j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 7, issue 3. 2024

Explainable Multimodal Deep Learning Model for Cyberbullying Detection (EMDL-CBD)

Mena Mohammed Abood *[1], Maha A. Al-Bayati [2]

[1&2] *Department of Computer Science, College of Science, Al-Mustansiriyah University, Baghdad, Iraq.*

mena.m.abood@uomustansiriyah.edu.iq

Abstract The increased use of social media and the internet is leading to an increase in cyberbully vulnerabilities as well as daily usage. Cyberbullying is a deliberate, aggressive behavior that can be committed by an individual or organization. It occurs when people communicate, post, and distribute damaging, false, or unfavorable content online. For individuals impacted, it results in emotional and mental health issues. Therefore, it is imperative to create automated techniques for the detection and prevention of cyberbullying. The majority of the research done on cyberbullying detection in recent years has been on text-based analysis. The two most significant media in incidents of cyberbullying are text and visual. This paper presents An Explainable Multimodal Deep Learning Model for Cyberbullying Detection include three steps The first step involves collecting datasets from different resources, which include images and their captions with binary classes (bullying and non-bullying). The second step applies two techniques of XAI: CNN+GradCam to analyze input images and produce visual explanations, and LSTM+LRP to analyze and interpret input text. The third step employs two techniques of data fusion (early and late). Final step represents the evaluation performance of the EMDL-CBD model based on a set of accuracy metrics

Crossref 10.36371/port.2024.3.6

Keywords: *Cyberbullying detection, multimodal data, Grad-Cam, LRP, LSTM.*

1. INTRODUCTION

Social media has grown in popularity as a handy means of communication for individuals of all ages as a result of the widespread use of the Internet. But social networking has brought about a number of issues [1]. These platforms have made it possible for individuals to interact and communicate in previously unimaginable ways, but they have also given rise to evil practices like cyberbullying. One form of psychological abuse that has a big effect on society is cyberbullying. It can be recognized by a pattern of offensive statements with harsh or derogatory language that are frequently posted.[2]

Events involving cyberbullying have been on the rise, especially among young people who frequently spend a lot of time switching between various social media sites. Because big social media platforms like Twitter are so widely used, abusers can remain anonymous on them, making them vulnerable to cyberbullying [2]. Not all tweets that use derogatory language are abusive, though. Numerous studies have been conducted on the automatic detection and prevention of cyberbullying, but more work needs to be done before a practical answer is reached [3].

Cyberbullying detection is important because it helps to identify and categories cyberbullying activities, deals with occurrences once they are recognized, and empowers users of the internet to take preventative measures against cyberbullying. It can be challenging to identify cyberbullying on social media platforms because different people interpret it differently, particularly when it comes to how severe it should be classified. For example, what one person considers to be intense bullying may not be for another.[3]

A detection model should be able to act instantly in order to stop cyberbullying instances, however this might be challenging to accomplish in real-world situations. Therefore, the propagation and influence of cyberbullying can be stopped more successfully if a cyberbullying detection model is able to classify cyberbullying episodes among distinct severity levels, allowing incidents to be prioritised. The primary objective of cyberbullying detection tasks is determining if text contains cyberbullying material.[4]

The main contributions of our research are as follows:

- Development an explainable multimodal deep learning model for cyberbullying detection. This model implementation on a new dataset are collect from different platform of social media with images and texts divided into classes for bullying and non-bullying.

- Adopts use two XAI methods (LSTM+LRP for text and CNN+GradCam for image), early and late data fusion, and accuracy measures to assess performance.

Mena Mohammed Abood, F Maha A. Al-Bayati. 2024, Explainable Multimodal Deep Learning Model for Cyberbullying Detection (EMDL-CBD). *Journal port Science Research*, 7(3), pp.268-280. https://doi.org/10.36371/port.2024.3.6

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 7, issue 3. 2024

2. LITERATURE REVIEW

A number of studies pertaining to the identification of cyberbullying have been proposed recently. Our study concentrates on both text- and image-based cyberbullying identification because no single medium can accurately identify the victim's motivation on social media.

The writers of Jadhav et al. (2023) [5] investigated text mining as a means of identifying cyberbullying on social media platforms. Convolutional neural networks, long short-term memory, bidirectional long short-term memory, CNN, and LSTM are employed in their techniques; LSTM obtained 66% accuracy.

Vishwamitra et al.'s study from 2021 [6] was centred on the notion that cyberbullying is more than just text or remark abuse. It may also appear in pictures. They gathered 19,300 legitimate photos in their attempt to gather actual data from social media platforms like Instagram. But much like in the earlier research, the focus of the researcher's examination was the narratives associated with the photos, therefore the text content was more important. Using their dataset of cyberbullying photos, the most popular classifier model, which was based on multimodal categorisation, produced a detection accuracy of 93.36%.

In order to detect and classify cyberbullying on Twitter based on many classes, researchers in Talpur and O'Sullivan (2020) [7] developed a supervised machine learning method. The dataset was created using text from the tweets. For instance, they employed Naïve Bayes, KNN, Decision Tree, Random Forest, and Support Vector Machine algorithms to leverage Embeddings for Sentiment, Lexicon characteristics for Lexicon, and PMI-semantic orientation. In instances of bullying rates, the accuracy ranged from 89% to 91%; the algorithms with the highest accuracy were SVM and Random.

In order to investigate cyberbullying on Twitter, researchers in Muneer and Fati (2020) [8] assembled a global dataset of 37,373 unique tweets from Twitter. Seven machine learning classifiers (092.8%) were utilised in the logistic regression process to obtain the highest accuracy (90.57%) and F1 score. The best precision (096.8%) and recall (100%) were achieved using the Stochastic Gradient Descent and Support Vector Machine.

In order to reduce occurrences of trolling, the system in Hitkul et al. (2019) [9] was designed to recognise photographs that are prone to trolling and alert users before content is posted online. It was found that traditional (i.e., state-of-the-art) photo categorisation methods were not useful in this context. The test accuracy for VGG16 was 61.81, and the validation accuracy for Inception V3 was 65.62, so the results were not very good.

Rosa et al. (2018) [10] have provided an outline of the varieties of cyberbullying and how to detect them. The review covered characteristics and classification strategies, as well as the detection of cyberbullying and the data sources that are available. The article contained the subjects on cyberbullying detection that this review paper addressed. The methods covered in the paper heavily rely on machine learning classifiers and natural language processing (NLP).

In order to identify cyberbullying in photo-sharing networks, Zhong et al. (2016) [11] conducted research. This essay focusses on early-warning techniques for identifying photos that could be attacked (on Instagram). Using a dataset of more than 3000 photos, they looked into the uploaded photos and captions to enhance bullying identification in reaction to shared content. To categorise and identify bullying in the photos, a variety of machine learning and deep learning techniques are used, and emergent remarks are made. They used classifiers such as Word2Vec, OFF, BoW, and captions to achieve 95.00% accuracy on a Natural Language Processing issue. With DL-FS (Stacked) and captions, they obtained an overall accuracy of 68.55 percent.

Table 1. *Literature review summary.*

Paper	Year	Type of Cyberbullying	Approach	Accuracy
[5]	2023	Text	LSTM, CNN, LSTM	66 %
[6]	2021	Images based on text	Multimodal classification	93.36 %
[7]	2020	Text	Machine Learning classifiers – supervised learning – SVM & Random Forest	89 % to 91 %
[8]	2020	Text	Machine learning – Logistic Regression	90.57 %
[9]	2019	Images	Transfer Learning and Machine Learning	Inception V3: 65.62 val_acc, VGG16: 61.81 test acc
[10]	2018	Text and images	Deep learning and machine learning (Support Vector Machines and Logistic Regression)	-
[11]	2016	Images with text	Natural Language Processing, BoW, and Word2Vec classifying. By using Captions and DL-FS.	Overall, 68.55 %, using captions 95.00 %.

Mena Mohammed Abood, F Maha A. Al-Bayati. 2024, Explainable Multimodal Deep Learning Model for Cyberbullying Detection (EMDL-CBD). *Journal port Science Research*, 7(3), pp.268-280. https://doi.org/10.36371/port.2024.3.6

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 7, issue 3. 2024

3. METHODOLOGY

The section before this one provides a detailed presentation and discussion of the suggested architecture. The suggested method for identifying cyberbullying in text and image data is essentially new. The field of explainable artificial intelligence (XAI) has come a long way, but its application to bullying detection is still in its early stages. The researcher in this study offers a sophisticated system that is specifically made for the detection of bullying and is based on XAI principles. "EMDL-CBD" stands for Explainable Multimodal Deep Learning Model for CyberBullying Detection.

The first step in the design of the proposed EMDL-CBD is the gathering of 1000 photographs and descriptions of cyberbullying from various social media sources, as shown in Figure (1). After that, the model combines Grad-CAM and CNN to analyses bullying images in-depth and identify patterns and visual indicators. Concurrently, bullying material is analyzed using LSTM and LRP (Layer-wise Relevance Propagation), which captures context and linguistic subtleties. These two lines of inquiry are used to provide a thorough knowledge of incidents of cyberbullying. Lastly, a fully linked layer for prediction is attached to the model.

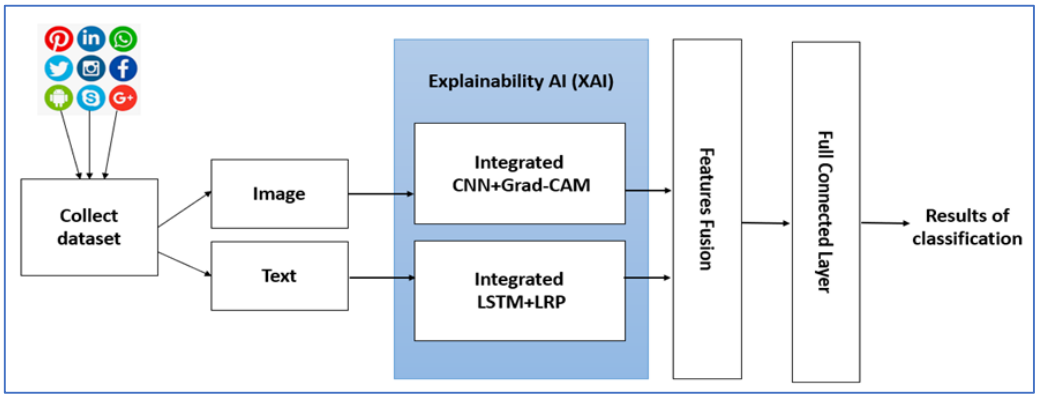

Figure 1 *Architecture of Proposed method*

The stages of the proposed EMDL-CBD system are discussed and explained as follows:

3.1 Dataset Collected They gather information from several social networking sites, including Twitter, Instagram, and Facebook. Ultimately, 849 examples were gathered so that the model could be trained. Each entry has two fields: an image and a written comment. The dataset are shown as examples in Figure.(2)

Sad teen with a phone in her bedroom-bulling image

Hacker doing his crime on a desktop computer in broad daylight- bulling image

Spring blossoms with butterfly- non bulling

Back view of young woman standing with outstretched hands against cloudy sky- non bulling

Figure 2. *An Examples of Images and text comment.*

Mena Mohammed Abood, F Maha A. Al-Bayati. 2024, Explainable Multimodal Deep Learning Model for Cyberbullying Detection (EMDL-CBD). *Journal port Science Research*, 7(3), pp.268-280. https://doi.org/10.36371/port.2024.3.6

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 7, issue 3. 2024

3.2 Explain ability AI (XAI)

This step shows how production (Guided backpropagation, Grad-Cam, and Guided-GradCam images) for image cyberbulling based on integrate CNN+GradCam and create relevance score for cyberbulling text based on integrate LSTM+LRP model.

3.2.1 XAI techniques to explain images:

3.2.1.1 Grad-CAM use case image classification

Using Gradient-based Class Activation Mapping (Grad-CAM), the model is analysed. In most cases, a machine learning model's interpretations are invisible to the human eye. This is known as the "black-box problem," in which the model's credibility is limited by its uninterpretability and opaqueness .[12]

Interpretability advances recently have proposed visualising the model's thinking behind the output by employing the pre-trained network's weights. While this method has increased the trustworthiness and dependability of neural networks, we think that illustrating the model's logic in challenging situations may have longer-term advantages, such as enlightening us about concepts that people are now unaware of. Here, we create coarse heatmaps that highlight the most discriminant regions of the provided images by using gradients of the pre-trained model to determine the weights of the feature maps on the class score [13].

The Grad-CAM is operated in three simple steps: First, the gradients of $\boldsymbol{y^c}$ (the score for a predicted class c) are calculated with respect to the feature map $\boldsymbol{A^k}$, which is $\frac{\partial y^c}{\partial A^k}$. Note that k is the number of feature maps produced from the respective convolutional layer, with width dimensions i and height dimensions j. Then, the gradients $\boldsymbol{y^c}$ are global-average-pooled with the height and width dimensions of the feature maps to obtain the neuron importance weights $\boldsymbol{\alpha_k^c}$,[13,14]

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A^k{}_{ij}} \qquad (1)$$

Where Z is the number of pixels in the concerned feature mapUltimately, a Rectifier Linear Unit (ReLU) is created by computing a weighted sum of the feature maps, which yields a course heatmap that highlights the most significant places [14].

$$L^c_{Grad-CAM} = ReLU\left(\sum_k \alpha_k^c\, A^k\right) \qquad (2)$$

The obtained coarse highlights of the discriminant regions are overlaid on the original images to visualize the most important regions.

3.2.1.2 Implementation integrate CNN +Grad-CAM foe providing Visual Explanantion

The integrate CNN-Grad-CAM works to produced three images. Heatmaps, also known as GradCam images, were created by Grad-CAM, which emphasized the key areas in the input picture that were crucial for class prediction. The most significant regions of a bulling image are visualized at the pixel level using high resolution heat maps produced by guided Grad-CAM, which combines Grad-CAM with guided backpropagation. The guided GradCam image can be obtained by multiply guided backpropagation and GradCam images.

The weight α in the integrated CNN-Grad-CAM network designs indicates the importance on feature map k towards the positive class. A weighted combination & a ReLU served to produce Grad-CAM. In the end, element-wise multiplication was used to combine Grad-CAM with Guided Backpropagation to create high-resolution Guided Grad-CAM maps.

3.2.2 LRP Analysis for LSTM Model:

3.2.2.1 Background on Long short-term memory networks (LSTM)

As a more sophisticated RNN network, LSTM has been used [15]. By using memory cells, commonly referred to as hidden layer units, it addresses the RNN's drawback. Three gates—the input, output, and forget gates—are used to control memory cells. They possess self-connections that allow them to store the network's temporal state [16]. Information flowing from memory cell input and output to the rest of the network is addressed and controlled by input and output gates. The information with larger weights is transferred from the preceding neurone to the subsequent neurone by the forget gate, also known as a remember vector. The memory cell gains the forget gate. Depending on the high activation results, the information is saved in memory; if the input unit has a high activation, the information will be stored in a memory cell. In the event that the output unit exhibits strong activation, the information will be transmitted to the subsequent neurone. In the absence of such, high-weight input data is stored in the memory cell[17] .Mathematically, LSTM network can be described as [17]:

$$\boldsymbol{h_t = f(W_h . x_t + U_h . h_{t-1} + b_h)} \qquad (3)$$

Where $W_h \in R^{m\,x\,d}$ and $U_h \in R^{m\,x\,m}$ indicates weight matrices, x_t denotes the current word embedding, $b_h \in R^m$ refers to bias term, whereas f(x) is a non-linear function.

LSTMs tend to retain information for a longer period of time and have a more intricate architecture with hidden states. This is how the hidden state of an LSTM is calculated:[18]

Mena Mohammed Abood, F Maha A. Al-Bayati. 2024, Explainable Multimodal Deep Learning Model for Cyberbullying Detection (EMDL-CBD). *Journal port Science Research*, 7(3), pp.268-280. https://doi.org/10.36371/port.2024.3.6

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 7, issue 3. 2024

$$f_t = \sigma(w_f[h_{t-1}, X_t] + b_f \quad (4)$$

$$i_t = \sigma(\omega_j.[h_{t-1}, X_t] + b_i) \quad (5)$$

$$O_t = \sigma(w_o.[h_{t-1}, X_t] + b_o) \quad (6)$$

$$C_t = f_t.C_{t-1} + i_t.\check{C}_t \quad (7)$$

$$h_t = O_t.\tanh(C_t) \quad (8)$$

Where f_t denotes the forget gate, i_t refers to the input gate, C_t denotes the cell state, O_t is the output gate, h_t is the regular hidden state, σ indicates sigmoid function, and ◦ is the Hadamard product.

3.2.2.2 General LRP (Layer-wise Relevance Propagation) theory

A recently developed method for acquiring these explanations is called LRP. It can be used with several classifiers for machine learning, including deep convolutional neural networks. The function value f(x) is decomposed by the LRP technique on its input variables in a way that satisfies the conservation property:[19]

$$f(x) = \sum_d R_d \quad (9)$$

By doing a backward pass on the network, the decomposition is achieved by redistributing each neuron's associated significance to its ancestors. Considering neurons mapping a set of n inputs $(x_i)_{i\in[1,n]}$ to the neuron activation $x_j$through the sequence of functions [20]:

$$Z_{ij} = x_i w_{ij} + \frac{b_j}{n}, \quad (10)$$

$$Z_j = \sum i^{Z_{ij}},$$

$$x_j = g(Z_j)$$

Where each input neurone has received the same amount of the neurone bias $b_j$for simplicity, and where g(·) is an activation function that increases monotonically. Denoting by $R_i$and R_j the relevance associated with X_i and X_j, the relevance is redistributed from one layer to the other by defining messages $R_{i\leftarrow j}$ ←j indicating how much relevance must be propagated from neuron X_j to its input neuron X_i in the lower layer. These messages are defined as[20]:

$$R_{i\leftarrow j} = \frac{Z_{ij} + \frac{s(Z_j)}{N}}{\sum_i Z_{ij} + S(Z_j)} R_j \quad (11)$$

here $S(Z_j)\,\epsilon.\left(1_{z_j\geq 0} - 1_{z_j<0}\right)$ is a stabilizing term that handles near-zero denominators, with ϵ set to 0.01. The intuition behind this local relevance redistribution formula is that each input xi should be assigned relevance proportionally to its contribution in the forward pass, in a way that the relevance is preserved ($\sum_i R_{i\leftarrow j} = R_j$).

Each neuron in the lower layer receives relevance from all upper-level neurons to which it contributes [21]

$$R_i = \sum_j R_{i\leftarrow j} \quad (12)$$

This pooling ensures layer-wise conservation [21]:

$$\sum_i R_i = \sum_j R_j \quad (13)$$

Finally, in a max-pooling layer, all relevance at the output of the layer is redistributed to the pooled neuron with maximum activation.

3.2.2.3 Apply LRP for LSTM

In multilayer perceptron architectures, RNN-LSTM and other gated neural networks include a unique calculation called multiplicative interaction in addition to linear mapping computation. Two neurones are multiplied by one another in this computation: one acts as a signal, and the other as a gate that regulates how much the signal affects the output:

$$\alpha_p = f(Z_g).g(Z_s), \quad (14)$$

Where zg and zs are two neurone values supplied to the gate and signal units from previous layers, respectively, and (·) is the activation function for the gate unit and (·) is the activation function for the signal unit.

In contrast to linear mapping, the multiplicative interaction's nonlinearity presents unique challenges in terms of allocating relevance to the preceding layer. One often used redistribution technique in this situation, where an activation is obtained by multiplying the value of a gate neurone by the value of a signal neurone, is known as "signal-take-all." This refers to:

$$(R_g, R_s) = (0, R_p) \quad (15)$$

Where the relevance scores given to the gate and signal neurones are denoted by Rg and Rs, respectively. The gate neurone follows the conservation principle by taking zero, whereas the signal neurone receives all the relevance Rp from the top layer. This tactic can be understood as follows: while the gate regulates information flow, it is not information in and of itself. Instead, data is fully integrated into the signal. Even though it appears that zg is completely disregarded, zg's influence has actually been taken into account when determining the value of RP from the upper-layered structure.

3.2.3 Multimodal Data Fusion

Mena Mohammed Abood, F Maha A. Al-Bayati. 2024, Explainable Multimodal Deep Learning Model for Cyberbullying Detection (EMDL-CBD). *Journal port Science Research*, 7(3), pp.268-280. https://doi.org/10.36371/port.2024.3.6

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 7, issue 3. 2024

The proposed MDL-CBD explore Early and late technique of multimodal data fusion. The framework of multimodal data fusion using Early and late technique. Two primary methods for approaching the fusion process have been used in the work: feature level, also known as early fusion, and decision level, also known as late fusion. In early fusion, many raw data sources are combined to feed into models (integrate CNN+GradCom with image & integrate LST+LRP with text), which ultimately generates an inference.

4. Results of Explainability AI (XAI) Stage

The researcher applied the proposed EMDL-CBD to a newly collected dataset from various sources. This dataset comprises images and their descriptions, categorized into binary classes (cyberbullying and non-cyberbullying). The total number of samples is approximately 849, divided into 250 from the cyberbullying class and 599 from the non-cyberbullying class. The researcher divided the datasets into 60% training, 20% testing, and 20% validation (see details in Table 2).

Table 2. *Details of Splitting the Multimodal Dataset.*

Process	bullying	Non-bullying	total
60%training	150	361	**511**
20%validation	50	119	**169**
20%testing	50	119	**169**
Total	**250**	**599**	**849**

4.1 Results of Explainability AI (XAI) Stage

This step includes two types of XAI techniques for providing visual explanations: integrating CNN+GradCam with images and integrating LSTM+LRP with text data. This section presents the results of both techniques. Figure 3 illustrates the results of integrating CNN+GradCam. The integration of CNN and GradCam results in three images (guided backpropagation image, GradCam image, guided-GradCam image).

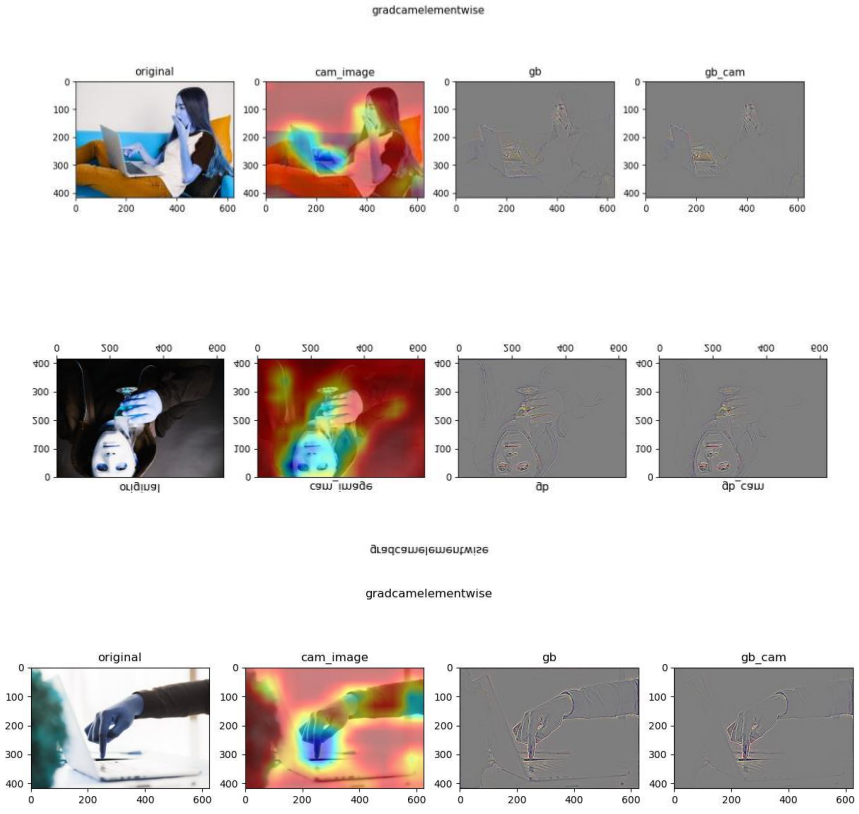

Figure 3. *Results of Integrating CNN+GradCam with Image.*

Figure 4 illustrates the results of integrating LSTM+LRP. The integration of LSTM and LRP produces color gradients to identify the most important words.

Mena Mohammed Abood, F Maha A. Al-Bayati. 2024, Explainable Multimodal Deep Learning Model for Cyberbullying Detection (EMDL-CBD). *Journal port Science Research*, 7(3), pp.268-280. https://doi.org/10.36371/port.2024.3.6

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 7, issue 3. 2024

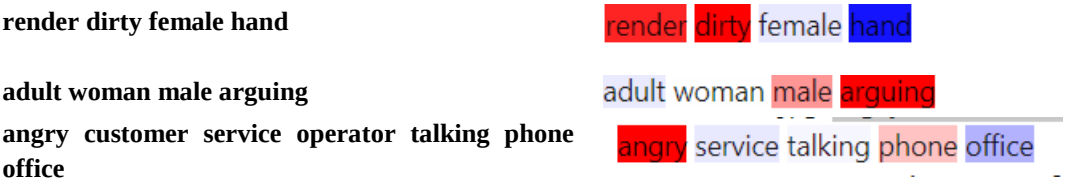

Figure 4. *Results of Integrating LSTM+LRP with Text.*

4.2 Results of the Multimodal Data Fusion (early and late)

Figures 5 and 6 display the results of early fusion for training and validation processes. Figure 5 illustrates the accuracy scores in the training and validation process, where the vector line refers to accuracy scores and the horizontal line refers to the number of epochs (50). The blue curve refers to the training process, while the green curve refers to the validation process. The train accuracy score is approximate (0.99987) and validated (0.99874).

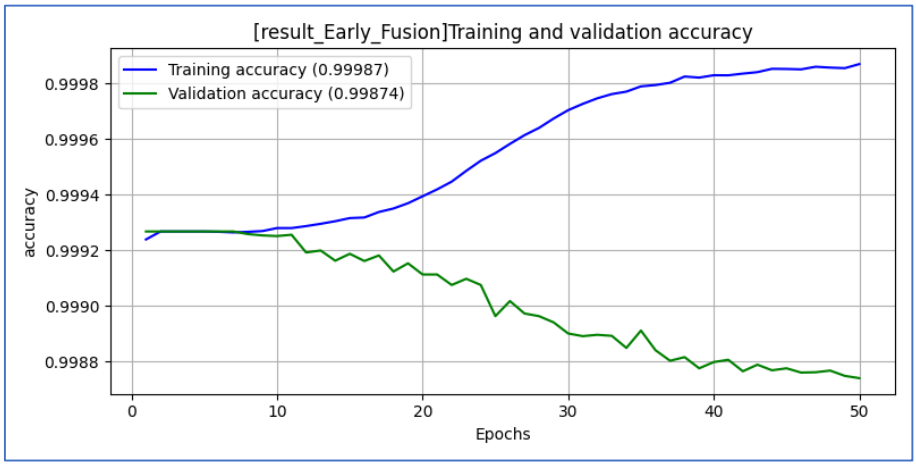

Figure 5. *Results of Accuracy of Traning and Validation Process of the Early Data Fusion.*

Figure 6 illustrates the Loss scores in the training and validation process, where the vector line refers to Loss scores and the horizontal line refers to the number of epochs (50). The train loss score is approximate (0.00040) and validated (0.0142).

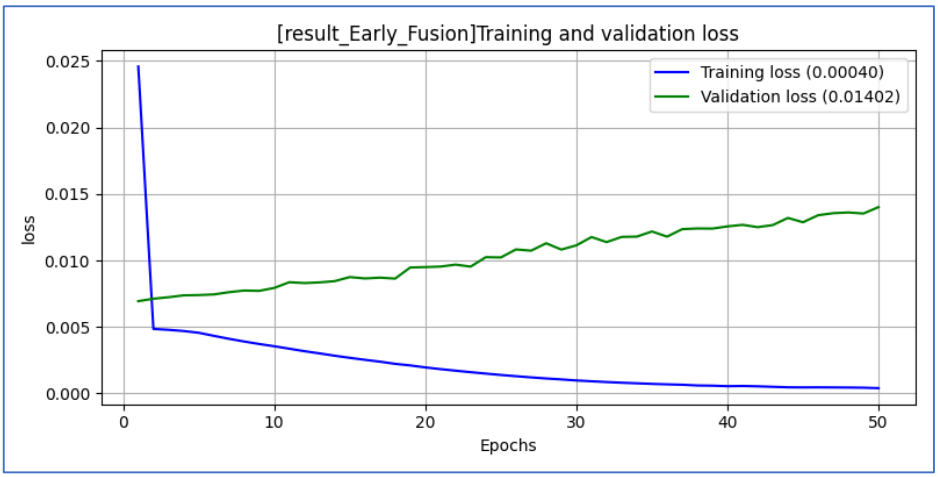

Figure 6. *Results of Loss of Traning and Validation Process of the Early Data Fusion.*

Figure 7 shows the confusion matrix of the early data fusion for the testing process. The total samples of testing are about 169, distributed as TP (114), TN(50), FP(5), and FN(0). The accuracy test was about 0.970414.

Mena Mohammed Abood, F Maha A. Al-Bayati. 2024, Explainable Multimodal Deep Learning Model for Cyberbullying Detection (EMDL-CBD). *Journal port Science Research*, 7(3), pp.268-280. https://doi.org/10.36371/port.2024.3.6

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 7, issue 3. 2024

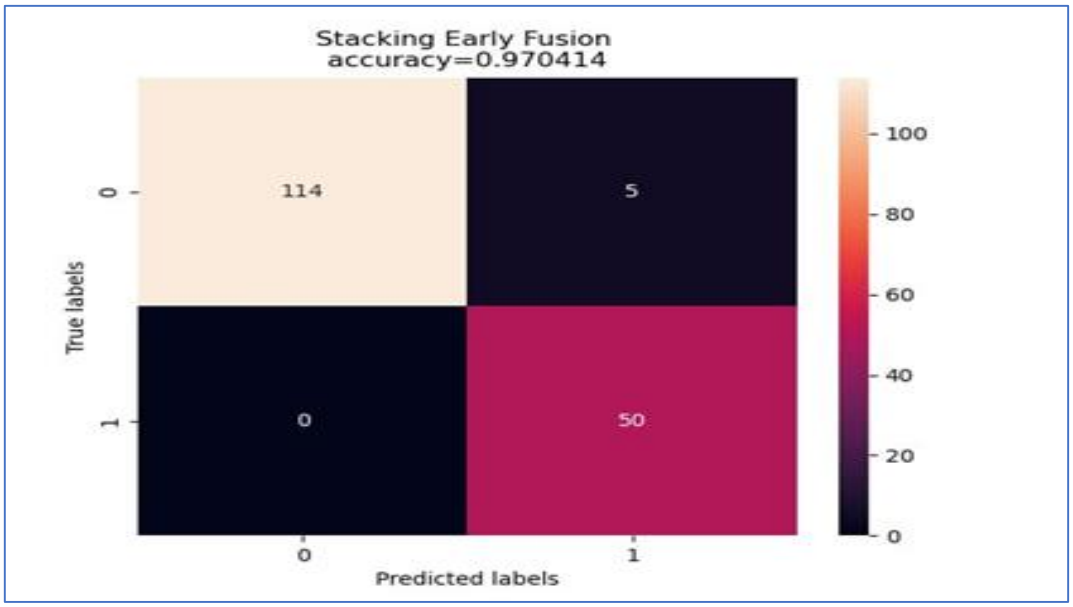

Figure 7. *Results of Confusion Matrix of the Early Data Fusion in Testing Process.*

Figures 8 and 9display the results of late fusion for training and validation processes. Figure 8 illustrates the accuracy scores in the training and validation process, where the vector line refers to accuracy scores and the horizontal line refers to the number of epochs (50). The train accuracy score is approximate (0.99963) and validated (0.99896)

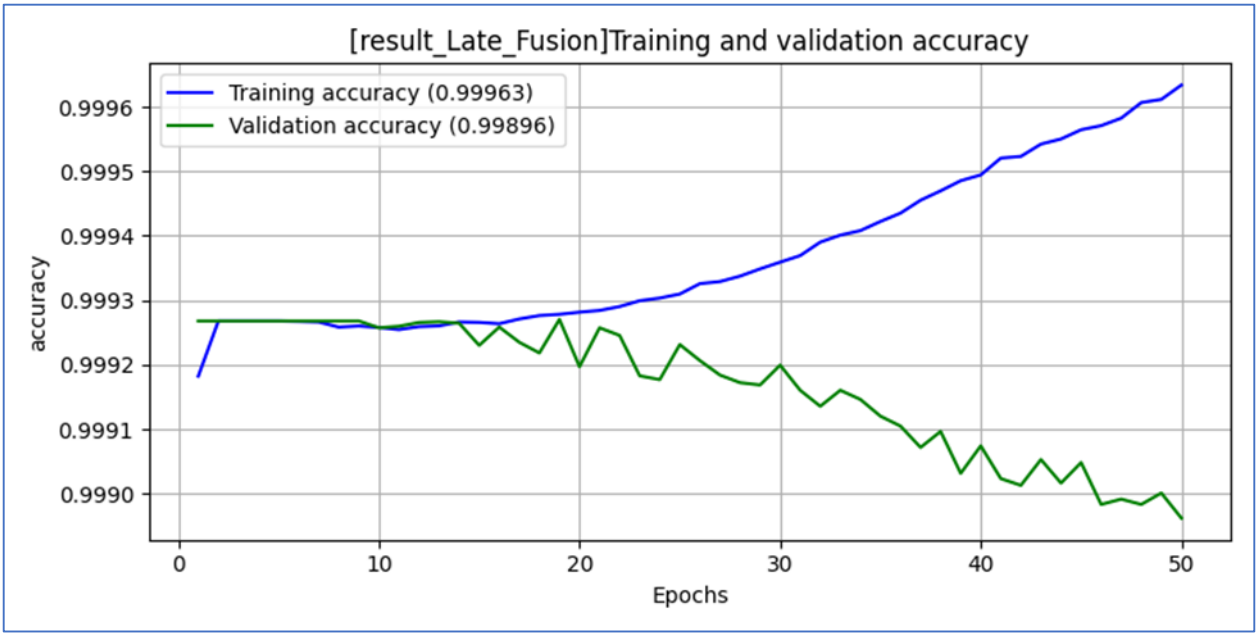

Figure 8. *Results of Accuracy of Traning and Validation Process of the Late Data Fusion.*

Figure 9 illustrates the Loss scores in the training and validation process, where the vector line refers to Loss scores and the horizontal line refers to the number of epochs (50). The train loss score is approximate (0.00114) and validated (0.01200).

Mena Mohammed Abood, F Maha A. Al-Bayati. 2024, Explainable Multimodal Deep Learning Model for Cyberbullying Detection (EMDL-CBD). *Journal port Science Research*, 7(3), pp.268-280. https://doi.org/10.36371/port.2024.3.6

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 7, issue 3. 2024

Figure 9. *Results of Loss of Traning and Validation Process of the late Data Fusion.*

Figure 10 shows the confusion matrix of the late data fusion for the testing process. The total samples of testing are about 169, distributed as TP (117), TN(49), FP(2), and FN(1). The accuracy test was about 0.982249.

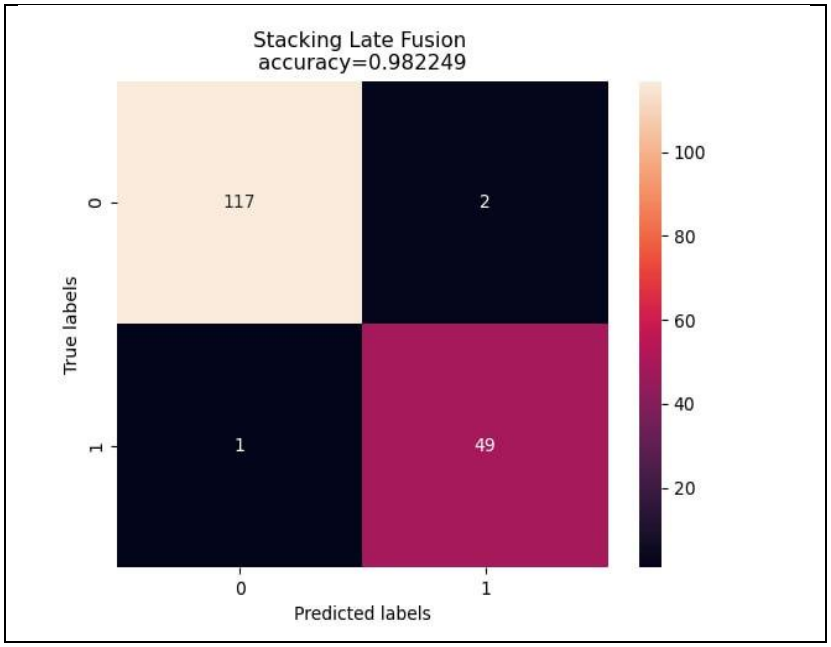

Figure 10. *Results of Confusion Matrix of the Late Data Fusion in Testing Process.*

4.3 Comparison between Multimodal Data Fusion Techniques

This section will compare the various multimodal data fusion techniques used in this work, taking into account evaluation metrics and testing results. Figure 11 presents a comparison of early and late fusion approaches, focusing on accuracy, recall, precision, f1 score, and specificity testing results.

The comparison demonstrated that late data fusion yielded better results in comparison with early fusion. Because late data fusion analyses each kind of data individually before integrating their findings. This makes it feasible to handle each data type as efficiently as possible, lowers noise mixing, and makes good use of each data type's advantages. So, there is an improvement in overall performance with respect to accuracy, recall, precision, F1 score, and specificity since the combined final predictions are more accurate and dependable.

Mena Mohammed Abood, F Maha A. Al-Bayati. 2024, Explainable Multimodal Deep Learning Model for Cyberbullying Detection (EMDL-CBD). *Journal port Science Research*, 7(3), pp.268-280. https://doi.org/10.36371/port.2024.3.6

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 7, issue 3. 2024

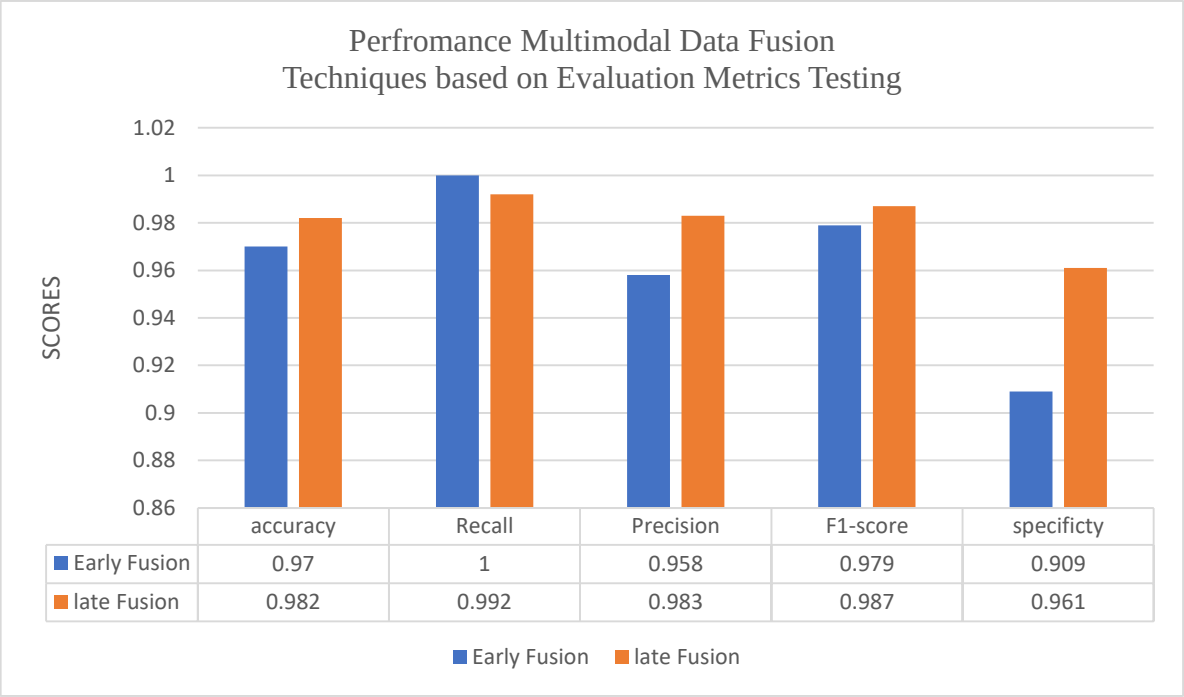

Figure 11. *Comparsion between Early and Late Data Fusion based on Evaluation Metrics Testing.*

4.4 Performance Evaluation Metrics-

The following equations define the evaluation metrics, where TP (true positive), TN (true negative), FP (false positive), and FN (false negative) are used [22]:

-Accuracy: this indicator assesses how closely the actual data values match the expected value. The definition of it can be found in the formula below:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \qquad (16)$$

tp: True Positive, tn: True Negative, f p: False Positive, f n: False Negative

-Classification error: This is the quantity of samples that have been misclassified (false negatives and false positives combined). It has the following definition:

Classification Error (Err) = 1 – Acc (17)

-Precision: The precision metric evaluates the classifier's capacity to omit unnecessary samples. This metric's formula can be defined as follows:

$$\text{Precision (Pre)} = \frac{\text{tp}}{tp + fp} \qquad (18)$$

-Sensitivity: The percentage of appropriately identified pertinent samples is measured by the sensitivity metric. The following is a representation of it:

$$\text{Sensitivity (Sn)} = \frac{tp}{tp + fn} \qquad (19)$$

-F1-Score: The weighted average of sensitivity (recall) and precision yields the F1-score, with recall and precision contributing the same proportion to the score. The definition of the F1-score is as follows:

$$\text{F1 Score} = \frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \qquad (20)$$

-Specificity: This refers to the classifier's capacity to identify the real negative rate. The following equation can be used to define the formula of specificity:

$$\text{Speci f icity (Sp)} = \frac{tn}{tn + fp} \qquad (21)$$

-False-Positive Rate (FPR): The percentage of negative cases that are mistakenly classified as positive. This measure, which is also referred to as the miss rate, is shown as follows:

$$\text{False} - \text{Positive Rate (FPR)} = \frac{fp}{fp + tn} \qquad (22)$$

-False-Negative rate (FNR): The percentage of negative examples that are mistakenly classified as positive. The fall-out rate is another name for this statistic. The following is an introduction to this evaluation criterion:

$$\text{False} - \text{Negative Rate (FNR)} = \frac{fn}{fn + tp} \qquad (23)$$

5. Future Performance of Explainable Multimodal Deep Learning for Cyberbullying Detection Using Swin Transformer and Wavelet Transform

Mena Mohammed Abood, F Maha A. Al-Bayati. 2024, Explainable Multimodal Deep Learning Model for Cyberbullying Detection (EMDL-CBD). *Journal port Science Research*, 7(3), pp.268-280. https://doi.org/10.36371/port.2024.3.6

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 7, issue 3. 2024

The integration of Swin Transformers and wavelet transforms within an explainable multimodal deep learning model for cyberbullying detection holds significant promise for enhancing performance and interpretability. Here's a look and details of potential future developments [23-35].:

1) **Enhanced Feature Extraction and Fusion** This can be achieved by using more sophisticated techniques for fusing information from different modalities (text, image, audio) will be explored. This could involve attention mechanisms or early fusion strategies to capture complex interactions between modalities.

2) **Improved Explainability** Mixing attention mechanisms within Swin Transformers to highlight the most relevant parts of the input data can provide valuable insights into the model's decision-making process.This can help understand the model's sensitivity to different input features.

3) **Real-Time Detection and Adaptation** By developing more efficient Swin Transformer-based models will enable real-time detection of cyberbullying incidents. Models will be capable of adapting to evolving cyberbullying tactics by incorporating new data and retraining models incrementally.

4) **Multimodal Data Fusion** Therefore, exploring ways to share information between different modalities within the Swin Transformer architecture can lead to improved feature representation and classification accuracy.

6. Conclusion and future work

Cyberbullying incidents are on the rise as a result of social media users using text and image-based communication more frequently. Prior to causing harm to users, these occurrences must be identified and stopped. In this paper, this paper presents An Explainable Multimodal Deep Learning Model for Cyberbullying Detection such as image and text. The data extracted from different sources the model developed with by employing two advanced XAI techniques, including CNN-Gardham for image analysis and LSTM-LRP for text analysis, the proposed EMDL-CBD model achieves optimal classification performance. These approaches, as evidenced by the results in Results of Early Data Fusion Technique the train accuracy score is approximate the train accuracy score is approximate (0.99963) and validated (0.99896). Effectively extract and utilize important features for the best results. In the future, cyberbullying detection can be coupled with audio and video. Pictures Text is taken into account while looking for signs of cyberbullying. It is possible to include options for multilingual, cross-linguistic, and mix language.

REFERENCES

[1] J. L. Wu and C. Y. Tang, "Classifying the severity of cyberbullying incidents by using a hierarchical squashing-attention network," *Applied Sciences*, vol. 12, no. 7, p. 3502, 2022.

[2] N. Haydar and B. N. Dhannoon, "A comparative study of cyberbullying detection in social media for the last five years," *Al-Nahrain Journal of Science*, vol. 26, no. 2, pp. 47-55, 2023.

[3] M. H. Obaid, S. K. Guirguis, and S. M. Elkaffas, "Cyberbullying detection and severity determination model," *IEEE Access*, 2023.

[4] N. A. Azeez, S. Misra, O. I. Lawal, and J. Oluranti, "Identification and detection of cyberbullying on Facebook using machine learning algorithms," *Journal of Cases on Information Technology (JCIT)*, vol. 23, no. 4, pp. 1-21, 2021.

[5] M. A. Khan, S. Kadry, P. Parwekar, R. Damaševičius, A. Mehmood, J. A. Khan, and S. R. Naqvi, "Human gait analysis for osteoarthritis prediction: a framework of deep learning and kernel extreme learning machine," *Complex Intell. Syst.*, 2021. Available: https://doi.org/10.1007/s40747-020-00244-2.

[6] R. Jadhav, G. Chaudhari, and S. Rane, "Cyber bullying detection," *Int. J. Creative Res. Thoughts (IJCRT)*, pp. 1–6, 2023.

[7] N. Vishwamitra, H. Hu, F. Luo, and L. Cheng, "Towards understanding and detecting cyberbullying in real-world images," in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2021.

[8] B. A. Talpur and D. O'Sullivan, "Cyberbullying severity detection: A machine learning approach," *PLoS ONE*, vol. 15, no. 10, Art. no. e0240924, 2020.

[9] A. Muneer and S. M. Fati, "A comparative analysis of machine learning techniques for cyberbullying detection on Twitter," *Future Internet*, vol. 12, no. 11, p. 187, 2020.

[10] H. Hitkul, R. R. Shah, P. Kumaraguru, and S. Satoh, "Maybe look closer? Detecting trolling prone images on Instagram," in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pp. 448–456, IEEE, 2019.

[11] H. Rosa, D. Matos, R. Ribeiro, L. Coheur, and J. P. Carvalho, "A 'deeper' look at detecting cyberbullying in social networks," in *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2018.

[12] K. Hara, T. Kojima, K. Kutsukake, H. Kudo, and N. Usami, "3D CNN and grad-CAM based visualization for predicting generation of dislocation clusters in multicrystalline silicon," APL Machine Learning, vol. 1, no. 3, 2023.

[13] F. Vaquerizo-Villar, G. C. Gutiérrez-Tobal, E. Calvo, D. Álvarez, L. Kheirandish-Gozal, F. Del Campo, et al., "An explainable deep-learning model to stage sleep states in children and propose novel EEG-related patterns in sleep apnea," Computers in Biology and Medicine, vol. 165, p. 107419, 2023.

[14] L. D. Quach, K. N. Quoc, A. N. Quynh, N. Thai-Nghe, and T. G. Nguyen, "Explainable deep learning models with gradient-weighted class activation mapping for smart agriculture," IEEE Access, vol. 11, pp. 83752-83762, August 2023.

[15] Kareem, M. R., "Image analysis and detection of olive leaf diseases using recurrent neural networks," Al-Mustansiriyah Journal of Science, vol. 35, no. 1, pp. 60-65, 2024.

[16] M. H. Al-Tai, B. M. Nema, and A. Al-Sherbaz, "Deep learning for fake news detection: Literature review," Al-Mustansiriyah Journal of Science, vol. 34, no. 2, pp. 70-81, 2023.

[17] M. T. Hasan, M. A. E. Hossain, M. S. H. Mukta, A. Akter, M. Ahmed, and S. Islam, "A review on deep-learning-based cyberbullying detection," Future Internet, vol. 15, no. 5, p. 179, 2023.

[18] A. Dewani, M. A. Memon, and S. Bhatti, "Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data," J. Big Data, vol. 8, no. 1, p. 160, 2021.

[19] A. Bennetot et al., "A practical tutorial on explainable ai techniques," arXiv, 2021, arXiv:2111.14260.

[20] T. R. Dieter and H. Zisgen, "Evaluation of the Explanatory Power Of Layer-wise Relevance Propagation using Adversarial Examples," Neural Processing Letters, vol. 55, no. 7, pp. 8531-8550, 2023.

[21] R. Singh, "Understanding Image Classification Tasks Through Layerwise Relevance Propogation," in 2022 IEEE 18th International Conference on Intelligent Computer Communication and Processing (ICCP), September 2022, pp. 199-203.

[22] N. M. Khassaf and S. H. Shaker, "Image Retrieval based Convolutional Neural Network," Al-Mustansiriyah Journal of Science, vol. 31, no. 4, pp. 43-54, 2020

[23] Afrah U Mosaa, Waleed A Mahmoud Al-Jawher "A proposed Hyper-Heuristic optimizer Nesting Grey Wolf Optimizer and COOT Algorithm for Multilevel Task" Journal Port Science Research, Vol. 6, PP. 310,317, 2023.

[24] [24] Walid A Mahmoud, Majed E Alneby, Wael H Zayer "2D-multiwavelet transform 2D-two activation function wavelet network-based face recognition" J. Appl. Sci. Res, Vol. 6, Issue 6, PP. 1019-1028, 2010.

[25] Maryam I Al-Khuzaie, Waleed A Mahmoud Al-Jawher "Enhancing Medical Image Classification: A Deep Learning Perspective with Multi Wavelet Transform" Journal Port Science Research, Vol. 6, Issue 4, PP. 365-373, 2023.

[26] A. H. Salman, W. A. Mahmoud Al-Jawher "A Hybrid Multiwavelet Transform with Grey Wolf Optimization Used for an Efficient Classification of Documents" International Journal of Innovative Computing 13 (1-2), 55-60, 2022.

[27] Saadi M Saadi and Waleed A Mahmoud Al-Jawher "Image Fake News Prediction Based on Random Forest and Gradient-boosting Methods" Journal Port Science Research, Vol. 6, Issue 4, PP. 357-364, 2023.

[28] Ahmed Hussein Salman, Waleed Ameen Mahmoud Al-Jawher "A Hybrid Multiwavelet Transform with Grey Wolf Optimization Used for an Efficient Classification of Documents" International Journal of Innovative Computing, Vol. 13, Issue 1-2, PP. 55-60, 2022.

[29] Sarah H Awad Waleed A Mahmoud Al-Jawher "Precise Classification of Brain Magnetic Resonance Imaging (MRIs) using Gray Wolf Optimization (GWO)" HSOA Journal of Brain & Neuroscience Research, Volume 6, Issue 1, Pages 100021, 2022.

[30] W. A. Mahmoud, Jane Jaleel Stephan and A. A. W. Razzak "Facial Expression Recognition from Video Sequence Using Self Organizing Feature Map" Journal port Science Research و TRANSACTION ON ENGINEERING, TECHNOLOGY AND THEIR APPLICATIONS, Vol. 4, Issue 2, 2021.

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 7, issue 3. 2024

[31] Dihin, R. Al-Jawher, Waleed and Al-Shemmary "Implementation of The Swin Transformer and Its Application In Image Classification" Journal Port Science Research, vol. 6, Issue 4, PP. 318-331. 2023

[32] W. A. Mahmoud & Z. Ragib "Face Recognition Using PCA and Optical Flow" Engineering Journal, Vol. 13, Issue 1, PP. 35-47, 2007.

[33] Waleed A Mahmoud Al-Jawher, Sarah H Awad "A proposed brain tumor detection algorithm using Multi wavelet Transform (MWT)" Materials Today: Proceedings, Volume 65, Pages 2731-2737, 2022.

[34] Rasha Ali Dihin, Waleed A Mahmoud Al-Jawher, Ebtesam N AlShemmary "Diabetic Retinopathy Image Classification Using Shift Window Transformer", International Journal of Innovative Computing, Vol. 13, Issue 1-2, PP. 23-29, 2022.

[35] AHM Al-Heladi, W. A. Mahmoud, HA Hali, AF Fadhel "Multispectral Image Fusion using Walidlet Transform" Advances in Modelling and Analysis B, Volume 52, Iss. 1-2, pp. 1-20, 2009.

[36] Maryam I Mousa Al-Khuzaay, Waleed A Mahmoud Al-Jawher, "New Proposed Mixed Transforms: CAW and FAW and Their Application in Medical Image Classification" International Journal of Innovative Computing, Volume 13, Issue 1-2, Pages 15-21, 2022.