.

This paper presents a Python-based pre-processing system that uses CNNs for image data and LSTM networks for time-series data, addressing the complexities of multimodal datasets. The system uses a background subtraction algorithm to isolate foreground elements in video scenes, and integrates a logistic sigmoid with LSTM models to extract video-level features. Multimodal data fusion enhances recognition capabilities, and key frame extraction techniques ensure consistent handling across formats [13, 14].

The model architecture consisted of multiple layers, including an input layer, hidden layers with activation functions, and an output layer designed for a specific task. The choice of activation functions, such as ReLU for hidden layers and softmax for classification tasks, ensures nonlinearity and efficient gradient propagation. Additionally,

Figure 1
Proposed improved web content mining system



batch normalization and dropout layers were incorporated to improve generalization and mitigate overfitting.

4.2. Pre-processing

Many classification models require pre-processing steps, such as handling missing data, cropping images, and reducing noise. This data-cleaning phase often requires specialized knowledge and can be challenging to apply universally. However, multimodal classification provides greater flexibility, as each modality can be processed separately. For example, multimedia data may require unique approaches for cropping, scaling, and noise filtering. In some cases, pre-processing can be skipped for one or more modalities, allowing raw data to be used instead. Because this approach is often dependent on the user and the specific domain, the pre-processing stage should be evaluated on a case-by-case basis [15, 16].

In terms of hyperparameter tuning, various optimization techniques, including grid search and random search, have been explored to identify the best configuration. The learning rate, batch size, and number of epochs were fine-tuned based on the empirical analysis and validation performance. Moreover, adaptive optimizers such as Adam were utilized to accelerate convergence and improve model stability. These considerations collectively contribute to enhancing the accuracy, robustness, and computational efficiency of the model.

Multimodal classification models leverage feature selection in different ways, including through manual feature engineering, deep learning techniques, and integration within the classifier algorithm itself. This process can be performed individually for each modality or as part of a multistep approach in the overall model design [17]. Deep learning methods such as CNNs are commonly employed for extracting features, with the added benefit of performing classification in parallel. Classifiers such as Random Forests contribute to feature selection by determining key decision boundaries during the creation of decision trees.

4.2.1. Background and foreground modeling

Background modeling methods are crucial for identifying foreground objects in dynamic scenes owing to the complexity of video sequences and application environments. They are based on the corner frequency, descriptions, and correlation matrix between previous and current frames. Detection involves comparing current video frames with the background model in terms of color and brightness differences [18, 19]. If incoming pixels satisfy specific criteria, they are categorized as background, characterized by color distortion below the detection limit and luminance within the designated codeword range. Otherwise, they are classified as foreground, as illustrated in Figure 2.

This procedure is swift and shows negligible fluctuation in identification compared with the likelihood estimate. The action to deduct an arriving pixel value xxx in the evaluation set is defined as:

Algorithm for background subtraction

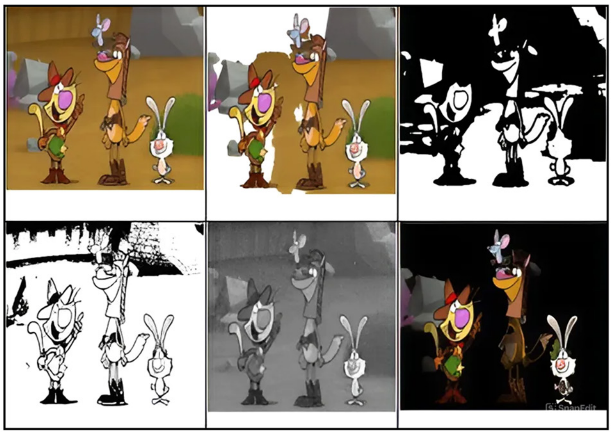
```
1) $X = (R, G, B),\ I - \sqrt{R^2 + G^2 + B^2}$
2) For all codewords in the equation, find the codeword $c_m$ matching to $x$ based on two conditions:
   a. colordist $(x, c_m) \leq \acute{\varepsilon}_2$
   b. brightness $(I, (I_m, I_m)) = \text{true}$
   c. Revise the corresponding codeword as described in Step II of the codebook formation process.
3) $BGS(x) = \begin{cases} foreground & if\,not\,match \\ background & otherwise \end{cases}$
```

If no valid coterminous codeword is found, then the pixel is identified as the foreground. Otherwise, it was categorized as a background.

4.2.2. Logistic Sigmoid Long Short-Term Memory ((LS)2TM)

Logistic regression is a probabilistic model that classifies instances based on their probabilities. Because classification relies on probabilities, an intuitive approach to optimize the parameters is to maximize the predicted probability of the correct class for each training example [20, 21].

Figure 2
The left column: Original images, the middle column: Color-labeled non-permanent backgrounds, the rightmost column: Detected foreground



Logistic regression is considered to be a reference point for processing video-level attributes. Given the averaged RGB value for all frames as a video-level feature xi of testing video *i* and the probability of entity *j* as $\_\left(w_T^j x_i\right)$, we trained this model by reducing the overall log loss of the training dataset with fixed parameter weights *w*. Here, the sigmoid function _(*x*) is given as:

$$L(w_j) = \lambda \|w_j\|_2^2 + \sum_{i=1}^{N} L\left(y_i, \sigma\left(w_j^T x_i\right)\right) \tag{1}$$

and the cross entropy loss function is:

$$\sigma(x) = \frac{1}{1+\exp(-x)} \tag{2}$$

1) $\sigma(x)$: Sigmoid function that converts input *x* to a probability value between zero and one.

and the cross entropy loss function is:

$$L(\hat{y}) = \frac{1}{N}\sum_{i=0}^{N} y_i log\left(\grave{y}_i + \epsilon\right) + (1 - y_i) log\left(1 - \left(\grave{y}_i + \epsilon\right)\right) \tag{3}$$

1) $\grave{y}_1$: Forecasted likelihood for category *i*.
2) y_i: The true label of category *i*.
3) ϵ: A minor value included for computational stability.

Typically, it involves a linear transformation of the video-level features into label space, followed by a sigmoid function that converts the log values into probabilities.

4.2.3. Long Short-Term Memory (LSTM)

LSTM is a type of RNN architecture designed for capturing and modeling successive data with long-range dependencies. In multimedia applications, audio features are extracted using a CNN–Long Short-Term Memory (CNN–LSTM) network, whereas video features are processed using the corresponding LSTM. In this study, we developed a classifier that takes a series of observations $(x(1), \ldots, x(t))$ and utilizes hidden states to predict multimedia labels y and $c^{(t)}$ (where $c^{(t)}$ refers to the cell state at time t). LSTM can incorporate or eliminate data from the cell state, enabling adaptive modifications to the model at each sequence step, as demonstrated in the formula.

The sigmoid layer acts as a discard gate layer, determining which data will be eliminated from the cell state. It looks at $h^{(t-1)}$ and $x^{(t)}$ and generates a value between 0 and 1 for each element in the cell state $c^{(t-1)}$.

$$g_l^{(t)} = \sigma\left(W_l^g x_{(l-1)}^{(t)} + W_l^g h_l^{(t-1)} + b_l^g\right) \tag{4}$$

1) $g_l^{(t)}$: Dismissal unit at moment t for layer l.
2) W_l^g: Matrix of coefficients for the dismissal unit.
3) $x_{(l-1)}^{(t)}$: signal at time t for layer l-1.
4) $h_l^{(t-1)}$: Concealed state from the preceding moment in layer l.

In this context, *t* represents the position within a sequence phase, while the sequence length is denoted by *t* for various video formats like F4V, FLV, and MP4The entry portal layer, functioning as a logistic activation, identifies which data points will make alterations to within the framework. Meanwhile, a tanh layer generates a vector of new data values, which could potentially be incorporated into the state, enriching the model with relevant information at each time step [22].

$$f_l^{(t)} = \sigma\left(W_f f x h_{(l-1)}^{(t)} + W_f f h h_l^{(t-1)} + b_{\uparrow}^{f}\right) \tag{5}$$

1) $f_l^{(t)}$: Forget gate at time step t for layer *l*.
2) $W_f f$: weight matrix for the input gate.

A layer of memory cells can be calculated using the equations below $h_i^{(t)}$, where $hh_i^{(t-1)}$ indicates the preceding layer at the equivalent sequence step, perhaps an earlier LSTM module or the input $h_i^{(t)}$ and $hh_i^{(t-1)}$ represents the same layer equation.

$$o_l^{(t)} = \sigma\left(W_l o x h_{(l-1)}^{(t)} + W_l^{oh} h_l^{(t-1)} + b_l^{o}\right) \tag{6}$$

1) $o_l^{(t)}$: output gate at time step *t* for layer *l*.
2) $W_l o$: Matrix of weights for the output unit.
3) b^{o}_{l}: offset term for the output gate.

The closing step involved selecting the output. First, a sigmoid activation layer determines which segments of the cell state are passed through as the output. In these equations, the sigmoid (logistic) function is applied element wise as denoted by σ, the tanh function is applied element wise as denoted by the symbol $\odot$. Equations are as follows:

$$\begin{gathered} s_l^{(t)} = g_l^{(t)} \odot i_l^{(t)} + s_l^{(t-1)} \odot f_l^{(t)}, \\ h_l^{(t)} = \gamma\left(s_l^{(t)}\right) \odot o_l^{(t)} \end{gathered} \tag{7}$$

1) $s_l^{(t)}$: Memory cell state at time step t for layer l.
2) $i_l^{(t)}$: Input gate activation.
3) γ: Non-linearity function (e.g., tanh).
4) $\odot$: Element-wise multiplication.

Here, *g* denotes the input node, and employs a tanh activation function. The terms "i," "o," and "f" signify the input gate, output gate, and forget gate, respectively [23].

4.2.4. Audio and video feature extraction

The general architecture of the voice deep learning multimodal fusion feature extraction model was segmented into three key components: attribute extraction, attribute selection, and categorization. The feature extraction module, which serves as the core of the multidimensional video recognition model (MVRM), defines specialized extraction schemes for each modality. Audio features are extracted using a CNN-LSTM network, where the CNN captures spectral representations and the LSTM models temporal dependencies. Video features are extracted using an LSTM model that effectively learns sequential patterns and motion dynamics. Frequency spectrum analysis plays a key role in audio feature extraction, where spectrogram-based representations (e.g., MFCCs and STFT) help capture critical acoustic patterns. CNNs process these spectrograms to extract spatial-frequency features, whereas LSTMs analyzes temporal variations. Similarly, in the same manner, video feature extraction includes the Aesthetic Visual Features (AVF) of color, texture, dynamics of motion, spatial composition, and algorithmic layout that aid in understanding a given scene. The LSTM model interprets these AVFs for better accuracy in temporal interdependence yielding higher video classification precision.

In multimodal learning, data integration is considered one of the most important components, as it entails the fusion of data from various sources into a consolidated model. There are traditional approaches to fusion that comprise of early and late fusion. However, these terms within the context of multimodal integration seems too simplistic. All multimodal data is fed to the learning model in early fusion, thus enabling cross-modal interactions at the initial stage; whereas in late fusion, each modality is processed independently before their results are merged, thus preserving modality specific learning, as each modality accounts for learned data. By using a hybrid fusion strategy that incorporates both early and late fusion techniques of our proposed model, classification performance is further enhanced. Early fusion captures deep interdependencies between modalities and late fusion helps each modality retain its contributions. This combined approach significantly improves the classification accuracy, making it superior to unimodal or conventional fusion techniques [24, 25].

To ensure robustness against noise, irrelevant data, and missing modalities, our system incorporates multiple preprocessing and data-filtering mechanisms. For audio data, noise reduction techniques such as spectral subtraction and wavelet denoising are applied before feature extraction. Video data underwent frame selection and motion smoothing to minimize irrelevant information. Missing modality scenarios are handled using imputation techniques, where missing audio/video features are estimated using attention-based deep learning models or data augmentation methods. Our hybrid fusion strategy also provides redundancy, allowing the system to rely on the most informative modality when the other modality is unreliable or absent. These measures collectively enhance the reliability and generalization performance of the system in real-world scenarios.

4.2.5. Key frame extraction

The video downloaded over the Internet contains different types of videos: f4v, flv, and mp4. The approach under consideration disregards the format and streaming video layout to enhance the accessibility of the video key-extraction technique. Instead, the video was decoded before undergoing processed frame decomposition [26]. The original difference between the video frames was utilized to obtain the key-frame sequence. The subsequent step entails obtaining the essential frame sequence by analyzing the structural feature variations among the intermediate key frames. Finally, the assessment of the quantity of keyframes within the derived sequence ensures the preservation of their effectiveness, as illustrated in Figure 3.

4.2.6. The proposed Logistic Sigmoid and Long Short-Term Memory (LS)2TM model

In the recognition of video–audio, (LS)2TM is used. A schematic diagram of our (LS)2TM model is shown in Figure 4. (LS)2TM was used in this model to obtain the desired results. In the audio-only section, features were extracted in a manner similar to the audio model. A Logistic Sigmoid function was then applied using a CNN. This audio-visual feature extraction model was designed to capture both provincial and planetary feature representations from the primary visual and sound recordings within a video, with the goal of recognizing personality traits, as outlined below.

$$\mathrm{Y}^5 = \mathrm{R}\left(\mathrm{W}^5 {*} \mathrm{y}^4 + \mathrm{b}^5\right) \tag{8}$$

1) Y^5: Output of audio-visual feature extraction model.
2) W^5: Weight matrix for the audio-visual network.
3) y^4: Input audio features extracted from model.
4) b^5 Bias term for the network output.
5) R: Nonlinear transformation function (e.g., activation function).

In the video-only segment, the video features were extracted in a manner similar to that of the video model. Subsequently, the (LS)²TM model was constructed. In the audio–video configuration, there were 129

Figure 3
Multimodal fusion for audio–video feature extraction

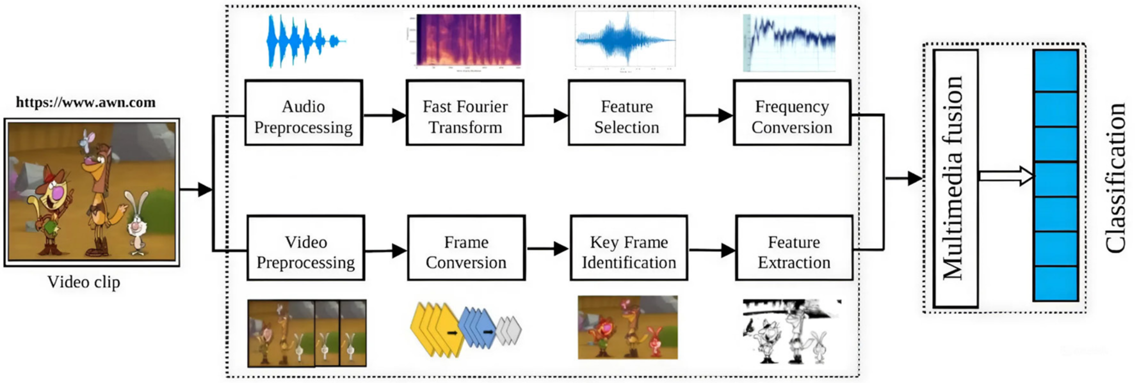
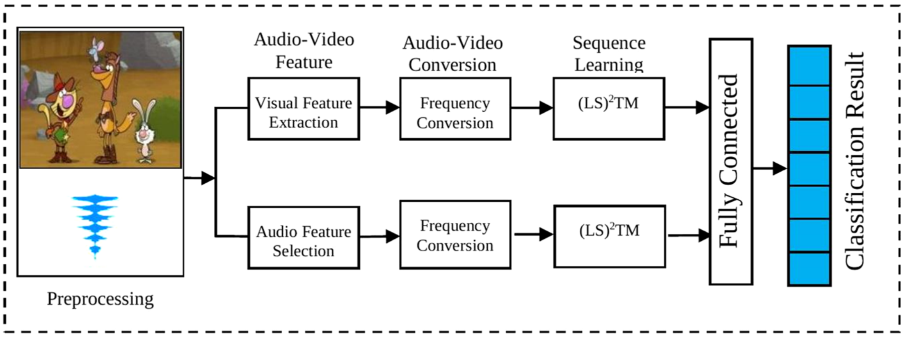


Figure 4
The architecture and proposed methodology of the (LS)^2 TM model



hidden units [27]. This (LS)²TM is then succeeded by a fully connected layer and the resulting equation derived from this layer is as follows:

$$Y_v = \mathrm{R}\left(W_2 a_1^{(t)} + b_2\right), \tag{9}$$

1) Y_v: Output of the video-based model.
2) R: Nonlinear transformation function (e.g., activation function).
3) W_2: Weight matrix of the final dense layer.
4) B_2: Bias term.

The response map from the first fully connected layer from the audio-only part is concatenated with the ingredient map from the first (LS)2TM layer from the video-only part in the implementation of the method "Integration of audio and visual features into the same operation."

$$\mathrm{A_c} = \left[a_1^{(t)}, y_2\right] \tag{10}$$

1) $\mathrm{a_c}$: Concatenated audio–video feature vector from both audio and video modalities.
2) $a_1^{(t)}$: Audio features from time step t.
3) y_2: Video features from the corresponding video modality.

To address these two issues, this study proposes a multimodal personality trait recognition method that integrates audio and video modalities based on an (LS)²TM architecture. As shown in Figure 3, the proposed method amalgamates different models, including (LS)²TM, a recently emerged transformer for learning high-level audio–video feature representations, followed by a decision-level fusion strategy for final personality trait recognition.

In audio characteristic distillation, the trained audio A-CNN version was used to analyze excessive-degree segment-level audio features. Video features are extracted from live films to separately learn excessive-stage image features from each body in animation succession. Subsequently, the extracted audio-visual features are input into an (LS)²TM framework to independently capture long-term temporal dynamics, thereby producing the final global audio and visual representations for subsequent tasks. This study used a linear regression method to discover six self-reliant character progressions based on audio and visual responsibilities [28, 29]. This study introduced a multimodal approach for character trait popularity analysis by integrating auditory and visual modalities. It is built upon a framework in which (LS)²TM and transformer architectures are merged to extract advanced spatio-temporal audio-visual feature representations, facilitating the recognition of personality traits.

4.2.7. Comparison of methodology and effectiveness

The comparison section has been significantly enhanced to provide a complete evaluation of the methodology in relation to the previously established techniques. Especially, the relationships in the updated content are linking the proposed approach with the existing methods in the area in a way that can be articulated more clearly to explain the contribution made by the framework.

The research survey reviews literature on multimodal personality trait recognition, including studies that use text, video, audio, or telematics data for personality assessment. In contrast, the "comparison targets" refer to established benchmark methods from prior research, which are used to evaluate the effectiveness of the approach presented in this paper. By comparing results in this way, the study not only integrates findings from the survey with these benchmarks but also provides a comparative narrative that highlights the strengths and weaknesses of different models.

To clearly distinguish the proposed method from existing approaches, the comparison includes a thorough evaluation of fundamental performance metrics such as accuracy, sensitivity, and specificity. These indicators help assess how well the classification model performs, offering insights into the types of conditions it can handle and the range of scenarios where it is applicable.

While the specific methodology does not have a promising outcome, its critique will definitely reveal some shortcomings. One of the most noticeable problems is the heavily synthetic datasets on which the method is based, which are rudimentary and might lack depth. These same issues combined with some other challenges stand to weaken the overall robustness and generalizability of the model in practical application contexts. Fluctuations in data streams, incomplete data, and real-world noise are other considerations that will impede the model’s operational efficiency. These challenges require validation on real world datasets, and multiple test environments in order to enhance study reliability. These explanations and some contextual reasoning will allow readers to appreciate the sophisticated logic behind the method outshining existing approaches to multimodal data fusion. A comprehensive analysis which is incorporated here creates space for a tacit but important discussion that strengthens the belief of applying the proposed framework for ubiquitous multi-camera surveillance systems.

5. Result and Discussion

In the present study, for tasks involving web scraping, machine learning model development, and page content retrieval through URLs, the Python language was utilized in this research. With the help of Urllib and Selenium, webpage content was extracted from the URLs and relevant webpages were created [19, 30, 31]. The Numpy library was used to accelerate mathematical calculations after the qualitative pre-processing step performed on texts from the websites.

This study presents a new method for personality detection using feature-level fusion. It uses audio and video data from (LS)^{2}TM and equipment networks to predict temperament traits. The model's performance is compared to state-of-the-art techniques, highlighting the trade-off between accuracy and computational cost. The new log-sigmoid activation function is implemented, allowing for better classification accuracy. The study emphasizes the need for a more structured and transparent data presentation.

5.1. Performance analysis of proposed classification technique

The proposed (LS)^{2}TM long- and short-time dependency relations, as well as the handling of gradient explosion difficulties, are some of the topics that the LSTM seeks to improve. Long-distance correlations in a sequence of variable lengths were captured using the LSTM model. The log-sigmoid activation function serves as an alternative to the conventional hard-sigmoid activation in LSTM models that interpret audio-visual information. Hard-sigmoid activation is applied to the LSTM cell forget, input, and egress gates. This activation method is preferable because it can easily calculate gradients [32, 33].

It is simultaneously validated with methods such as the DBN, A-CNN, Brownian Motion and Continuous Wavelet Analysis into Recurrent Neural Networks (BM-CWRNNs), along with LSTM, with respect to a variety of performance metrics, including responsiveness, exactness, and a combined measure of precision, F-measure, True Positive (TP), and False Positive (FP). Current methodologies are used in conjunction with a comparative evaluation to determine the effectiveness of the model.

To address the imbalanced class distributions, the system utilizes a combination of data augmentation, adaptive loss functions, and performance evaluation metrics. Data augmentation techniques were applied to enhance the representation of underrepresented web categories and ensure sufficient feature diversity. Additionally, the model employs class weighting during training, assigning higher importance to minority classes to balance the learning. The evaluation process considers precision, recall, and F1-score rather than accuracy, ensuring that classification remains effective across all categories despite the imbalance.

The baseline models, which included the CNN, DBN, BM-CWRNN, LSTM, and (LS)^{2}TM proposed model, were compared as fairly as possible by training and testing them under identical conditions. All baseline models were provided with a dataset in accordance with our prior work. Because of the similarity in training, validation and testing, a proportionality for the training, validation and testing was fixed for all baseline models. Normalization and data augmentation on all baseline models were applied in the same manner. In addition, all baseline models were compiled by using the Adam optimizer and trained at a constant learning rate of 0.001% using a 32-batch size. All baseline models were evaluated for their performance metrics of accuracy, precision, recall, and F1 score using the same set of formulas [34, 35] controlled to ensure that any discrepancies in performance could be attributed to learning performance of the baseline models rather than to experiment design.

To evaluate the effectiveness of the proposed (LS)^{2}TM model, we considered various performance metrics including accuracy, sensitivity (recall), and specificity. These are defined as follows.

Accuracy: Measures the overall correctness of the model's predictions

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

Sensitivity (Recall): Measures the proportion of actual positives correctly identified.

$$Sensivity(Recall) = \frac{TP}{TP+FN} \tag{12}$$

Specificity: Measures the proportion of actual negatives correctly identified.

(LS)^{2}TM as shown in Table 1 is a new method for web content mining, which can be applied to the classification of all website classes,

Table 1
Performance analysis of proposed (LS)^{2}TM based on accuracy, sensitivity, and specificity

	Performance metrics (%)		
Techniques	**Accuracy**	**Sensitivity**	**Specificity**
Proposed Logistic Sigmoid Long Short-Term Memory ((LS)^{2}TM)	88.09	89.14	88.01
LSTM	87.12	88.44	87.02
BM-CWRNN	87.34	87.54	86.05
A-CNN	86.15	87.25	89.12
DBN	88.09	89.14	88.01

which is more precise, specific, and sensitive than any other recently developed methods. The performance of the proposed method is 89.14%: sensitivity at 89.14%, 88.09%: accuracy at 88.9%, and 89.01%: specificity at 87.18%, 87.34% and 88.02%, which shows that the proposed (LS)^{2}TM method improves web content mining efficiency as well as to classify the website classes on the basis of video feature analysis [36].

This paper improves the classification performance through different models including, but not limited to, DBN, A-CNN, BM-CWRNN, LSTM, and (LS)^{2}TM, which improves categorization by embedding words and deep learning even if the URL information is closed. The proposed (LS)2TM model applies URL-based pull-out of important elements of URLs, to achieve more precise ranking. This method outperforms previous ranking metrics and effectively classifies websites based on domains or classes as shown in Figure 5.

It increases the classification performance using various classification models including DBN, A-CNN, BM-CWRNN, LSTM, and (LS)^{2}TM; improve the categorization of embedded words and deep learning even when the URL information is closed [37, 38]. An ablation study evaluates the impact of key components by comparing the performance with and without LS-DSOA, alternative optimization techniques, and different clustering methods. The significance of the audio and video analyses was assessed by selectively removing them. A tradeoff analysis between computational cost and accuracy is also provided, and the proposed (LS)2TM model uses a URL-based feature derivation strategy from URLs, aiming for better accuracy. The methodology outperformed previous metrics [39], effectively classifying websites based on domains or classes, as shown in Figure 5.

Additionally, precision and F1-score are critical in classification tasks:

Precision: Measures the proportion of correctly predicted positive instances.

$$Precision = \frac{TP}{TP+FP} \tag{14}$$

F1-Score: Provides a balance between precision and recall using their harmonic mean.

$$F1 - Score = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{15}$$

The evaluation of precision, recall, and F-measure for (LS)^{2}TM is presented in Table 2. The higher rates of precision, recall, and F-measure 93.15%, 92.84%, and 93.25%, respectively determined the model's value. The precision, recall, and F-measure rates of the current work, in comparison, range from 86.52% to 92%, 88% to 92.15%, and 89.24% to 93.34%, respectively. Comparing these rates with those of the suggested methodologies, they are lower. Consequently, (LS)^{2}TM streamlines details and improves robust web data mining. The proposed methodology has higher metric rates than those in previous works, as shown in Figure 6.

Figure 5
Graphical representation of the proposed (LS)^{2}TM based on accuracy, sensitivity, and specificity

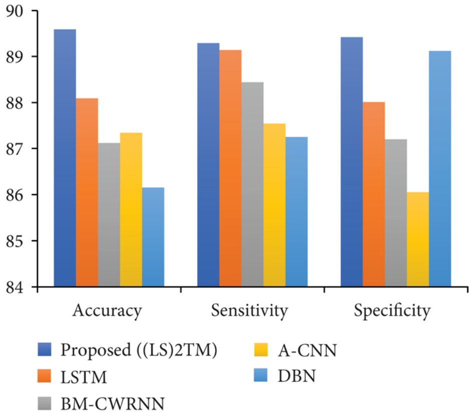


Table 2
Performance analysis of proposed BM2-CWRNN based on precision, recall, and F-measure

Techniques	Performance metrics (%)		
	Precision	Recall	F-measure
Proposed (LS)^{2}TM	93.15	92.84	93.25
LSTM	92.01	92.15	93.34
BM-CWRNN	91.85	92.24	91.22
A-CNN	87.20	89.21	90.28
DBN	86.52	88.05	89.24

The study analyzed multiclass classification models trained on webpage videos, with F1 scores displayed in Table 3. All models used the extracted images and appropriate hyperparameters. The ramification showed that the (LS)^{2}TM model outperformed existing models in terms of classification accuracy, indicating its superior performance.

By analyzing the models' training durations and contrasting them according to their individual F1 scores and training durations, it is clear that the (LS)^{2}TM model performs significantly better than the other models.

Figure 6
Graphical representation of the proposed (LS)^{2}TM based on precision, recall and F-measure

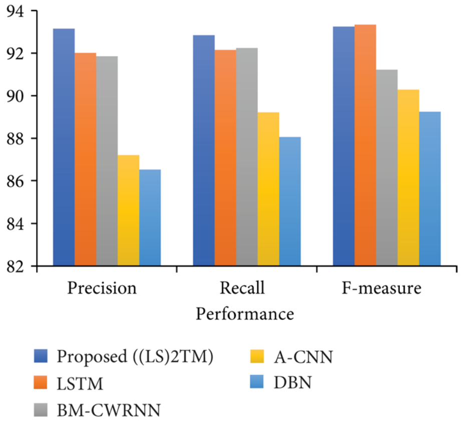


Table 3
Results of multiclass classification models

Models	F1 score
Proposed (LS)^{2}TM	90.12
LSTM	89.37
BM-CWRNN	88.46
A-CNN	87.24
DBN	85.86

6. Conclusions

This study presents a novel algorithm for key-frame extraction from videos, focusing on integrating video and audio web content for efficient webpage classification. The proposed (LS)²TM model demonstrates superior performance compared to existing models, such as DBN, BM-CWRNN, A-CNN, and LSTM. It achieved accuracy (88.09%), sensitivity (89.14%), and specificity (89.01%), with improvements of approximately 2.7% in sensitivity, 1.1% in specificity, and 1.2% in accuracy. Furthermore, the (LS)²TM model enhances classification metrics with precision (93.15%), recall (92.84%), and F-measure (93.25%), reflecting an increase of up to 5% over traditional methods. These outcomes demonstrate how well the model integrates multimodal data and achieves a high classification accuracy. However, this study's reliance on simulated datasets highlights the need for real-world validation using diverse datasets and dynamic environments. Future work should involve real-world case studies, performance evaluation across varied data, and inclusion of user feedback to enhance usability and scalability for real-time applications.

Acknowledgement

The authors acknowledge the support of MSRIT, REVA University, for the facilities provided to carry out the research.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available at https://www.kaggle.com/datasets/shaurov/website-classification-using-url. The data that support the findings of this study are openly available at https://doi.org/10.1109/78.650093, reference number [12].

Author Contribution Statement

Manjunath Pujar: Methodology, Software, Validation, Resources, Writing – original draft. **Monica Mundada:** Conceptualization, Formal analysis, Writing – review & editing, Project administration. **Sowmya B. J.:** Methodology, Software, Validation, Investigation, Data curation. **Supreeth Shivashankar:** Writing – review & editing, Supervision, Project administration. **Ganesh Dalappagari Ramanjinappa:** Data curation, Visualization. **Shambulingana Gouda:** Visualization, Supervision.

References

[1] Markkandeyan, S., & Indra Devi, M. (2015). Efficient machine learning technique for web page classification. *Arabian Journal for Science and Engineering*, *40*(12), 3555–3566. https://doi.org/10.1007/s13369-015-1844-1

[2] López-Sánchez, D., Arrieta, A. G., & Corchado, J. M. (2018). Deep neural networks and transfer learning applied to multimedia web mining. In *Distributed Computing and Artificial Intelligence, 14th International Conference*, 124–131. https://doi.org/10.1007/978-3-319-62410-5_15

[3] Afzal, M., Wu, X., Chen, H., Jiang, Y. G., & Peng, Q. (2016). Web video categorization using category-predictive classifiers and category-specific concept classifiers. *Neurocomputing*, *214*, 175–190. https://doi.org/10.1016/j.neucom.2016.06.004

[4] Li, H., Xu, Z., Li, T., Sun, G., & Choo, K. K. R. (2017). An optimized approach for massive web page classification using entity similarity based on semantic network. *Future Generation Computer Systems*, *76*, 510–518. https://doi.org/10.1016/j.future.2017.03.003

[5] Hernández, I., Rivero, C. R., Ruiz, D., & Corchuelo, R. (2014). CALA: An unsupervised URL-based web page classification system. *Knowledge-Based Systems*, *57*, 168–180. https://doi.org/10.1016/j.knosys.2013.12.019

[6] Ali, F., Khan, P., Riaz, K., Kwak, D., Abuhmed, T., Park, D., & Kwak, K. S. (2017). A fuzzy ontology and SVM-based web content classification system. *IEEE Access*, *5*, 25781–25797. https://doi.org/10.1109/ACCESS.2017.2768564

[7] Aung, S. S., & Kyu, Z. M. (2017). Modified codebook algorithm with Kalman filter for foreground segmentation in video sequences. In *International Conference on Signal Processing and Communication*, 332–336. https://doi.org/10.1109/CSPC.2017.8305864

[8] An, E., Ji, A., & Ng, E. (2017). *Large scale video classification using both visual and audio features on YouTube-8M dataset*. Stanford University. https://cs231n.stanford.edu/reports/2017/pdfs/702.pdf

[9] Salehin, I., Islam, M. S., Amin, N., Baten, M. A., Noman, S. M., Saifuzzaman, M., & Yazmyradov, S. (2023). Real-time medical image classification with ML framework and dedicated CNN–LSTM architecture. *Journal of Sensors*, *2023*(1), 3717035. https://doi.org/10.1155/2023/3717035

[10] Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

[11] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

[12] Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681. https://doi.org/10.1109/78.650093

[13] Hossain, S., Deb, K., Sakib, S., & Sarker, I. H. (2025). A hybrid deep learning framework for daily living human activity recognition with cluster-based video summarization. *Multimedia Tools and Applications*, *84*(9), 6219–6272. https://doi.org/10.1007/s11042-024-19022-0

[14] Zhou, Y., Zhang, Z., Wang, X., Sheng, Q., & Zhao, R. (2025). Multimodal archive resources organization based on deep learning: A prospective framework. *Aslib Journal of Information Management*, 77(3), 530–553. https://doi.org/10.1108/AJIM-07-2023-0239

[15] Pujar, M., Mundada, M. R., Sowmya, B. J., Supreeth, S., & Shruthi, G. (2023). An efficient framework for web content mining systems using improved CD-PAM clustering and the A-CNN technique. *SN Computer Science*, *4*(5), 692. https://doi.org/10.1007/s42979-023-02137-w

[16] Hong, D., Zhang, B., Li, X., Li, Y., Li, C., Yao, J., ..., & Chanussot, J. (2024). SpectralGPT: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(8), 5227–5244. https://doi.org/10.1109/TPAMI.2024.3362475

[17] Luo, S., & Yu, J. (2024). ESGNet: A multimodal network model incorporating entity semantic graphs for information extraction from Chinese resumes. *Information Processing & Management*,

61(1), 103524. https://doi.org/10.1016/j.ipm.2023.103524

[18] Kothuri, S. R., & RajaLakshmi, N. R. (2022). MALO-LSTM: Multimodal sentiment analysis using modified ant lion optimization with long short term memory network. *International Journal of Intelligent Engineering and Systems*, *15*(5), 326–335. https://doi.org/10.22266/ijies2022.1031.29

[19] Wu, X., Hong, D., & Chanussot, J. (2022). Convolutional neural networks for multimodal remote sensing data classification. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–10. https://doi.org/10.1109/TGRS.2021.3124913

[20] Alfartosy, H., & Khafaji, H. (2023). A new feature extraction, reduction, and classification method for documents based on Fourier transformation. *International Journal of Intelligent Engineering and Systems*, *16*(5), 586–597. https://doi.org/10.22266/ijies2023.1031.50

[21] Wang, C., Zhang, M., Shi, F., Xue, P., & Li, Y. (2022). A hybrid multimodal data fusion-based method for identifying gambling websites. *Electronics*, *11*(16), 2489. https://doi.org/10.3390/electronics11162489

[22] Chaithra, Gowdru Lingaraju, S., Sreenivasaiah, J. (2022). Classification of web pages using the machine learning algorithms with web page recommendations. *International Journal of Intelligent Engineering and Systems*, *15*(4), 637–650. https://doi.org/10.22266/ijies2022.0831.57

[23] Sun, W. (2022). Data mining in the big data era. In *Proceedings of the 2022 8th International Conference on Humanities and Social Science Research*, *664*, 2107–2111. https://doi.org/10.2991/assehr.k.220504.381

[24] Yin, X., & Chen, L. (2022). A cross-modal image and text retrieval method based on efficient feature extraction and interactive learning CAE. *Scientific Programming*, *2022*(1), 7314599. https://doi.org/10.1155/2022/7314599

[25] Belfedhal, A. E. (2023). Multi-modal deep learning for effective malicious webpage detection. *Revue d'Intelligence Artificielle*, *37*(4), 1005–1013. https://doi.org/10.18280/ria.370422

[26] Johnson, F., & Gupta, S. K. (2012). Web content mining techniques: A survey. *International Journal of Computer Applications*, *47*(11), 44–50. https://doi.org/10.5120/7236-0266

[27] Shahabaz, A., & Sarkar, S. (2024). Increasing importance of joint analysis of audio and video in computer vision: A survey. *IEEE Access*, *12*, 59399–59430. https://doi.org/10.1109/ACCESS.2024.3391817

[28] Sulun, S., Viana, P., & Davies, M. E. P. (2024). Movie trailer genre classification using multimodal pretrained features. *Expert Systems with Applications*, *258*, 125209. https://doi.org/10.1016/j.eswa.2024.125209

[29] Alsaedi, M., Ghaleb, F. A., Saeed, F., Ahmad, J., & Alasli, M. (2024). Multi-modal features representation-based convolutional neural network model for malicious website detection. *IEEE Access*, 12, 7271–7284. https://doi.org/10.1109/ACCESS.2023.3348071

[30] Raj, S., Agrawal, P., Saha, S., Singh, B., & Pedanekar, N. (2024). Optimizing movie selections: A multi-task, multi-modal framework with strategies for missing modality challenges. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, 910–918. https://doi.org/10.1145/3605098.3635947

[31] Ding, Y., Ren, K., Huang, J., Luo, S., & Han, S. C. (2024). PDF-MVQA: A dataset for multimodal information retrieval in PDF-based visual question answering. *arXiv Preprint: 2404.12720*. https://doi.org/10.48550/ARXIV.2404.12720

[32] Sowmya, B. J., Meeradevi, Alex, S. A., Kanavalli, A., Supreeth, S., Shruthi, G., & Rohith, S. (2024). Machine learning model for emotion detection and recognition using an enhanced Convolutional Neural Network. *Journal of Integrated Science and Technology*, *12*(4), 786. https://doi.org/10.62110/sciencein.jist.2024.v12.786

[33] Dentamaro, V., Impedovo, D., Musti, L., Pirlo, G., & Taurisano, P. (2024). Enhancing early Parkinson's disease detection through multimodal deep learning and explainable AI: Insights from the PPMI database. *Scientific Reports*, *14*(1), 20941. https://doi.org/10.1038/s41598-024-70165-4

[34] Yin, M., Lin, J., Wang, Y., Liu, Y., Zhang, R., Duan, W., ..., & Zhu, J. (2024). Development and validation of a multimodal model in predicting severe acute pancreatitis based on radiomics and deep learning. *International Journal of Medical Informatics*, *184*, 105341. https://doi.org/10.1016/j.ijmedinf.2024.105341

[35] Duy, P. T., Minh, V. Q., Dang, B. T. H., Son, N. D. H., Quyen, N. H., & Pham, V. H. (2024). A study on adversarial sample resistance and defense mechanism for multimodal learning-based phishing website detection. *IEEE Access*, *12*, 137805–137824. https://doi.org/10.1109/ACCESS.2024.3436812

[36] Ekenel, H. K., & Semela, T. (2013). Multimodal genre classification of TV programs and YouTube videos. *Multimedia Tools and Applications*, *63*(2), 547–567. https://doi.org/10.1007/s11042-011-0923-x

[37] Li, J., Qi, G., Zhang, C., Chen, Y., Tan, Y., Xia, C., & Tian, Y. (2023). Incorporating domain knowledge graph into multimodal movie genre classification with self-supervised attention and contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3337–3345. https://doi.org/10.1145/3581783.3612085

[38] Wang, J., Huang, J. X., Tu, X., Wang, J., Huang, A. J., Laskar, M. T. R., & Bhuiyan, A. (2024). Utilizing BERT for information retrieval: Survey, applications, resources, and challenges. *ACM Computing Surveys*, *56*(7), 185. https://doi.org/10.1145/3648471

[39] Choudhary, M., & Chouhan, S. S. (2025). Enhancing credibility assessment in online social networks using multimodal deep learning. *Applied Soft Computing*, *171*, 112796. https://doi.org/10.1016/j.asoc.2025.112796

How to Cite: Pujar, M., Mundada, M., J, S. B., Shivashankar, S., Ramanjinappa G. D., & Gouda, S. (2025). Enhanced Multimodal Webpage Classification Using Deep Learning for Efficient Information Retrieval. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE52025256