International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An Efficient Model for Visual Sentiment Analysis using Hybrid Feature Extraction and Fusion-Based Classification

Siddhi Kadu*[1], Bharti Joshi [2], Pratik K. Agrawal [3]

Submitted: 28/01/2024 **Revised**: 06/03/2024 **Accepted**: 14/03/2024

Abstract: The exponential growth of the internet technology industry has led users to share their opinions/sentiments through an online platform not only in the form of text but also in the form of images, speech, and videos in variety of applications. Researchers are focusing on building a sentiment analysis model based on image data, as it offers a more effective method for analyzing sentiments. The need is addressed with a model for visual sentiment analysis which uses hybrid feature extraction and fusion-based classification method The proposed approach strategically combines the strengths of multiple visual features, selected through the effective Dual Moth Flame Optimization (DMFO) model. To effectively leverage the selected features, a customized Fusion-based Convolutional Neural Network (CNN) architecture is specifically designed for visual analysis for multiclass sentiment classification, categorizing visual sentiments into positive, negative, and neutral. Our proposed model is superior to existing approaches, as shown by empirical evaluations on multiple datasets and achieves outperforming efficiency as compared to existing methods. In addition, the model's applicability to real-time scenarios is promising. The approach ensures robust performance and holds promise for applications in social media analysis, marketing, user experience assessment which increases the model's adaptability levels.

Keywords: *Visual Sentiment Analysis, Visual Features, Convolutional Neural Networks*

1. Introduction

Sentiment analysis systems use publicly accessible data to analyse user data to understand attitudes, ideas, and trends, aiming to gauge opinions and emotional impact on web users. The tasks in this study field are both hard and practical. It has a wide range of practical applications because sentiments [1], [2], [3] are influenced in both corporate and social environments. Many human decisions rely on user feedback to evaluate a product before purchasing it. Individuals and businesses are increasingly depending on public opinion to make decisions, thanks to the advent of social media (for example, feedback, forums, and Facebook and Twitter) [4]. Furthermore, people make use of videos and images along with text messages on the different social platforms to express themselves [5], [6], [7]. The images contain not only semantic content like objects or activities but also affect the sentiment represented by the exhibited scene. This information is crucial for comprehending the emotional impact beyond the semantic level. Existing approaches to visual sentiment analysis frequently ignore the rich visual information embedded in images in favour of textual data [8], [9], [10]. While textual analysis techniques provide valuable insights, they are incapable of capturing the intricate details and subtleties conveyed by visual characteristics. This limitation hinders the efficacy and precision of visual analysis [11], [12], [13].

To address this deficiency, a novel multidomain visual sentiment model with fusion-based deep learning is proposed for the analysis. Our model aims to overcome the limitations of existing approaches by incorporating a wide variety of visual characteristics. The Fourier Transform(FT), Wavelet Transform(WT), Discrete cosine transform(DCT), and Convolutional Transform(CT) are utilized to analyze diverse visual characteristics in images. The Dual Moth Flame Optimization (DMFO) Model is a method that combines inter-class variance maximization and intra-class variance minimization to optimize features. This intelligent feature selection procedure ensures that our model can distinguish between distinct sentiment categories while remaining robust against within-class variations. In addition, to fully exploit the potential of the chosen visual characteristics, a Fusion-based CNN architecture that is specifically tailored for visual sentiment analysis is designed. This architecture is composed of multiple convolutional and pooling layers with finely tuned hyperparameters and dropout regularization. The traditional Fully Connected Layer is replaced by a combination of Naive Bayes(NB), K-Nearest Neighbour(KNN), Support Vector Machine(SVM), Deep Forest(DF), and Logical Regression(LR) classifiers. This integration permits efficient classification while enhancing the model's

[1] *Ramrao Adik Institute of Technology, D Y Patil deemed to be University Nerul, Navi Mumbai,India*
ORCID ID : 0000-0003-4735-6218
[2] *Ramrao Adik Institute of Technology, D Y Patil deemed to be University Nerul, Navi Mumbai,India*
ORCID ID :0000-0001-8082-3450
[3] *Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, Maharashtra, India*
ORCID ID : 0000-0003-1578-308X
** Corresponding Author Email: siddhi.k1121989@gmail.com*

interpretability on the CK+, FER2013, and JAFFE datasets [14]. The system's goal is to identify the polarity—that is, positive, neutral, or negative—of the emotion that an image represents [15]. Here are the contributions that stand out the most.

• The development of the model which combines the various visual characteristics of images using Fourier Transform, Wavelet Transform, DCT, and Convolutional Transform.

• The research introduces a novel method that combines the robust capabilities of Dual Moth Flame Optimization (DMFO) for feature selection.

• The development of a Fusion-based Convolutional Neural Network (CNN) architecture where the conventional Fully Connected Layer is replaced by a combination of Naive Bayes, KNN, SVM, Deep Forest, and Logical Regression classifiers is used for the classification purpose.

• Our proposed model is superior to existing approaches, as evidenced by empirical evaluations conducted on a comprehensive datasets & samples.

• Our model surpasses the most advanced techniques in terms of precision, accuracy, recall, and F1-score.

The organisation of the following sections of the paper is as follows: The several sentiment analysis techniques now in use will be briefly reviewed in the next section. In Section 3, a specialised Fusion-based CNN architecture designed for visual sentiment analysis is succinctly introduced. Section 4 presents the results of the experimental trials. component 5, the final component of the paper, contains the conclusion and recommendations for further research.

2. Literature Review

Existing models for image sentiment analysis have made substantial strides in comprehending different sentiments for social media data and enhancing quality of levels. However, they have limitations that diminish their effectiveness and precision levels that can be improved via use of different techniques. Using the popular deep convolutional neural network Inception-v3 and extra deep features, the authors' work [14] enhances image classification performance on the FER2013, JAFFE, and CK+ datasets. According to the study, the suggested approach can achieve a 99.5% accuracy rate. The authors [16] suggested a genetic algorithm and Gabor filters. For choosing the best features for the SVM, a genetic algorithm can be used to optimise the SVM hyperparameters. On the CK and CK+ datasets, the recognition rate was 94.20% and 94.26% respectively. An AlexNet-DCNN model was presented by the authors [17] in order to learn characteristics associated with various emotion classes. The CK+ and CK dataset has an average recognition accuracy of 93.66%. The authors [18] have developed an ensemble deep learning algorithm using CNN. The proposed technique joins three independent sub-networks to create a larger network for emotion recognition system called HERO. The influence of kernel size and number of filter these two parameters of CNN are considered for the classification of FER-2013 dataset and the accuracy is explored. In this study as part of the authors [19] attempt to create new CNN models. Numerous existing methods primarily analyse textual data, such as customer reviews and ratings, while ignoring the valuable visual information contained in product images. While textual analysis provides insight into opinions, it is incapable of capturing the subtleties and fine-grained details conveyed by visual features. As a result, both the comprehensive understanding of opinions of users and the capacity to accurately predict sentiments are compromised.

In addition, some models use insufficiently complex image representation techniques, such as raw pixel values or handcrafted features, to capture the complex visual characteristics of an image. These techniques lack the ability to extract high-level characteristics and patterns. As a result, the representational capacity of these models is limited, resulting in suboptimal performance in visual analysis tasks [20], [21], [22].

In addition, some models use conventional machine learning classifiers with fully connected layers, which may not fully exploit the representational power of deep learning models. In the absence of more sophisticated architectures and techniques, models are unable to learn complex relationships and capture intricate data patterns, resulting in subpar performance in visual sentiment analysis tasks [23], [24], [25].

Another obstacle is the absence of effective fusion strategies for integrating diverse visual characteristics. Various domains, such as Fourier Transform, Wavelet Transform, DCT, and Convolutional Transform, provide distinctive perspectives on the visual content of images. Existing models, however, frequently struggle to effectively combine these diverse features, resulting in suboptimal performance and limited discrimination capabilities [26], [27], [28], [31].

To improve the performance of visual analysis system authors are using Nature-inspired algorithms that can be employed for image feature selection to identify a subset of pertinent features that enhance precise image categorization or analysis [32], [33], [37].

Lastly, many existing models lack robustness and scalability, especially for real-time scenarios. When processing large volumes of data and delivering timely

insights to businesses, the computational efficiency and speed of the models become crucial. Nevertheless, several models suffer from computational bottlenecks, resulting in delays that limit their applicability for real-time visual sentiment analysis [29], [30], [38].

In order to develop more effective models for visual sentiment analysis, it is crucial to address these limitations. It is possible to improve the accuracy, representational ability, and scalability of these models by leveraging advanced deep learning techniques, incorporating diverse visual features, and designing efficient fusion strategies.

3. Proposed Methodology

After briefly reviewing the existing models used for identification of sentiments from visual datasets and samples, it is noticed that these models are either extremely sophisticated or extremely simple. It is evident that these models perform less well in large-scale use cases or are extremely sophisticated in real-time circumstances. This section addresses these concern and discusses the advancement of an effective multidomain framework for visual sentiment analysis of images using fusion-based deep learning technique. As per Figure 1, the proposed model that represents visual datasets into multidomain feature sets. These features are selected using an efficient Dual Moth Flame Optimizer (DMFO), which assists in increasing interclass variance, while minimizing intraclass variance levels. These features are classified into different sentiment classes via an efficient mechanism, which fuses CNN with an ensemble set of classifiers.

To perform multidomain feature analysis, the model initially extracts frequency components by applying Fourier Transform on the input data samples via equation 1,

$$F(x) = \sum_{j=0}^{N-1} x_j * \left[\cos\left(2 * pi * i * \frac{j}{N}\right) - \sqrt{-1} * \sin\left(2 * pi * i * \frac{j}{N}\right)\right] \quad (1)$$

Where, x represents intensity levels of pixels for the visual datasets, while N represents their cumulative dimensions across rows & columns.

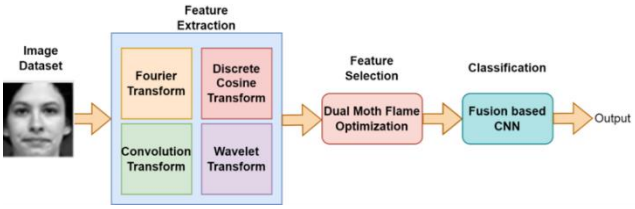

Fig 1. Design of Proposed Visual Sentiment Analysis Process

This is cascaded with Discrete Cosine Transform, which assists in estimating entropy levels via equation 2,

$$D(x) = \frac{1}{\sqrt{2 * N}} * x_i \sum_{j=1}^{N} x_j * \cos\left[\frac{\sqrt{-1} * (2 * i + 1) * \pi}{2 * N}\right] \quad (2)$$

Both these features are extended via a 2D Convolutional Layer, which uses Leaky Rectilinear Unit (LReLU) to activate the visual image pixels and is estimated via equation 3,

$$C(x) = \sum_{a=0}^{2m} \sum_{b=0}^{2n} x(m + a, n + b) * LReLU\big(x(m + a, n + b)\big) \quad (3)$$

Where, m, n represents the 2D Window Dimensions, which are varied between 2x2 to 256x256, while a, b are similar stride dimensions which are varied between 3x3 to 9x9 and assist in the extraction of feature sets with a high density. The activation function via equation 4 is applied,

$$LReLU(x) = l_a * x, when\ x < 0\ , else\ LReLU(x) = x \quad (4)$$

Here l_a represents an augmented factor used to keep positive feature value sets. While these features assist in representing visual datasets into spatial & frequency domains. The Fourier, Cosine and Convolutional Components are extended via extraction of Approximate & Detail Wavelet Coefficients. This is done via equations 5 and 6, where input image pixels and their consecutive pixels are used for analysis.

$$W(Approx) = \frac{x_i + x_{i+1}}{2} \quad (5)$$

$$W(Detail) = \frac{x_i - x_{i+1}}{2} \quad (6)$$

The need for these features arises from their collective ability to provide a holistic understanding of the sentiments conveyed in visual data. While each type of

feature offers unique insights, their combination enables a more comprehensive and accurate sentiment analysis. This multidimensional approach is particularly important in complex scenarios where sentiments are subtly embedded in visual cues across different domains.

All these features are fused in order to form an augmented Visual Sentiment Feature Vector (VSFV), which might contain feature-level redundancies. These redundancies are removed using novel Dual MFO Model, which assists in retaining high variance features via the following process,

- The DMFO Model Initializes an augmented set of NM Moths via equation 7,

$$NF = STOCH(LM * N(VSFV), N(VSFV)) \quad (7)$$

Where, LM represents Learning Rate of the MFO process, while NF represents total Number of Features extracted via stochastic (STOCH) operations.

- Based on these features, inter-class variance & intra-class variance is estimated via equation 8 and 9 as follows,

$$InterC(k) = \frac{1}{N-1} * \sum_{i=1}^{N} \left(x_i(j) - \sum_{j=1}^{N} \frac{x_j(k)}{N} \right)^2 \quad (8)$$

$$IntraC(k,l) = |Inter(k) - Inter(l)| \quad (9)$$

- Based on this evaluation, Dual Moth Fitness is estimated via equation 10,

$$fm = \frac{\sum_{i=1}^{NC} Inter(i)^2}{\sum_{i=1}^{NC} \sum_{j=1}^{NC} Intra(i,j)} \quad (10)$$

Where, NC represents total Number of Classes which are used for sentiment analysis.

- After repetition of the process for NM Moths, the threshold value for fitness is calculated through equation 11,

$$fth = \frac{1}{NM} \sum_{i=1}^{NM} fm(i) * LM \quad (11)$$

- Moths with $fm > fth$ are passed to Next Iteration, while other are discarded and replaced with New Moths via equations 7, 8 and 9, which replicates the Moth Flame characteristics.
- After NI Iterations, this procedure is repeated to create new
  Moths, which helps to identify highly variable feature set.

At the end of NI Iterations, features with highest Moth fitness levels are selected, and used for further classification operations. Due to this, the model is able to identify features which have higher interclass variance and low intraclass variance levels. These features are classified via a fusion of CNN and Ensemble classification process. This process is represented in figure 2, where the final layer of CNN process is replaced with Ensemble Classification operations for identification of sentiments.

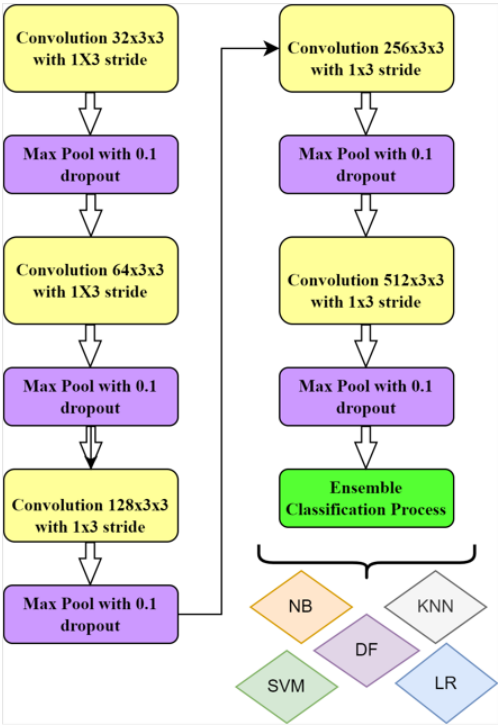

Fig.2. Design of the Fusion based Classification Process

The model replaces Fully Connected Neural Network with an ensemble classification process is illustrated in Figure 2. The CNN Model also uses Convolutional, Max Pooling, and Drop- out layers in order to improve the feature efficiency levels. This is done via equation 12, where

Convolutional Operations are applied along with LReLU activation process.

$$C(p) = \sum_{a=0}^{2m}\sum_{b=0}^{2n} p(m+a, n+b) * LReLU\big(p(m+a, n+b)\big) \quad (12)$$

Where, p are the high variance features which are extracted the DMFO process. In order to preserve high variance feature sets, the convolutional features are passed through an enhanced set of Max Pooling & Dropout layers. This process is repeated for different set of layers. Once the features pass through multiple layers, then an ensemble classification process is used to classify the processed features into different sentiment classes. Configurations of these classifiers is discussed in Table 1 as follows,

Table 1. Hyper parameters for different classifiers

Classifier	Parameter Used
Naïve Bayes	Prior Configurations (P) are calculated via equation 13, $P = \frac{1}{N-1} * \sum_{i=1}^{N}\left(x_i - \sum_{j=1}^{N}\frac{x_j}{N}\right)^2 \dots (13)$ Smoothing Value (SV) is calculated via equation 14, $SV = \frac{P}{\sum_{j=1}^{N}\frac{x_j}{N}} \dots (14)$
Deep Forest	Total Forests (TF) are estimated via equation 15, $TF = 10 * N(Features) \dots (15)$ Split Sample Configuration (SSC) is calculated via equation 16, $SSC = \frac{N(Feat)}{N(Classes)} \dots (16)$ Where, $(Features)\ and\ N(Classes)$ represents total features estimated by the convolutional layers and sentiment classes present in the datasets & their respective samples.
kNN	k=5, for single feature mapping during classification operations
Logistic Regression	Tolerance of Error (E), is estimated via equation 17, $E = \frac{0.01}{N(Classes)} \dots (17)$ Number of Iterations (NI) is estimated via equation 18, $NI = 10 * N(Features) \dots (18)$
Support Vector Machine	Level of Regularization (C) is calculated via equation 19, $C = \frac{1}{N(Classes)} \dots (19)$ Tolerance of Error ($tol = 0.0001$)

The output of each classifier is estimated individually, and based on this output the final sentiment class is evaluated via equation 20,

$$c(out) = A(NB) * C(NB) + A(DF) * C(DF) + A(kNN) * C(kNN) + A(LR) * C(LR) + A(SVM) * C(SVM) \quad (20)$$

Where, C represents the output class, while A represents the testing accuracy of different models & processes. The model can distinguish between various sentiment classes with a high level of efficiency based on this evaluation method. This model's performance is estimated using different efficiency metrics and in next section, they are compared with conventional models.

4. Experimental Results Analysis and Comparison

The proposed model in order to identify different visual sentiments from images uses a combination of multidomain features, Dual Moth Flame Optimization for feature selection and fusion-based CNN which fuses CNN with ensemble classification operations.

4.1. Dataset Description

The literature contains a large number of datasets for sentiment analysis of images. The datasets listed below are used for the research: Using a number of datasets facial expression are studied including the CohnKanade Dataset (CK+) [34], the JAFFE dataset [35], and the FER2013 dataset [36]. The FER2013 dataset consists about 35,887, 48 x 48-pixel grayscale images from which 28,709 were used for training and 3589 were used for testing [36]. The JAFFE dataset was assembled by staff psychologists from Kyushu University for their studies. They used Japanese women models to construct the dataset which contains facial expression with variations. For a collection this size, there aren't nearly enough frontal images in JAFFE (213 of 10 persons) [35]. CK+ [34] is a commonly used investigator database that includes seven gestures for 123 different individuals. It had 593 picture data with subjects that reflected the seven key emotional categories. For this work, out of different emotional categories considering angry, disgust as negative sentiment category, happy as positive and neutral. The system's main objective is to determine the polarity, or whether a face expression is positive, neutral, or negative, in an image. The sample of the dataset is given below.

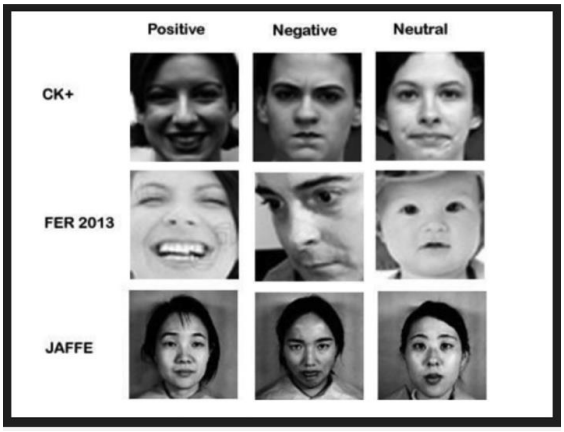

Fig.3 Datasets sample [14].

4.2 Evaluation Metrics

The model's performance was assessed using Precision (P), Accuracy (A), Recall (R), and F1-score (F1) equations. 21, 22, 23 & 24 as follows,

$$P = \frac{t_p}{t_p + f_p} \quad (21)$$

$$A = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \quad (22)$$

$$R = \frac{t_p}{t_p + f_n} \quad (23)$$

$$F1 = \frac{2 * P * R}{P + R} \quad (24)$$

Where, t_p, f_p, t_n & f_n represent values for true & false rates. To evaluate the potential performance of the Fusion-based CNN, the outcomes are contrasted with numerous existing machine learning models.

4.2. Results analysis and comparison

The proposed model designed for Visual sentiment analysis outperforms well on the following observations shown below: The very first observation is how the feature selection plays an important role in the design of system. Figure 4 presents accuracy percentages for different methods of feature selection when Moth flame optimization (MFO)is applied as feature selection, without MFO(WMFO) i.e. when no feature selection was applied and when proposed DMFO applied which is the improved version of MFO which includes the concept of increasing interclass variance, while minimizing intraclass variance levels on three datasets: CK+, FER2013, and JAFFE.

When no feature selection was used i.e. WMFO the model achieves 94.2% accuracy on CK+, 67.2% on FER2013, and 85.8% on JAFFE. When MFO was used the model achieves 97.5% accuracy on CK+, 71.5% on FER2013, and 89.4% on JAFFE. When DMFO was used the model achieves 99.5% accuracy on CK+, 73.5% on FER2013, and 90.85% on JAFFE. The accuracy consistently improves across the three methods, with DMFO-Fusion Based CNN achieving the highest accuracy. Figure 4 shows the model outperforms well when the novel Dual MFO is used for feature selection when compare with base Moth Flame Optimization (MFO) is used and when it is not used i.e. without MFO (WMFO).

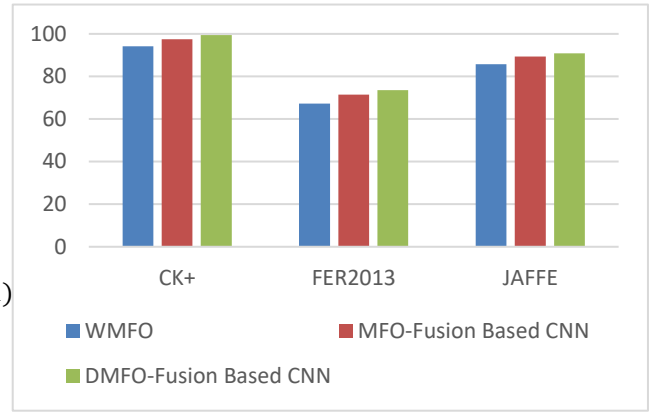

Fig.4 Performance comparison of optimization proposed model

The other observations is when performance of model was compared with different machine learning algorithms and also with proposed visual sentiment classification models like Gabor filters [16], AlexNet-DCNN[17], Inception V3 based model [14], Resnet101[18], CNN[19] and showcase high performance levels.A graphic illustration of the comparison is shown in Fig.5,6,7. Well-known machine learning techniques including LR, KNN, NB, SVM, DF and ensemble model all are used to assess the accuracy, F1-score, precision and recall values of the suggested model. Every machine learning model is evaluated using the FER2013, CK+, and JAFFE to gauge their level of performance. Our Fusion-based CNN model outperforms various other machine learning methods in terms of performance.

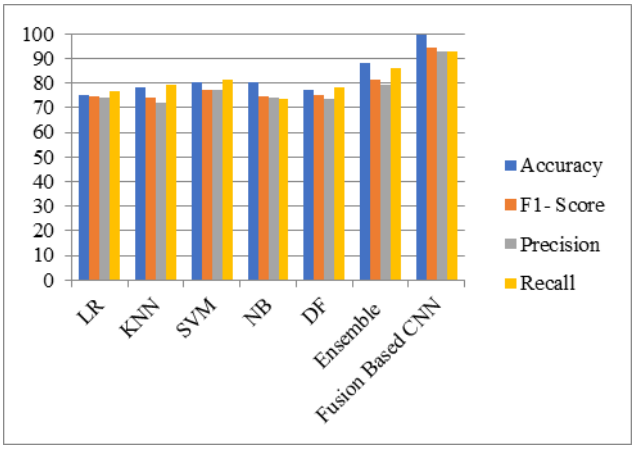

Fig 5: Comparative analysis of machine learning models with the Fusion-based CNN for CK+ dataset

Figure 5 presents a comparative examination of machine learning models using the fusion-based CNN on the CK+ dataset. The LR model demonstrates a satisfactory overall performance, with a balanced precision and recall. On the other hand, the KNN model exhibits competitive accuracy but somewhat lower precision, suggesting the possibility of misclassifications. The Support Vector Machine (SVM) demonstrates strong performance in all evaluation metrics, exhibiting a favourable equilibrium between precision and recall. The Naive Bayes (NB) classifier achieves good accuracy, although there is a minor disparity between precision and recall. The Decision Forest (DF) model performs well, albeit slightly lower than SVM and NB in terms of accuracy. The ensemble model surpasses the individual models by a significant margin, highlighting the efficacy of combining diverse classifiers. Lastly, the Fusion-based Convolutional Neural Network (CNN) achieves exceptional performance across all metrics, underscoring the potency of deep learning with fusion strategies.

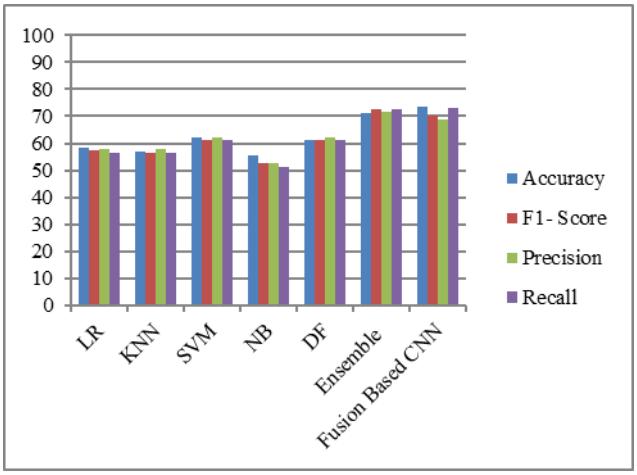

Fig 6: Comparative analysis of machine learning models with the Fusion -based CNN for FER2013 dataset.

Figure 6 presents a comparative analysis of machine learning models using the Fusion-based CNN for the FER2013 dataset. The results show that LR (Logistic Regression) achieves moderate accuracy with balanced precision and recall. KNN (K-Nearest Neighbours) performs similarly to LR, with a trade-off between precision and recall. SVM (Support Vector Machine) achieves slightly higher accuracy and balanced precision-recall. NB (Naive Bayes) exhibits lower accuracy with a trade-off between precision and recall. DF (Decision Forest) performs well with balanced precision-recall, similar to SVM. The ensemble model outperforms individual models, indicating the effectiveness of combining different classifiers. The Fusion Based CNN achieves the highest accuracy among individual models, demonstrating the power of deep learning with fusion strategies.

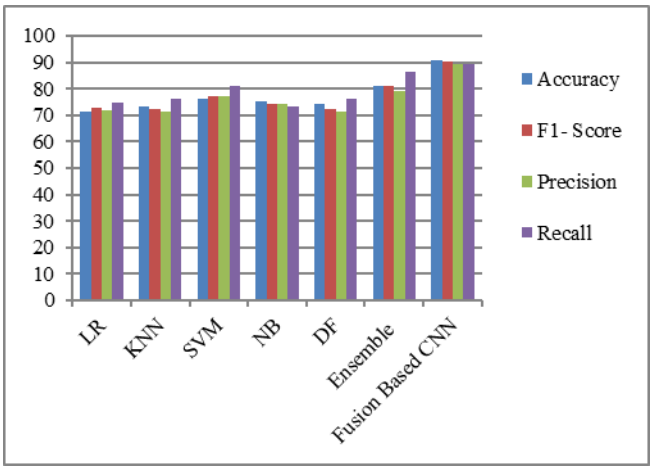

Fig 7: Comparative analysis of machine learning models with the Fusion- based CNN for JAFFE dataset.

Figure 7 presents a comparative analysis of machine learning models using the Fusion-based CNN for the JAFFE dataset. The LR model demonstrates moderate accuracy with balanced precision and recall. The KNN model exhibits higher accuracy with balanced precision-recall. The SVM model achieves higher accuracy and balanced precision-recall. The NB model shows good overall performance with a balance between precision and recall. The DF model performs well with balanced precision-recall. The ensemble model outperforms individual models, indicating the effectiveness of combining diverse classifiers. Finally, the Fusion based CNN model achieves the highest accuracy among all

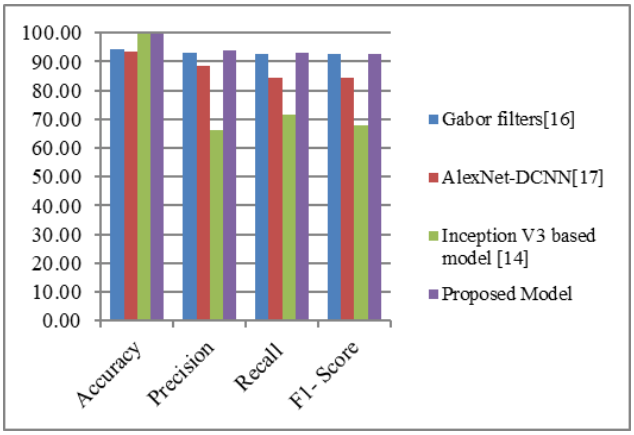

Fig 8. Comparative analysis of existing work with the Proposed model for CK+ dataset.

models, showcasing the power of deep learning with fusion strategies.

As per figure 8 it was determined that the proposed framework exhibited better accuracy for CK+ dataset when compared with different existing models Gabor filters [16], AlexNet-DCNN[17], Inception V3 based model [14] . Here, Inception V3 has commendable accuracy, although encounters challenges in maintaining balanced precision and recall. This indicates potential

obstacles in effectively managing specific classes or patterns within the data. The Proposed Model exhibits robustness in attaining a high level of accuracy while simultaneously preserving a balanced level of precision and recall. This demonstrates the successful acquisition of useful features and the design of an effective model structure.

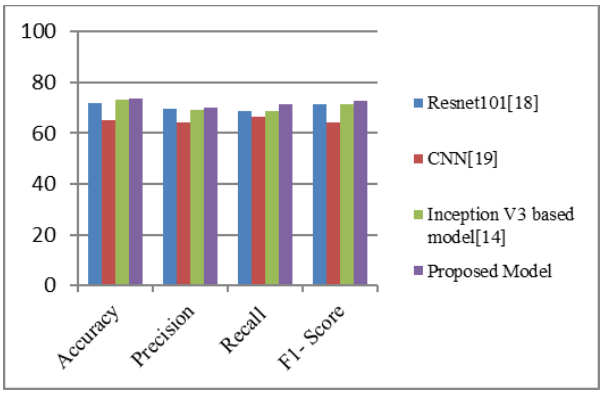

Fig 9. Comparative analysis of existing work with the Proposed model for FER2013 Dataset

Figure 9 shows The accuracy of Inception V3 and the Proposed Model is similar, but the Proposed Model has slightly superior precision and recall because to the potential incorporation of ensemble approaches or fusion strategies, which combine the capabilities of numerous models to enhance overall performance.

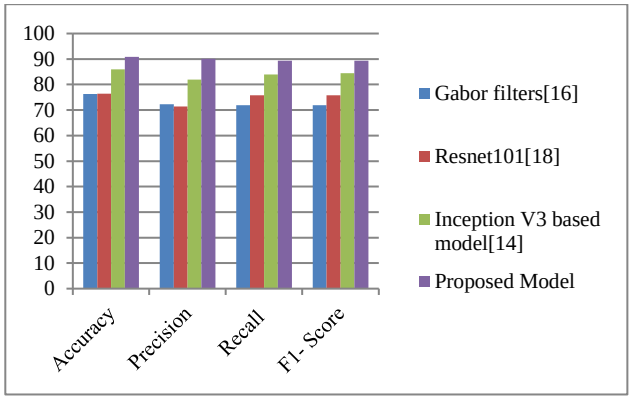

Fig 10. Comparative analysis of existing work with the Proposed model for JAFFE Dataset.

As per figure 10, it was determined that the Proposed model exhibited better accuracy for JAFFE dataset when compared with different existing models Gabor filters [16], Resnet101[18], Inception V3 based model [14]. Both Inception V3 and the Proposed Model exemplify the efficacy of deep learning in image classification problems, attaining superior accuracy and enhanced overall performance in contrast to conventional approaches. This research makes a substantial contribution to the field of sentiment analysis by introducing a novel, efficient, and robust multidomain visual sentiment analysis model. The innovative use of DMFO for feature selection, combined with a customized CNN architecture, positions this work as a significant advancement in harnessing the power of image data for understanding user sentiments.

Table 2: Performance Results

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CK+	99.5	94.1	92.9	92.6
FER2013	73.5	70.1	71.4	72.9
JAFFE	90.85	90.1	89.4	89.4

A thorough representation of the performance indicators is shown in Table 1. This table presents the findings of a study that used an image sentiment analysis model that combines several visual features and employing an effective Dual Moth Flame Optimisation (DMFO) Model for feature selection. Furthermore, to enhance interpretability and classification performance, a unique CNN architecture is built and coupled with a set of classifiers. The three distinct datasets are analysed CK+, FER2013, JAFFE which achieve 99.5%, 73.5% and 90.85% of accuracy. Despite its lower size, the CK+ dataset consistently yields the best results across all research. Similarly, the others evaluation parameters are also considered like precision, recall and F1 score and the fusion-based CNN exhibited better recall, precision and

F1-score values making it highly suitable to a wide range of situations of high-scalability real-time classification scenarios.

5. Conclusion and Future Scope

In this paper, an efficient multidomain model for visual sentiment analysis using fusion-based deep learning technique is proposed. The shortcomings of existing models in capturing and accurately representing the diverse visual properties of product images is addressed. Combining multiple visual features using an efficient Dual Moth Flame Optimization (DMFO) Model for feature selection, our model overcomes these limitations. In addition, a custom fusion-based CNN architecture for visual sentiment analysis and integrated it with a group of classifiers to improve interpretability and classification performance is designed.

Multiple empirical dataset and sample evaluations demonstrated the superiority of our Fusion based CNN. It outperformed previous techniques by achieving about 99.5%, 73.5% and 90.85% of accuracy on CK+, FER2013, JAFFE datasets. Due to its capacity to capture and represent multidomain visual features, the model's applicability to real-time scenarios is promising. Using our proposed model, businesses can efficiently analyse product feedback and make well-informed decisions regarding product enhancements.

Even though our model represents a significant advance in visual sentiment analysis, there are numerous avenues for future investigation and development such as extending the evaluation on more extensive and diverse datasets can provide a more thorough understanding of its performance across a variety of domains. It can be advantageous to investigate and implement advanced fusion strategies beyond the ensemble of classifiers. Also, exploring and developing more advanced feature selection techniques may improve the efficiency and efficacy of capturing pertinent visual features. By investigating these avenues, researchers can advance the field of visual sentiment analysis, allowing users to gain deeper insights of visual contents which can subsequently improve the analysis of the opinions of the user reviews.

Author contributions

Siddhi Kadu: Conceptualization, Methodology, Software, Field study. **Bharti Joshi:** Data curation, Writing-Original draft preparation, Software, Field study.**Pratik Agrawal:** Visualization, Investigation, Writing-Reviewing and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Xue, L. Y., Mao, Q. R., Huang, X. H., & Chen, J. (2020). NLWSNet: a weakly supervised network for visual sentiment analysis in mislabeled web images. Frontiers of Information Technology & Electronic Engineering, 21(9), 1321-1333.

[2] Mittal, N., Sharma, D., & Joshi, M. L. (2018, December). Image sentiment analysis using deep learning. In 2018 IEEE/WIC/ACM international conference on web intelligence (WI) (pp. 684-687). IEEE.

[3] Ortis, A., Farinella, G. M., & Battiato, S. (2020). Survey on visual sentiment analysis. IET Image Processing, 14(8), 1440-1456.

[4] She, D., Yang, J., Cheng, M. M., Lai, Y. K., Rosin, P. L., & Wang, L. (2019). Wscnet: Weakly supervised coupled networks for visual sentiment classification and detection. IEEE Transactions on Multimedia, 22(5), 1358-1371.

[5] J. Xu, Z. Li, F. Huang, C. Li and P. S. Yu, "Visual Sentiment Analysis With Social Relations-Guided Multiattention Networks," in IEEE Transactions on Cybernetics, vol. 52, no. 6, pp. 4472-4484, June 2022, doi: 10.1109/TCYB.2020.3027766.

[6] Chen, J., Mao, Q., & Xue, L. (2020). Visual sentiment analysis with active learning. IEEE Access, 8, 185899-185908.

[7] Mamatha, M., Shivakumar, S., Thriveni, J., & Venugopal, K. R. (2022, July). Visual Sentiment Classification of Restaurant Review Images using Deep Convolutional Neural Networks. In 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) (pp. 1-6). IEEE.

[8] S. Urolagin, J. Nayak and U. R. Acharya, "Gabor CNN Based Intelligent System for Visual Sentiment Analysis of Social Media Data on Cloud Environment," in IEEE Access, vol. 10, pp. 132455-132471, 2022, doi: 10.1109/ACCESS.2022.3228263.

[9] Wu, Z., Meng, M., & Wu, J. (2020). Visual sentiment prediction with attribute augmentation and multi-attention mechanism. Neural Processing Letters, 51, 2403-2416.

[10] Song, K., Yao, T., Ling, Q., & Mei, T. (2018). Boosting image sentiment analysis with visual attention. Neurocomputing, 312, 218-228.

[11] T. Zhou, J. Cao, X. Zhu, B. Liu and S. Li, "Visual-Textual Sentiment Analysis Enhanced by Hierarchical Cross-Modality Interaction," in IEEE Systems Journal, vol. 15, no. 3, pp. 4303-4314, Sept. 2021, doi: 10.1109/JSYST.2020.3026879.

[12] S. Lee, D. K. Han and H. Ko, "Multimodal Emotion Recognition Fusion Analysis Adapting BERT With

Heterogeneous Feature Unification," in IEEE Access, vol. 9, pp. 94557-94572, 2021, doi: 10.1109/ACCESS.2021.3092735.

[13] Truong, Q. T., & Lauw, H. W. (2017, October). Visual sentiment analysis for review images with item-oriented and user-oriented CNN. In Proceedings of the 25th ACM international conference on Multimedia (pp. 1274-1282).

[14] Meena, G., Mohbey, K. K., & Kumar, S. (2023). Sentiment analysis on images using convolutional neural networks based Inception-V3 transfer learning approach. International Journal of Information Management Data Insights, 3(1), 100174.

[15] J. Zhang, X. Liu, Z. Wang and H. Yang, "Graph-Based Object Semantic Refinement for Visual Emotion Recognition," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 5, pp. 3036-3049, May 2022, doi: 10.1109/TCSVT.2021.3098712

[16] Boughida, A., Kouahla, M. N., & Lafifi, Y. (2022). A novel approach for facial expression recognition based on Gabor filters and genetic algorithm. Evolving Systems, 13(2), 331-345.

[17] Fallahzadeh, M. R., Farokhi, F., Harimi, A., & Sabbaghi-Nadooshan, R. (2021). Facial expression recognition based on image gradient and deep convolutional neural network. Journal of AI and Data Mining, 9(2), 259-268.

[18] Hua, W., Dai, F., Huang, L., Xiong, J., & Gui, G. (2019). HERO: Human emotions recognition for realizing intelligent Internet of Things. IEEE Access, 7, 24321-24332.

[19] Agrawal, A., & Mittal, N. (2020). Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. The Visual Computer, 36(2), 405-412

[20] F. Alzamzami and A. E. Saddik, "Transformer-Based Feature Fusion Approach for Multimodal Visual Sentiment Recognition Using Tweets in the Wild," in IEEE Access, vol. 11, pp. 47070-47079, 2023, doi: 10.1109/ACCESS.2023.3274744.

[21] S. Urolagin, J. Nayak and U. R. Acharya, "Gabor CNN Based Intelligent System for Visual Sentiment Analysis of Social Media Data on Cloud Environment," in IEEE Access, vol. 10, pp. 132455-132471, 2022, doi: 10.1109/ACCESS.2022.3228263.

[22] Y. Su, W. Zhao, P. Jing and L. Nie, "Exploiting Low-Rank Latent Gaussian Graphical Model Estimation for Visual Sentiment Distributions," in IEEE Transactions on Multimedia, vol. 25, pp. 1243-1255, 2023, doi: 10.1109/TMM.2022.3140892.

[23] J. Kossaifi et al., "SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 3, pp. 1022-1040, 1 March 2021, doi: 10.1109/TPAMI.2019.2944808.

[24] L. Stappen, A. Baird, L. Schumann and B. Schuller, "The Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements," in IEEE Transactions on Affective Computing, vol. 14, no. 2, pp. 1334-1350, 1 April-June 2023, doi: 10.1109/TAFFC.2021.3097002.

[25] J. He, S. Mai and H. Hu, "A Unimodal Reinforced Transformer With Time Squeeze Fusion for Multimodal Sentiment Analysis," in IEEE Signal Processing Letters, vol. 28, pp. 992-996, 2021, doi: 10.1109/LSP.2021.3078074.

[26] Qayyum, H., Majid, M., Anwar, S. M., & Khan, B. (2017). Facial expression recognition using stationary wavelet transform features. Mathematical Problems in Engineering, 2017.

[27] Avcı, D., Sert, E., Özyurt, F., & Avcı, E. (2024). MFIF-DWT-CNN: Multi-focus ımage fusion based on discrete wavelet transform with deep convolutional neural network. Multimedia Tools and Applications, 83(4), 10951-10968..

[28] You, N., Han, L., Zhu, D., & Song, W. (2023). Research on image denoising in edge detection based on wavelet transform. Applied Sciences, 13(3), 1837.

[29] J. Zeng, J. Zhou and C. Huang, "Exploring Semantic Relations for Social Media Sentiment Analysis," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2382-2394, 2023, doi: 10.1109/TASLP.2023.3285238.

[30] K. Ye, N. H. Nazari, J. Hahn, Z. Hussain, M. Zhang and A. Kovashka, "Interpreting the Rhetoric of Visual Advertisements," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 4, pp. 1308-1323, 1 April 2021, doi: 10.1109/TPAMI.2019.2947440.

[31] Truong, Q. T., & Lauw, H. W. (2023, February). Concept-oriented transformers for visual sentiment analysis. In Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (pp. 1111-1119).

[32] Chatterjee, S., Saha, D., Sen, S., Oliva, D., & Sarkar, R. (2023). Moth-flame optimization based deep feature selection for facial expression recognition using thermal images. Multimedia Tools and Applications, 1-24.

[33] Mirjalili, S. (2015). Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm. Knowledge-based systems, 89, 228-249.

[34] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010, June). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In 2010 ieee computer society

conference on computer vision and pattern recognition-workshops (pp. 94-101). IEEE.

[35] Akhand, M. A. H., Roy, S., Siddique, N., Kamal, M. A. S., & Shimamura, T. (2021). Facial emotion recognition using transfer learning in the deep CNN. Electronics, 10(9), 1036.

[36] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Bengio, Y. (2013). Challenges in representation learning: A report on three machine learning contests. In Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20 (pp. 117-124). Springer berlin heidelberg.

[37] Deshmukh, P. S., Gaikwad, A. K.., & Agrawal P. K. (2024). Deep Learning Based Dual-level Bioinspired Model for Parkinson's Disease Detection. International Journal of Intelligent Systems and Applications in Engineering, 12(13s), 179–187.

[38] Agrawal, P.K. A Novel Mapper Machine Learning Algorithm for Semantic Domain Mapping for Domain Database Updation. SN COMPUT. SCI. 4, 536 (2023).