Multimedia Tools and Applications (2025) 84:24457–24485
https://doi.org/10.1007/s11042-024-20518-y

1251: COMPUTER VISION WITH SMALL DATA: A FOCUS ON HUMAN AND ANIMALS

# FSTL-SA: few-shot transfer learning for sentiment analysis from facial expressions

Gaurav Meena[1] · Krishna Kumar Mohbey[1] · K. Lokesh[1]

Received: 7 May 2024 / Revised: 4 October 2024 / Accepted: 12 December 2024 /
Published online: 19 December 2024

**Abstract**
The primary objective of sentiment analysis is determining a person's viewpoint on a subject or the document's overall contextual polarity. When a significant quantity of labeled data is provided for the target task, deep learning is demonstrated to be successful for sentiment analysis in facial expressions. Nonetheless, there is ongoing research toward training a deep learning model with few observations of labeled data such that it may generalize effectively on a novel task called Few-shot Learning. This research proposes a unique few-shot transfer learning (FSTL-SA) framework for nonverbal communication sentiment polarity categorization. First, a deep learning model is trained using a large publicly available dataset of CK + and FER2013 on anger, fear, and surprise in the source domain. The trained model was then finetuned in the target domain using the N-way-k-shot approach on happy, sad, and neutral classes and FER2013 and CK +. In addition, we employed two-stage semi-supervised few-shot learning to address the labeled data scarcity. The proposed framework performed better than cutting-edge methods on nonverbal Sentiment Analysis. The proposed deep convolutional neural network (DCNN) achieved an accuracy of 75.33% in the source domain, and the FSTL-SA method achieved an average accuracy of 61% for 100 shots. Additionally, we achieved an accuracy of 82% on a single semi-supervised approach for 60-shot.

**Keywords** Transfer learning · Few-shot learning · Deep Learning · Sentiment Analysis

## 1 Introduction

Sentiment recognition based on facial expressions is a fascinating study area with implications in various sectors, including health, security, and human–machine interactions. The fundamental objective of sentiment analysis (SA) is to identify an individual's perspective on a topic or the overall contextual polarity of a document. With the rise of

✉ Gaurav Meena
gaurav.meena@curaj.ac.in

[1] Department of Computer Science, Central University of Rajasthan, Ajmer, India

social networks and mobile devices, consumers are capturing many photographs and videos to document all aspects of their lives daily and globally. People discuss their travel experiences, thoughts on events, and more. While humans primarily communicate through words, they also express emotions and emphasize specific points in their speech using facial expressions. Facial expressions are a vital communication component and serve as one of the main ways people convey their feelings. Figure 1 shows samples of facial expressions.

The expressions on a person's face convey nonverbal messages. Moreover, it is utilized in nonverbal communication and plays a significant role in recognizing emotions. The amount of data generated by these social media sites is expanding exponentially [1], encompassing more than just words. Users use significant visual data to express their thoughts and feelings intuitively. This information can be implemented for analysis. Traditionally, SA is carried out through text. Considering that feelings and emotions are communicated visually, traditional SA algorithms based on text are simply inappropriate.

It is a complex undertaking since it requires recognizing emotions through visual representations. There is no doubt that visual sentiment analysis (VSA), the study of predicting people's emotions based on visual representations, has garnered much attention [2]. Several machine-learning approaches have been utilized for VSA [3]. Recently, deep neural networks, especially CNNs [4], have received significant interest in visual sentiment analysis and have shown significant results due to exceptional feature representations [5]. However, training deep learning models requires large amounts of labeled data, which is time-consuming and expensive. In contrast, finetuning, an essential aspect of Transfer Learning, requires less labeled data than training a model from scratch and leverages knowledge gained from the previously trained model to adapt to new tasks more efficiently.

Few-shot learning, a technique that leverages finetuning in transfer learning, enables models to adapt efficiently to new tasks with minimal labeled data. This approach offers practical applications in public safety, healthcare, and business, where collecting large datasets is often time-consuming and costly. Conventional image classification algorithms typically require substantial training data to develop effective models. However, in practical scenarios, the availability of representative data is sometimes inadequate, leading to overfitting during network construction. Research on few-shot learning is

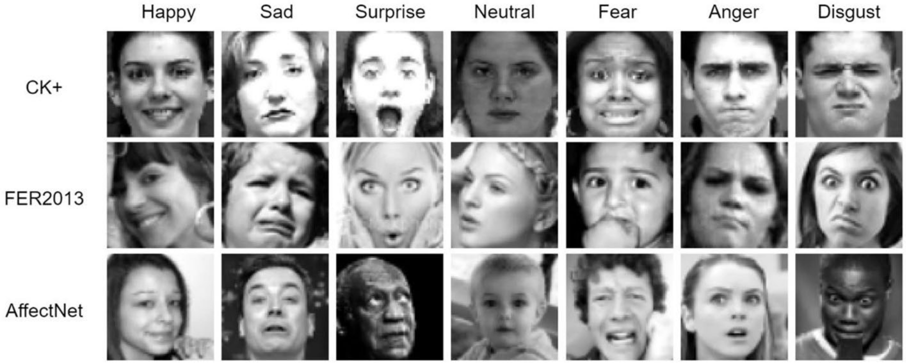


**Fig. 1** Facial expression example

well-recognized for effectively addressing this issue. Several deep learning models that exhibit exceptional performance cannot be deployed in this field due to the annotated data scarcity. This challenge can be overcome by employing N-way-k-shot few-shot learning techniques, which expand the application of these high-performance models across various domains.

In 2003, Fei et al. proposed the concept of few-shot learning, highlighting the primary difficulty of effectively utilizing acquired information to learn a new category [6]. This strategy addresses the persistent need for extensive and all-encompassing datasets. Few-shot learning often involves learning the distinguishing features of a small set of labeled pictures to classify a new image. However, abundant samples of facial images are available online, which can be leveraged to reduce the scarcity of labeled data by using pseudo-labeled data. This involves assigning labels to unlabelled data and incorporating it into the training process. The technique, known as semi-supervised learning, enables models to utilize both labeled and unlabelled data, thereby improving performance and reducing the dependency on large, annotated datasets while still learning new categories effectively.

We proposed FSTL-SA (Few-Shot Transfer Learning for Sentiment Analysis) by incorporating FSTL with a two-stage semi-supervised learning approach, using pseudo-labeled data with a 99% confidence interval. The two stages consist of single semi-supervised learning and iterative semi-supervised learning. In the first stage, pseudo-labeled data is introduced to finetune the model. In contrast, in the second stage, the model iteratively refines its predictions, progressively improving accuracy and robustness in sentiment analysis tasks. The key contributions of the study are as.

- We proposed a convolutional neural network (CNN) to extract facial features and perform robust visual sentiment analysis.
- To the best of our knowledge, we introduced the first Few-Shot Transfer Learning framework for identifying visual sentiment using semi-supervised learning.
- Introduced a novel adaptive function for selecting pseudo-labeled data based on confidence intervals, improving the model's accuracy and robustness.
- Extensive experiments were conducted considering various factors to verify the proposed framework's generalization ability.

The rest of the paper is organized as follows: Sect. 2 contrasts our work with the existing literature. Section 3 describes the proposed framework for visual sentiment analysis using semi-supervised few-shot learning. Subsequently, the data processing, experimental setup, and findings are explained in detail in Sect. 4. Section 5 concludes our study and suggests future paths for few-shot learning.

## 2 Related work

Emotions are valuable and self-explanatory in daily interactions between people. People's facial expressions convey emotions. Though difficult and time-consuming, facial expression recognition (FER) is helpful in many fields, including human–computer interaction, emotionally charged robots, and healthcare [7–9]. Even with FER's advancements increasing its efficacy, obtaining high precision remains challenging [10]. The six universal human emotions are fear, surprise, disgust, happiness, sadness, and anger. SA seeks to ascertain people's attitudes toward a subject or the desired emotional response the writer hopes to elicit from the audience. This research field's tasks are both demanding and practical. SA has various practical uses, as opinions impact many human actions in both economic and social situations. There has been a minimal attempt to extract emotions from visual input, in contrast to the development of SA algorithms for text analysis. Even though scientific research has already accomplished notable results in the field of textual sentiment analysis in a variety of contexts [11, 12], it is still challenging to understand the mood of a text due to the inherent ambiguity of different languages (for example, ironic sentences), cultural factors, linguistic subtleties, and the difficulty of generalizing any text analysis solution. All of these factors contribute to the difficulty of understanding the mood of a text. Most individuals utilize SA, which analyses a message to ascertain the underlying emotion and categorize messages. Determine if a specific phrase (a review, a tweet, or a comment) represents a neutral, negative, or positive feeling. This is the fundamental purpose of SA. To be more specific, the objective of face image SA in this context is to determine if the input picture in question indicates positive, neutral, or negative emotion [13, 14]. Many studies have conducted experiments to categorize facial images into one of these categories.

VSA has several applications, including resource building. Its goal is to build dictionaries, corpora, and lexica that label opinion statements as either pro or con-polarity. Developing resources is not explicitly related to SA but can aid in SA and emotion detection in other ways. The work in this category has faced substantial obstacles due to word ambiguity, multilingualism, granularity, and differences in expressing thought among literary genres [15]. Hybrid, lexicon-based, and ML methods are the three primary categories of thought regarding sentiment classification [16]. Concerning opinion recognition, SA is the most famous and researched component [17]. SA did not gain traction as a hot new area of study in information management until 2000. Improved information management methods for business use might result from this SA. Psychology, psychiatry, and mental health researchers have recently focused on image-based sentiment analysis [18]. Automated emotion identification from images is critical for various current applications, including assisted living, health care, autism spectrum disorder diagnosis, human–computer interface, and social welfare initiatives [19]. As a result, the scientific community has turned its attention to SA in anticipation of potential applications.

Feature extraction and sentiment classification are the two main components of the conventional method for image-based SA. Furthermore, image preprocessing is necessary, which includes functions like cropping, scaling, normalizing, and face recognition. In a conventional sentiment analysis system, feature extraction from the processed image is essential. The existing methods use specialist techniques like linear discriminant analysis, discrete wavelet transform, and related techniques [20]. In the last stage, sentiments are classified using the retrieved attributes to understand them better. Neural networks, DL, transfer learning, and other ML techniques are frequently used [21]. Many systems, like patient assistance robots, medical treatment decision support, and information management, use SA. The ability to use visual information may potentially increase the quality of data. SA may be utilized for interdisciplinary marketing, information management research, etc.

The concept of VSA is currently being researched. Earlier work on emotional semantic image retrieval, which develops links between emotions and low-level visual properties to improve automatic picture categorization and retrieval, provides the foundation for most research in this new field. According to the research that was conducted in 2010, the objective of visual sentiment analysis is to classify images as either "positive" or "negative" [22]. The authors of this study researched the connections between the emotional content of a picture and its visual content. Their quantitative evaluations of feelings were derived from the accompanying text of each photograph, which is referred to as meta-data.

Learning from experience and generalizing concepts are both made possible by DL, a branch of AI. Computers can learn from real-world examples and grasp an issue thoroughly before concluding, all without human intervention [23]. In DL, "deep" means more than one underlying data layer. An extensive dataset containing many labeled data is required to train DL networks to autonomously extract features or parameters from a dataset. Using RNNs, CNNs, and Deep NN are only a few of the approaches made possible by DL, making it an essential component for image VSA. Many people turn to CNNs, a feed-forward model, for problems with image processing, identification, and prediction. Despite much study in this field, learning over new classes with few samples is a significant challenge in deep learning tasks. Learning with zero shots and one shot are two examples of few-shot learning. With zero-shot learning, the most limited premise is that there are no training samples to train the model. Lampert et al. [24] propose an attribute-based classification system that performs object identification using a human-specified high-level semantic attribute rather than training photos. Zhang et al. [25] constructed a max-margin framework for learning semantic similarity embedding using the seen class proportion as a similarity measure for unseen classes. In order to determine the connections between the visual classes produced by an algorithm, some research considers supplementary data like visual traits or plain language semantics. As supplementary data, Jetley et al. [26] use visual archetypal conceptions to deduce previously undetected classes. Recognization using learned classifiers is made possible by a novel one-step zero-learning architecture that Guo et al. [27] provide. The framework recommends utilizing pseudo labels to move samples from source classes.

## 2.1 Sentiment analysis using machine learning approaches

Analyzing the emotional content of an image is a challenging task in artificial intelligence, especially in the machine learning sector. The groundbreaking approach to SA proposed by [28] began with incorporating a wavelet energy characteristic into a face image. After extracting characters using Fisher's linear discriminants, the researchers used the KNN approach to categorize the afflicted person's emotional state. Classification in face recognition also made use of KNN [29]. Additionally, feature extraction was done using principal component analysis and non-negative matrix factorization. After [30] gathered local binary pattern histograms from numerous separate small portions of the image, combined them into a single feature histogram, and used the histogram to classify the subject's emotional state, a linear programming approach was employed to categorize their emotional state. They used an improved wavelet transform for the 2D picture [31]. They developed the contourlet transform to use a boosting technique to extract picture features for classification purposes. Facial recognition was integrated into the SA process, and the radial basis function was used for classification [32].

Several classification algorithms use SVM to predict an individual's emotional state based on the returned feature values. The researchers [33] investigated various face representations using several different SVM forms. These representations were based on local statistical characteristics and binary patterns. [34] Investigated a method for determining the qualities of an object based on its appearance from the outside. The local directed pattern was the name given to this methodology. Recent research [35] utilized SVM to investigate the utilization of two feature descriptors, namely the center of gravity descriptor and the facial landmarks descriptor, and their respective applications. [36] investigated several different categorization strategies for their proposed face geometry-based feature extraction. Feature extraction through the use of facial geometry was the primary focus of the investigations. Inadequate performance is the primary deficiency that is shared by all age-old methods.

With the help of the AlexNet-DCNN model [37], it is possible to identify the high-level characteristics associated with the different categories of feelings. After applying transfer learning to the proposed model, it is optimized. In contrast to the CK dataset, which has an accuracy of 93.66%, the CK+dataset has an average recognition accuracy of 93.66%. Based on the findings of the experiments carried out on the benchmark emotional dataset, it has been determined that the proposed model is effective and has the potential to enhance the functionality of existing FER systems. It was via the utilization of DCNN, which is well-known for its capability to work with image data, that they achieved their objectives [38]. Although GPUs can handle computation-intensive jobs for deep CNN, they consume little power due to their unparalleled performance. It is possible to get higher precision in identifying vital emotions through the FaceLiveNet Network using the Dense Face Live Net architecture [39]. This may be performed by increasing the accuracy from low to high. They utilized the JAFFE basic emotion identification model on the FER2013 primary emotion dataset for the first step of Dense Face Live Net for Two-Phase Transfer Learning. This provided the basis for the learning process. They achieved a seventy percent accuracy rate as a reward for their efforts. Furthermore, when a transfer learning model is utilized to acquire the ability to recognize emotions, the test's accuracy rate may reach 91.93%. The evidence suggests enhancing recognition accuracy by appropriately using transfer learning processes is possible.

### 2.2 Sentiment analysis using deep learning approaches

Deep learning models have made substantial advances in various fields, including natural language comprehension, object detection in pictures, semantic segmentation, and audio interpretation. However, one significant problem with deep learning models is their dependency on labeled data for optimal performance, which is frequently sparse in real-world circumstances. Neural networks are proposed as a sub-domain of ML called DL [40] to represent a high-level generalization of data processing via several layers of piling-up alternatively linear and nonlinear changes. Creating deep neural networks, composed of tens or even hundreds of layers arranged in a heap structure, has been one of the most important breakthroughs in processing speech, images, and text [41]. Different convolutional architectures were demonstrated to increase the FER2013 dataset's recognition accuracy [42]. Ensembles of the models under consideration were created using the bagging approach. Several fractalization and histogram equalization modifications were produced for the dataset under consideration. The prediction accuracy of the original ResNet50 Network was increased using transfer learning and simple topologies. A deep belief network (DBN) and NN were coupled in the study [43] for face recognition. While the NN was used to classify emotional feature qualities, the DBN was employed for unsupervised feature learning. In their self-selected facial expression images, [44] investigated the effectiveness of a standard CNN architecture with two convolutional-pooling layers.

Additionally, [45] considered the CNN ensemble; even though they trained one hundred CNNs, their final model only included a portion of those CNNs. After initializing its weights with the encoder weights of a stacked convolutional auto-encoder, researchers [46] trained their CNN using facial images. This CNN initialization has been demonstrated to outperform a CNN with a random beginning point. [47] Examined a hybrid deep learning architecture for image identification that included CNN and RNN. Examined a hybrid architecture that included transfer learning [48], SVM is used in this architecture to classify features from an AlexNet that has already been trained. The most recent study to examine the potential of CNN as a clustering approach was [49]. However, [50] evaluated several data augmentation techniques, one of which used artificial images to train a deep neural network. They discovered that combining synthetic pictures with various techniques enhanced the deep CNN's performance.

On the other hand, people are remarkably adept at picking up new abilities quickly and with little guidance. With just a few examples, a youngster who comprehends addition, for instance, may quickly use that understanding to learn multiplication. In the same way, a young toddler shown a few photographs of an unknown person may quickly recognize that person in many more pictures. One crucial area of research is closing the knowledge gap between AI systems and human learning capacities. Few-shot learning [51, 52] is a revolutionary ML technique devised to tackle the problem of learning from a small number of samples using supervised information.

The conventional approach commonly refers to a few-shot image classification task as an N*K-shot [53] challenge. The training set for few-shot learning is partitioned into categories, each consisting of several samples. During the training phase, a random selection of N categories of picture samples is made from the training set. The support set is created by choosing K samples (N*K pictures) from each category. Subsequently, a restricted quantity of samples is chosen from the remaining data in each of the N categories to act as the prediction object for the model, also known as the query set. When the value of K is very tiny, often fewer than ten, the classification job is known as a few-shot picture classification. When the value of K is set to 1, the job is simplified to a single instance picture classification task. When the value of K equals zero, the classification issue is called a zero-shot image classification problem. The fundamental technique employed in few-shot learning is episode training. An episode consists of a support set and a query set. Once the learning process on the support set is over, the model's performance is assessed on the query set. Thus, a few-shot learning assignment might be considered analogous to an episode. A few-shot image classification challenge aims to accurately categorize the photos in the query set using the currently provided supporting set. However, the model must acquire the ability to identify these N categories from the N*K cases.

In NLP, prompt-based language modeling has proven effective for applying pre-trained language models (PLM) to many few-shot situations [54]. The classification issue is approached in prompt-based techniques as a masked language modeling challenge. This approach uses a sequence of prompts to finetune the model and direct its predictions. Strategies that are based on prompts are initially introduced in order to solve text few-shot classification challenges. These strategies include LM-BFF [55], LM-SC [56], and others. Ehsan et al. [57] propose a generative language model to solve the difficulty, which they rephrase as a language generation problem for text categorization. However, the models described above only apply to jobs with text content. The construction of models capable of performing few-shot multimodal tasks has recently received increased interest. Already existing models for few-shot multimodal tasks, such as Frozen [58], PVLM [59], and UP-MPF [60], rely heavily on the addition of picture tokens to a language model that has been trained.

Currently, few-shot learning is heavily used in many image-processing applications, such as image recognition [61], image segmentation [62], image classification [63–65], and retrieval [63–65]. Also, learning how to classify few-shot pictures has a lot of real-world applications. Subpar performance for deep learning models results from the difficulty in collecting large-scale labeled data in public security [66] and health [67]. One possible solution to high-performance models' inability to generalize to new classes due to a lack of training data is limited-shot learning. This would make these models more applicable to other fields. Research on few-shot learning models and algorithms was detailed by Zhao et al. [68]. Methods including transfer learning, data augmentation, and model finetuning were employed. After reviewing the literature on few-shot learning, Wang et al. [69] classified it according to data, model, and approach. On the other hand, few-shot picture classification has received surprisingly little attention in the academic community [70].

A Few-shot Multimodal aspect-based sentiment analysis framework based on Contrastive Finetuning was created by Du et al. [71]. The image modality is first converted to the appropriate textual caption to get the involved semantic information. Next, a contrastive dataset is built using similarity retrieval for finetuning. A sentence encoder is also learned using SBERT, which accomplishes MABSA by fusing sentence-level multi-feature fusion with supervised contrastive learning. A Multi-Aspect Semantic Auxiliary Network (MASANet) is suggested by Cen et al. [72] for visual sentiment analysis. More specifically, MASANet offers cross-domain semantic assistance by achieving modality expansion through cross-modal creation. Next, for aspect-level and cross-modal interaction, respectively, an adaptive modal fusion module and a cross-modal gating module are provided. Furthermore, a specifically created semantic polarity constraint loss is showcased to enhance sentiment multi-classification efficacy.

In the first phases of a few-shot learning research, the model parameters are evaluated using the Bayesian framework, which combines the prior and posterior probabilities to derive the class probability reasoning for the sample [73]. Recent developments in deep learning and neural network architecture have allowed academics to propose neural network models to solve the problem of few-shot photo classification. Most few-shot learning approaches now use the deep learning methodology. Table 1 shows a comparison of papers that have been published before. Medical diagnostics [74], image classification [53, 75], HAR [76–78], and human activity recognition [79–81] are only a few of the many applications of few-shot classification that have been studied before; the applications and corresponding accuracy metrics are included in Table 1.

# 3 Proposed framework

## 3.1 Sentiment analysis using few-shot learning

Low-shot learning, or few-shot learning (FSL), refers to machine learning techniques designed to train systems using only a small number of labeled training samples. The goal is to enable the model to perform tasks effectively despite the limited availability of labeled data. In scenarios with only a few labeled samples, classical machine learning techniques like supervised learning often struggle to generalize. Using classical machine learning techniques like supervised learning (learning from labeled data), scenarios with few training instances usually result in overfitting (i.e., the learner cannot generalize the features of the training data) [82, 83]. Figure 2 shows the general layout of a typical few-shot transfer-learning classification. The dataset was partitioned into source and target domains. The source domain consists of fear, surprise, and anger and

**Table 1** Comparative review of existing works

| Author | Task | FSL technique | Classification accuracy |
|---|---|---|---|
| Mahajan et al. [74] | Skin disease identification | Reptile algorithm and prototypical network | For 2-way, 5-shot classification on the ISIC 2018 Skin Lesion dataset, 79.7% used the prototype network, and 82.1% used the Reptile approach |
| Frikha et al. [75] | Image classification | One-class MAML | 96.6% for the Omniglot dataset with two shots for OCC and 88% for the MNIST dataset |
| Deng et al. [76] | HAR | Weakly-supervised prototypical Networks | The 5-shot classification rates for the PAMAP2, UCIHAR, and Skoda datasets were 65.53%, 91.37%, and 80.17%, respectively |
| Nie et al. [77] | HAR | MAML-CFCNN | 91.17% for 5-way, 1-shot classification, 94.61% with 5-way, 5-shot classification |
| Vinyals et al. [53] | Image classification | Matching Network | One-shot classification had a success rate of 93.2% on the ImageNet dataset and 93.8% on the Omniglot dataset, respectively |
| Feng & Duarte et al. [78] | HAR | Source task to few-shot target task transfer | 69.05% and 68.97% for 5-shot classification on the PAMAP2 and OPP datasets, respectively |

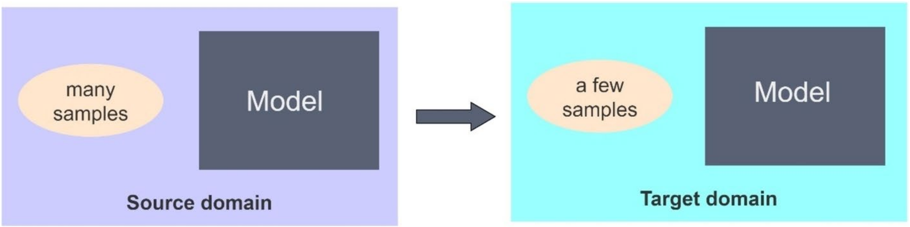


**Fig. 2** Few-shot classification using knowledge transfer

has many samples. The large amount of labeled data in the source domain makes it ideal for training the model. It may then be used to transfer knowledge to the target domain, where the model is finetuned using a limited number of labeled data points of happy, sad, and neutral classes. This classification requires few labeled samples, so the problem is termed a few-shot classification.

Our visual sentiment analysis framework utilizes CNN as its fundamental component. CNNs are extensively used in visual identification, pattern extraction, voice synthesis, and natural language processing (NLP). The proposed CNN model consists of convolution, pooling, and fully connected layers. The input image dimensions are $(n_H, n_W, n_C)$, and the image first passes through a convolution layer. Each convolutional layer is followed by an Exponential Linear Unit (ELU) activation function.

The exponential linear unit (ELU) with $0 < \alpha$ is:

$$f(x) = \begin{cases} x if x > 0 \\ \alpha(\exp(x) - 1) if x > 0 \end{cases} \tag{1}$$

The model in this study is constructed using a Deep convolutional neural network (DCNN). The model's architecture is seen in Fig. 3.

As shown in Fig. 2, the model architecture remains consistent across both the source and target domains. Taking the input image sizes into account, Fig. 4 shows the details of each model layer. Layers and their relative placements in the DCNN architecture, output format, parameters or weights for each layer, and the total number of parameters in the model are all part of the parameter configuration specification, which is based on the transfer learning finetuning parameter.

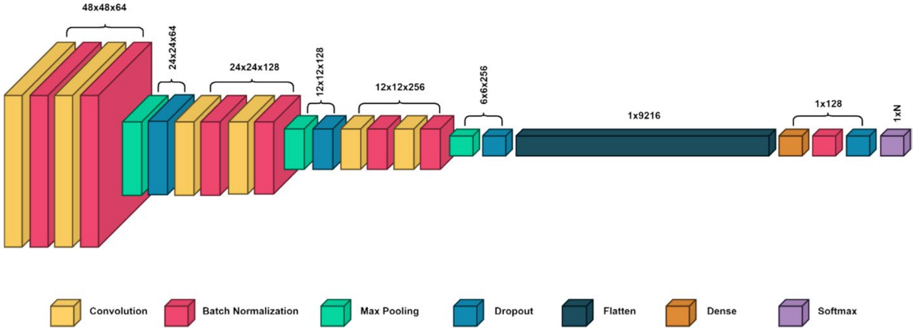
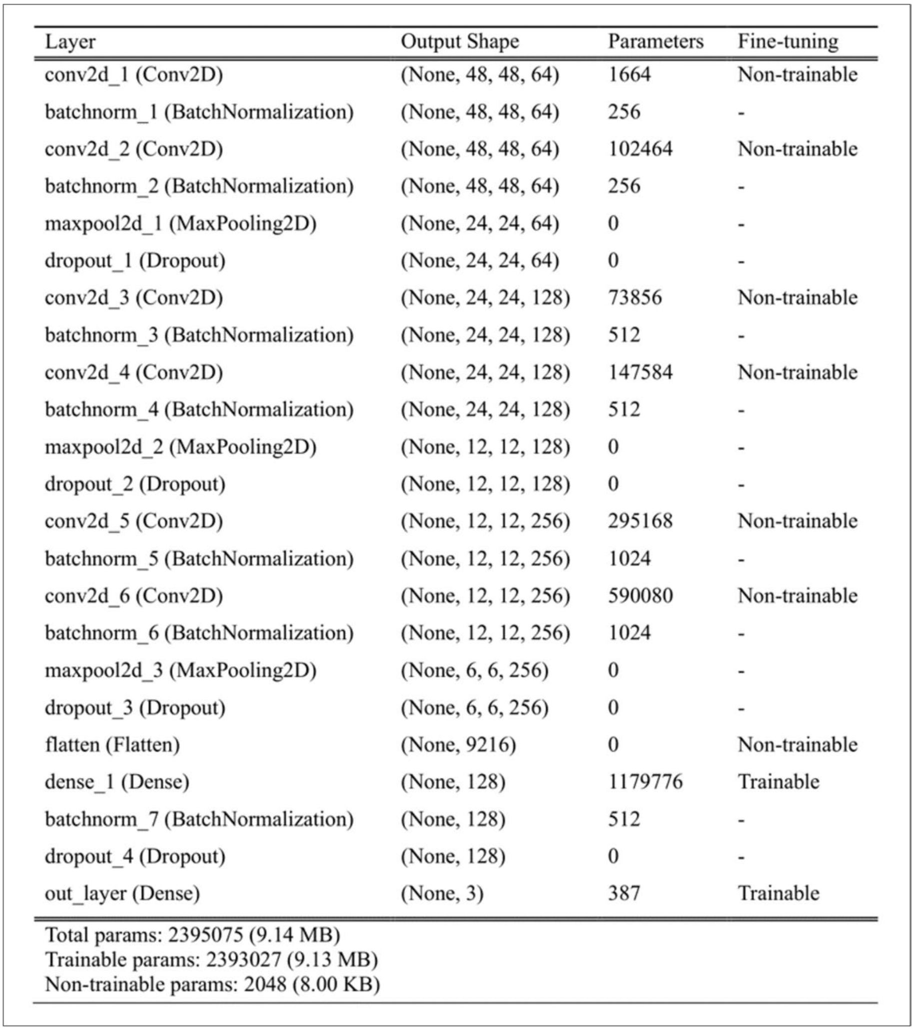


**Fig. 3** Model Architecture of DCNN



| Layer | Output Shape | Parameters | Fine-tuning |
|---|---|---|---|
| conv2d_1 (Conv2D) | (None, 48, 48, 64) | 1664 | Non-trainable |
| batchnorm_1 (BatchNormalization) | (None, 48, 48, 64) | 256 | - |
| conv2d_2 (Conv2D) | (None, 48, 48, 64) | 102464 | Non-trainable |
| batchnorm_2 (BatchNormalization) | (None, 48, 48, 64) | 256 | - |
| maxpool2d_1 (MaxPooling2D) | (None, 24, 24, 64) | 0 | - |
| dropout_1 (Dropout) | (None, 24, 24, 64) | 0 | - |
| conv2d_3 (Conv2D) | (None, 24, 24, 128) | 73856 | Non-trainable |
| batchnorm_3 (BatchNormalization) | (None, 24, 24, 128) | 512 | - |
| conv2d_4 (Conv2D) | (None, 24, 24, 128) | 147584 | Non-trainable |
| batchnorm_4 (BatchNormalization) | (None, 24, 24, 128) | 512 | - |
| maxpool2d_2 (MaxPooling2D) | (None, 12, 12, 128) | 0 | - |
| dropout_2 (Dropout) | (None, 12, 12, 128) | 0 | - |
| conv2d_5 (Conv2D) | (None, 12, 12, 256) | 295168 | Non-trainable |
| batchnorm_5 (BatchNormalization) | (None, 12, 12, 256) | 1024 | - |
| conv2d_6 (Conv2D) | (None, 12, 12, 256) | 590080 | Non-trainable |
| batchnorm_6 (BatchNormalization) | (None, 12, 12, 256) | 1024 | - |
| maxpool2d_3 (MaxPooling2D) | (None, 6, 6, 256) | 0 | - |
| dropout_3 (Dropout) | (None, 6, 6, 256) | 0 | - |
| flatten (Flatten) | (None, 9216) | 0 | Non-trainable |
| dense_1 (Dense) | (None, 128) | 1179776 | Trainable |
| batchnorm_7 (BatchNormalization) | (None, 128) | 512 | - |
| dropout_4 (Dropout) | (None, 128) | 0 | - |
| out_layer (Dense) | (None, 3) | 387 | Trainable |

Total params: 2395075 (9.14 MB)
Trainable params: 2393027 (9.13 MB)
Non-trainable params: 2048 (8.00 KB)

**Fig. 4** The details of each layer in the proposed model

The proposed CNN consists of three pooling and six convolution layers, as shown in Fig. 4. The first two convolution layers consist of 64 filters with the same padding, followed by max-polling to reduce the spatial dimension without sacrificing using information, followed by 128 and 256 filters. For classification, the Softmax activation was employed in the last dense layer with the output of size N, where N is the number of categories.

#### 3.1.1 Training, finetuning, and testing

In the source domain, the training step is carried out with a batch size of 32, using the Adam optimizer with a learning rate of 0.001 and categorical cross-entropy as the loss function. A validation set consisting of 10% of the data from the source domain is used to monitor the model's training. Once the training is complete, the trained model in the source domain is transferred to the target domain, retaining the same number of neurons in the final dense layer since the number of classes remains unchanged between domains in this study. During the finetuning process, there is a limited quantity of labeled data accessible in the target domain. No finetuning of the model parameters can escape the significant overfitting problem caused by the scarcity of labeled training data.

Figure 4 shows that the last two layers of the model can be finetuned. The finetuned model is then tested in the target domain. A few-shot classification problem is defined as an N-way k-shot problem. N represents the number of classes in the target domain, and k represents the number of samples used to finetune the transferred model. The task's difficulty level might vary depending on how similar or distinct the groups that need to be categorized.

### 3.1.2 Single semi-supervised few-shot classification

The typical approach for finetuning and testing is depicted in Fig. 5, which contrasts with the previous illustrations. The N*k samples were used to finetune the dense layers using N-way-k-shot samples and update the linear layers. The finetuned model is then fixed for testing. To assess the performance of the finetuned model from a few shots, we randomly selected 15 samples from each category, resulting in a total of N*15 samples for testing. The model was tested on a different test set each time, and the testing was conducted in the proposed framework to showcase the model's generalization.

In order to enhance the few-shot classification performance, we suggest utilizing the single semi-supervised approach depicted in Fig. 6. It is demonstrated that there are two stages to finish.

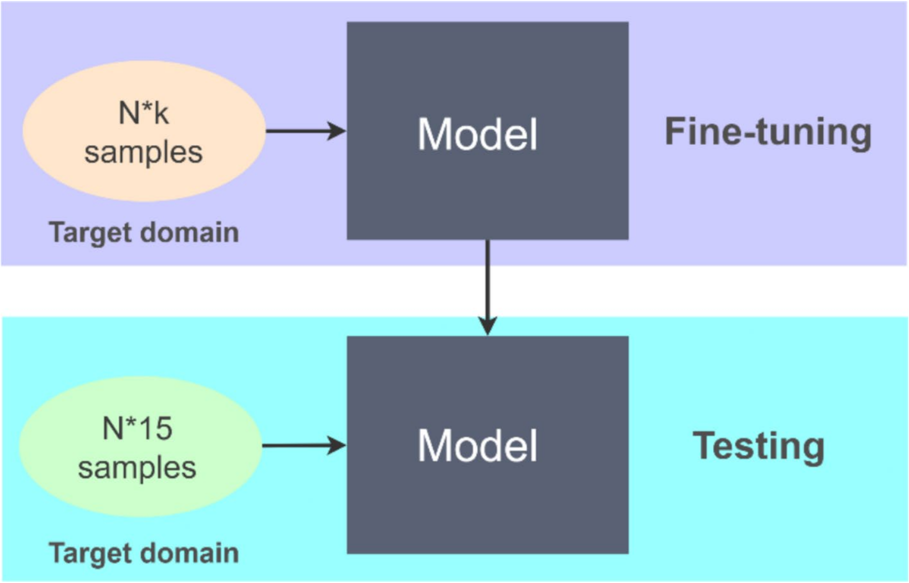


**Fig. 5** The finetuning and testing process

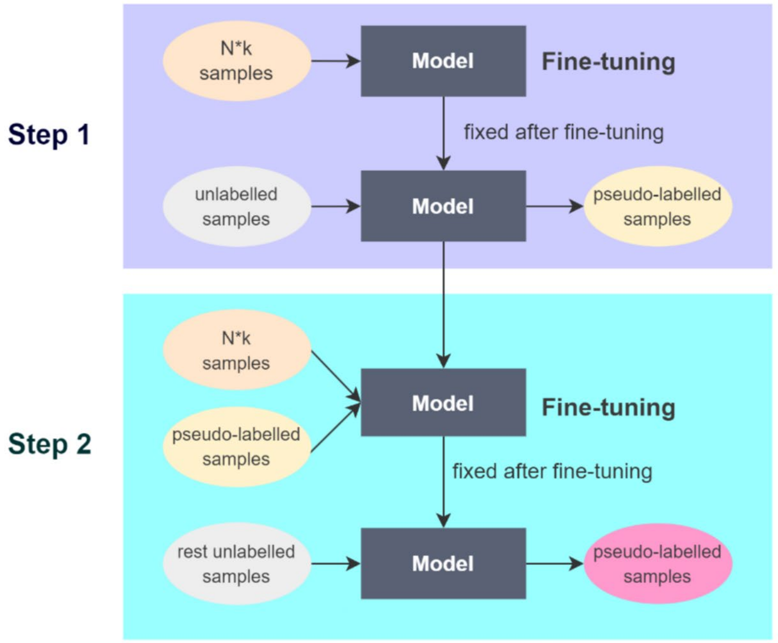


**Fig. 6** Single semi-supervised few-shot method

Step 1 involves finetuning the transferred model using the N*k samples with true labels while freezing the parameters. The fixed model is then fed all the unlabeled data to generate predictions and select pseudo-labeled samples with a prediction threshold of 99% accuracy. Given the model's prediction confidence, the prediction label assigned to a sample based on the model's assessment should closely resemble the true label.

Step 2 entails refining the model using the selected pseudo-labeled data and the N*k labeled samples from Step 1, with only the parameters of the two dense layers remaining trainable. Once the model has been finetuned, the parameters are again frozen and evaluated using the N*15 samples. This procedure is also known as single semi-supervised few-shot classification, as the pseudo-labeled samples in the semi-supervised technique are chosen only once.

### 3.1.3 Iterative semi-supervised few-shot classification

We also suggest an iterative semi-supervised few-shot classification, represented in Fig. 7, based on the single semi-supervised few-shot classification.

The main distinction between this iterative approach, which consists of three phases, and the single semi-supervised technique is the selection of pseudo-labeled samples twice.

The source domain model is finetuned and fixed in Step 1 using the N*k samples with true labels. The remaining unlabeled samples (apart from those chosen in Step 1) are loaded into the fixed model in Step 2 to pick further pseudo-labeled samples

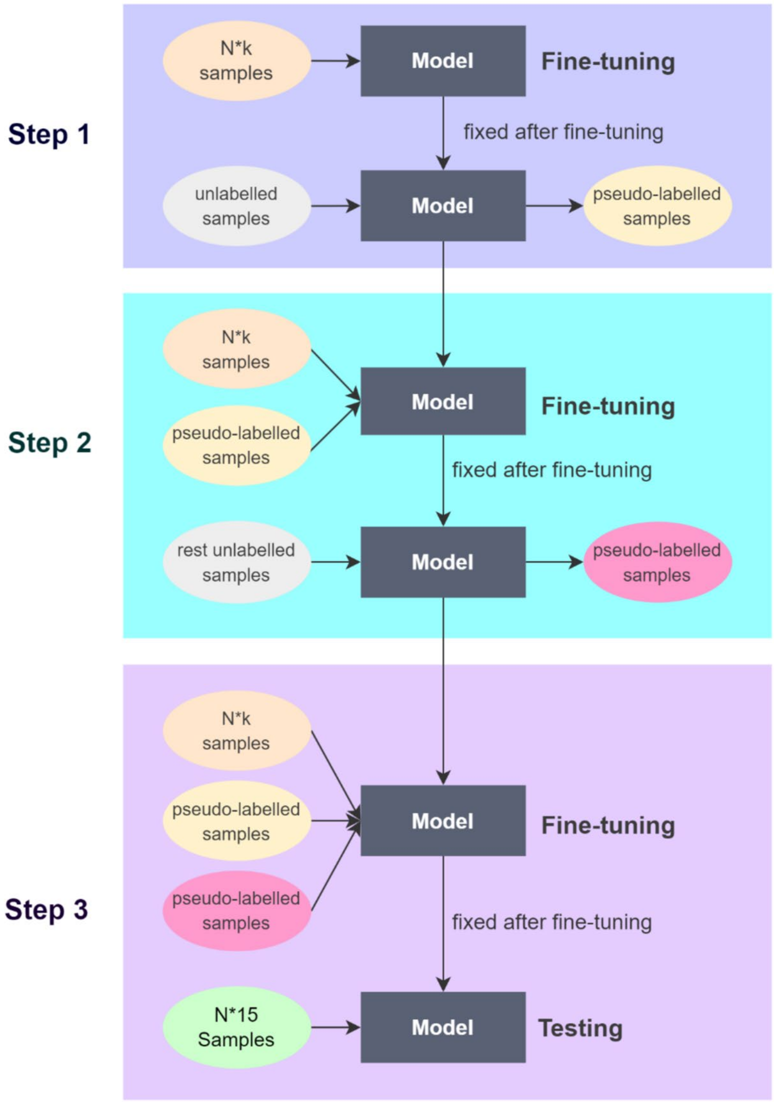


**Fig. 7** The iterative semi-supervised few-shot method

following the first finetuning with genuine labeled data and pseudo-labeled data from Step 1. In Step 3, the remaining two dense layers of the model are finetuned using the pseudo-labeled samples from Steps 1 and 2 and the N*k labeled samples. The model is then evaluated on the N*15 random samples.

### 3.1.4 Adaptive selection of pseudo-labeled samples

Figures 6 and 7 illustrate the process of selecting pseudo-labeled data by feeding the unlabelled data into the trained model. Selecting a fixed number, such as 5 or 10, is the simplest option for semi-supervised learning. Nevertheless, this choice is not intelligent due to the lack of suitability for several tasks. A word of caution: selecting pseudo-labeled samples is not without its limitations. Collecting several suitable pseudo-labeled samples using the semi-supervised method can compensate for the scarcity of data with original labels. On the other hand, if we get a lot of bad pseudo-labeled samples, the few-shot performance will take a nosedive. In addition, picking a limited number of highly cautious pseudo-labeled samples will result in a minor advantage.

To address the above issue, we propose a selection method based on the number of samples per class from the pseudo-labeled data, utilizing a threshold of 99% accuracy, the k-shot, and a multiplier. Despite being pseudo-labels, these given labels should align closely with the actual labels, given the model's high confidence in its predictions. In this scenario, the multiplier was set to 0.25 for the single semi-supervised model and 0.30 for the iterative semi-supervised model. The equation for selecting pseudo-labeled samples is provided in Eq. 2.

Let $n_{c1}, n_{c2}, andn_{c3}$ be the number of pseudo-labelled samples selected for the classes $c1, c2, andc3$ respectively, with a threshold of 99%.

$$psuedosamples(perclass) = \min\left(\min\left(n_{c1}, n_{c2}, n_{c3}\right), \lceil k * multiplier \rceil\right) \tag{2}$$

where ⌈x⌉ denotes the ceil function of x, after extensive experimentation on parameter tuning, Eq. 2 selects the suitable number of pseudo-labeled samples.

## 3.2 FSTL (Few Shot Transfer Learning) framework

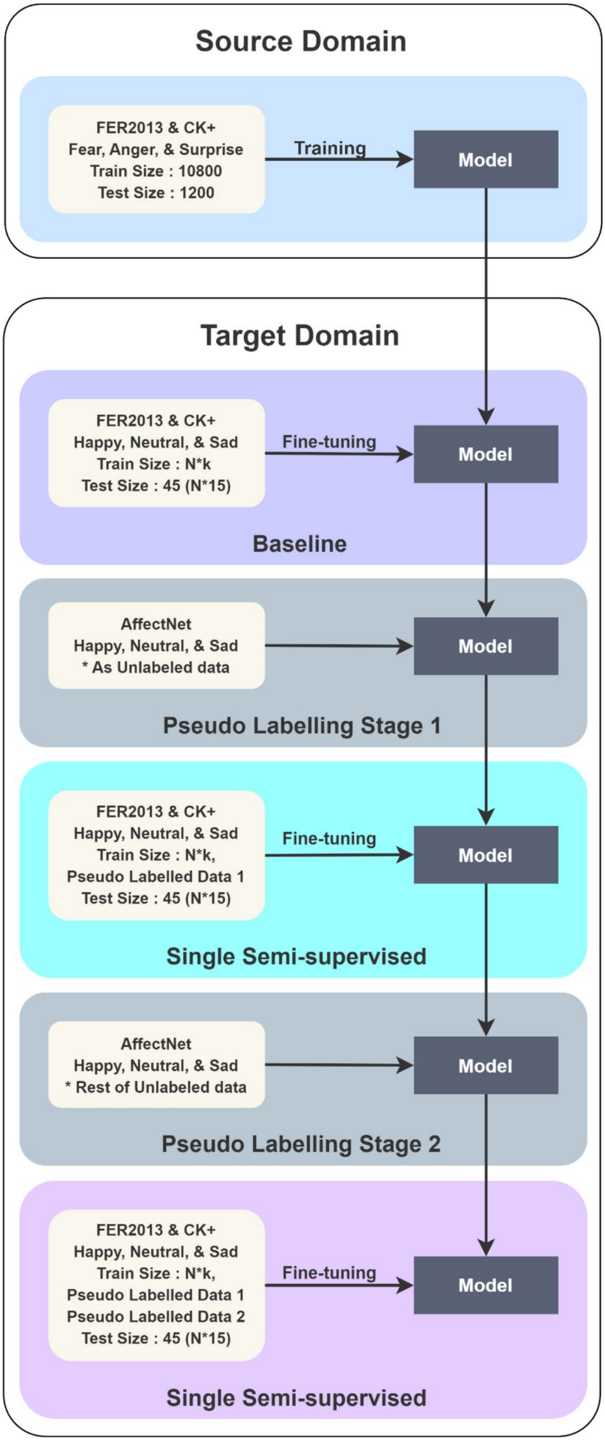
We decided to apply a standardized framework for sentiment analysis from nonverbal communication called FSTL. The framework comprises data preprocessing, model training, baseline (supervised), pseudo labeling, single semi-supervised few-shot learning, iterative semi-supervised few-shot learning, and pseudo labeling. Figure 8 shows the process in its entirety.

# 4 Experiments

## 4.1 Data source and description

Comprehensive data for our facial sentiment classification model is available in the FER2013 dataset, which was used for the Kaggle Competition. This database was created during the ICML 2013 Kaggle competitions [84]. Since then, scientific studies on facial sentiment detection have been assessed using the data gathered. The images are automatically registered to center each image's face and achieve roughly equal sizes. Based on the emotions conveyed by the facial expression, all faces are to be categorized into seven facial expressions. A total of 35887 grayscale images with a 48 × 48 pixel resolution are included in this dataset: 28709 for training and 3589 for testing. An illustration of the database in

**Fig. 8** Framework of the Proposed FSTL-SA



particular classes is shown in Fig. 1. This dataset's quantity of samples for each categorization group is wildly imbalanced. For instance, the class "disgust" makes up about 1% of the data set, but the class "happy" makes up nearly 25%.

The second database is CK + [85], a widely used resource for researchers that includes seven emotions from 123 different individuals. Additionally, 961 images were collected, with each participant representing one of the seven basic emotional categories based on their appearance. For pseudo-labeling, we utilized the AffectNet dataset developed by Mollahosseini et al. [86], an extensive in-the-wild dataset from the internet. It contains over 1 million facial images covering many emotions, including primary and compound expressions, all meticulously annotated with emotion labels and facial attributes. The diversity and detail of this dataset make it particularly suitable for real-world applications, as the images reflect genuine expressions captured in various contexts.

### 4.2 Data preparation

To perform the few-shot classifications, the FER2013 and CK + were combined to create a larger dataset that contains diverse facial expressions. The seven facial expressions are split into source and target domains. Due to the smaller number of samples in disgust, disgust was not considered in this experiment. Fear, surprise, and anger were used to train the source domain. Downsampling was used, and 4000 images were selected for each class. Among the 12000 images, 10%, 1200, was used for validation; the rest, 10800, was used for training. All the images are converted to grey-scale images, resized to 48 × 48 pixels, and normalized by 255. Data Augmentation increases the diversity of the training data without actually collecting more images. This helps the model learn better by seeing different variations of the same images. For the training process, the images were randomly rotated between + / − 15 degrees from the center of the image. The width and height of the images are shifted by 15%. The images were sheared and zoomed through 15%. Horizontal flipping was also employed randomly on the images. The same processing techniques were applied to the target domain to ensure consistency across both datasets, generalize the model for variations, and reduce bias.

### 4.3 Experimental setup

The experiments in this study were conducted using Python (version 3.10.7) on a Windows 11 operating system. The deep learning models were built with the Keras framework (version 2.10.0), utilizing the backend of TensorFlow (version 2.10.0). The high-performance computing environment included an Intel(R) Xenon(R) W-2255 CPU @ 3.70 GHz, 64 GB of RAM, an NVIDIA GEFORCE RTX A4000 GPU with 16 GB of memory, and DirectX-12 (version 12.1).

**Table 2** Confusion matrix

| Predicted Value | Actual Value | |
|---|---|---|
| | Positive | Negative |
| Positive | True Positive | False Positive |
| Negative | False Negative | True Negative |

**Table 3** Performance evaluation measures

| Metric Name | Definition |
|---|---|
| Accuracy | $\frac{TP+TN}{TP+TN+FP+FN}$ |
| Precision | $\frac{TP}{TP+FP}$ |
| Recall | $\frac{TP}{TP+FN}$ |
| F-Measure | $\frac{2*Precision*Recall}{Precision+Recall}$ |

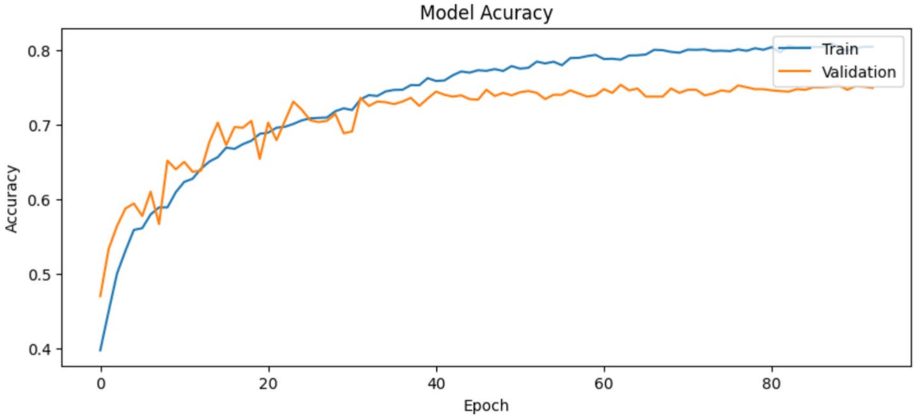


**Fig.9** Training and validation accuracy of the source domain

## 4.4 Performance evaluation

Primary key performance indicators are used to evaluate the efficacy of the classifier. Some measurements presented here include precision, recall, accuracy, and f-measure. Table 2 presents the parameter values for the confusion matrix measure [87]. These estimates will be considered during the upcoming performance review. Table 3 provides definitions for a performance measure used in this work [88].

## 4.5 Results

### 4.5.1 Source domain

For this experiment, we have chosen three facial emotions, fear, anger, and disgust, for the source domain. The number of samples from each of the above classes is over 4000. We have down-sampled the number of samples per class to 4000. The images are converted to grayscale images and normalized.

The images were processed and split with a test size of 10%, which is 1200 samples for the testing and 10800 for training the source domain. For training the model, we have employed model checkpoints, a learning rate scheduler with a factor of 0.5, patience of 7, and a minimum learning rate of $10^{-7}$, and early stopping with the patience of 30, the minimum delta of $10^{-5}$, all three to monitor the validation accuracy. The maximum number of

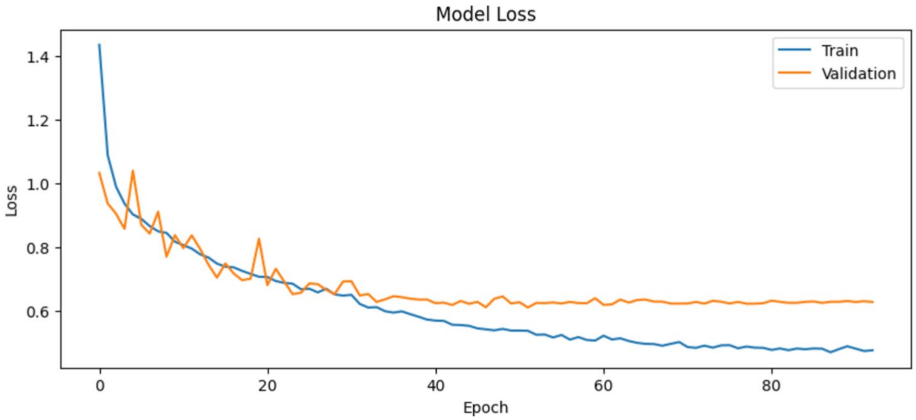


**Fig. 10** Training and validation loss of the source domain

**Table 4** Performance of DCNN model on Source Domain

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|---|
| DCNN | 75.33 | 75.33 | 75.33 | 75.33 |

epochs was set to 100 with a batch size of 32. The model stopped training at 93 epochs. The source domain achieved an accuracy of 75.33%, a loss of 0.6348. The accuracy and loss of the training and validation set are shown in Figs. 9 and 10, respectively. The accuracy, precision, recall, and F1-Score performance of the DCNN model on the source domain are defined in Table 4. To assist the reader in gaining a deeper comprehension of the classification abilities of the model, Fig. 11 presents the confusion matrix of the DCNN model on the test set.

### 4.5.2 Target domain

For the target domain, we have employed the N-way-k-shot few-shot learning. N represents the number of classes, and k is the number of samples per class used to finetune the model from the source domain. The N was set to 3, with three facial emotions: happy, sad, and neutral. For the k, we have used 1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100. The number of happy, sad, and neutral samples was 5044, 3091, and 5126, respectively, of CK + and FER2013. For pseudo-labeling, we have used AffectNet as unlabelled data in this experiment. The trained model from the source domain was used as the base learner in the target domain with all the layers till the flattened layer remained frozen, and only the fully connected layers were finetuned; for the finetuning process, checkpoint and learning rate scheduler with the patience of 5 was employed, monitoring the validation accuracy Tables 5 and 6. The models were finetuned with 20 epochs, and the batch size was set to $k$. For testing and validation, $N * 15$ samples were used. No single test and validation samples were used twice for testing and validation. To check the correctness of the pseudo-labeled data, we have calculated the accuracy of the pseudo-labeled data; the results are shown in Table 7. To check the

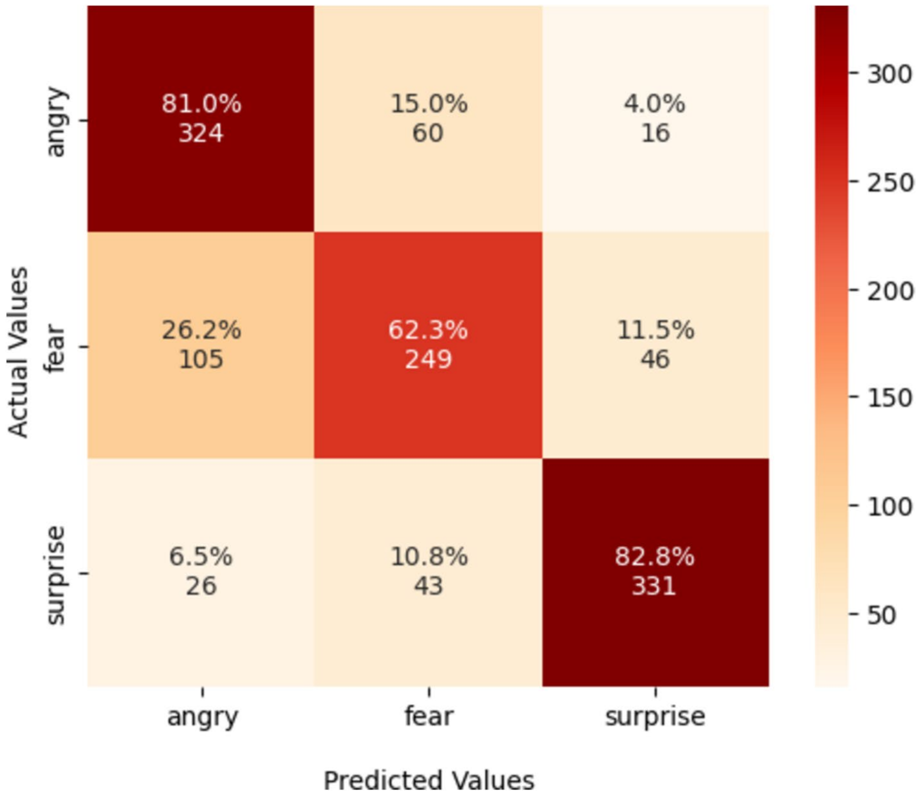


**Fig. 11** Confusion matrix of the source domain

performance, the few-shot model, accuracy, loss, precision, recall, and f1-score were employed for each emotion individually, and the three classes combined collectively. The results have shown improvement in training and testing accuracy on the baseline, single SS, and iterative SS method on different k-shots, as shown in Table 5.

In Fig. 5, we can observe the baseline, which stands for the standard few-shot transfer-learning categorization. Figure 6 shows the single SS strategy, while Fig. 7 shows the iterative SS method. These are the two suggested ways. Figures 12, 13, 14, and 15 depict the relationship between average accuracy and loss during training and testing using k-shot. The few-shot parameters are shown below. The N-way is 3; the k-shots are 1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100.

Two points can be intuitively represented in Fig. 13. Specifically, the iterative SS technique achieves better results than the alternatives, although it demands more operations. Therefore, the single SS technique might be a good option to balance performance gains with computational complexity.

The average improvement in accuracy for the different expressions of Happy, Neutral, and Sad by FSTL-based baseline, single SS method, and iterative SS method on different k-shots can be calculated from Table 6.

**Table 5** Comparison of Baseline, Single SS, and Iterative SS

(a): Accuracy of Baseline, Single SS, and Iterative SS on training data

| **Model** | **1** | **5** | **10** | **20** | **30** | **40** | **50** | **60** | **70** | **80** | **90** | **100** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | 0.00 | 93.3 | 96.6 | 88.3 | 96.6 | 94.1 | 93.3 | 86.6 | 82.3 | 85.0 | 85.5 | 78.0 |
| Single SS | 33.3 | 100 | 94.8 | 89.3 | 92.9 | 90.6 | 85.1 | 89.7 | 81.8 | 83.3 | 79.6 | 76.0 |
| Iterative SS | 33.3 | 100 | 87.5 | 92.4 | 86.5 | 89.2 | 82.4 | 84.2 | 82.2 | 82.7 | 79.2 | 81.5 |

(b): Loss of Baseline, Single SS, and Iterative SS on training data

| **Model** | **1** | **5** | **10** | **20** | **30** | **40** | **50** | **60** | **70** | **80** | **90** | **100** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | 2.86 | 0.07 | 0.04 | 0.34 | 0.08 | 0.13 | 0.18 | 0.44 | 0.55 | 0.43 | 0.37 | 0.62 |
| Single SS | 1.79 | 0.00 | 0.28 | 0.20 | 0.21 | 0.28 | 0.32 | 0.32 | 0.47 | 0.45 | 0.55 | 0.59 |
| Iterative SS | 3.51 | 0.02 | 0.29 | 0.23 | 0.31 | 0.29 | 0.50 | 0.37 | 0.50 | 0.43 | 0.50 | 0.44 |

(c): Accuracy of Baseline, Single SS, and Iterative SS on testing data

| **Model** | **1** | **5** | **10** | **20** | **30** | **40** | **50** | **60** | **70** | **80** | **90** | **100** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | 26.6 | 60.0 | 51.1 | 55.5 | 62.2 | 60.0 | 57.7 | 57.7 | 57.7 | 62.2 | 62.2 | 57.7 |
| Single SS | 31.1 | 46.6 | 48.8 | 55.5 | 55.5 | 62.2 | 53.3 | 82.2 | 71.1 | 64.4 | 60.0 | 60.0 |
| Iterative SS | 40.0 | 57.7 | 35.5 | 53.3 | 60.0 | 57.7 | 53.3 | 60.0 | 48.8 | 64.4 | 57.7 | 66.6 |

(d): Loss of Baseline, Single SS, and Iterative SS on testing data

| **Model** | **1** | **5** | **10** | **20** | **30** | **40** | **50** | **60** | **70** | **80** | **90** | **100** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | 3.39 | 1.99 | 1.72 | 1.34 | 1.90 | 1.60 | 1.54 | 1.59 | 2.03 | 1.37 | 1.26 | 1.37 |
| Single SS | 2.78 | 2.49 | 2.87 | 2.17 | 2.19 | 1.65 | 1.99 | 0.74 | 1.17 | 1.23 | 1.46 | 1.49 |
| Iterative SS | 4.08 | 1.35 | 3.19 | 1.83 | 1.27 | 1.08 | 2.29 | 1.25 | 1.69 | 1.10 | 1.44 | 1.10 |

**Table 6** Accuracy of different classes using FSTL

(a): Accuracy of Happy expression using FSTL

| **Model** | **1** | **5** | **10** | **20** | **30** | **40** | **50** | **60** | **70** | **80** | **90** | **100** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | 40.0 | 53.3 | 40.0 | 66.6 | 80.0 | 73.3 | 86.6 | 66.6 | 53.3 | 73.3 | 80.0 | 53.3 |
| Single SS | 26.6 | 60.0 | 80.0 | 66.6 | 66.6 | 73.3 | 53.3 | 86.6 | 80.0 | 66.6 | 80.0 | 66.6 |
| Iterative SS | 20.0 | 60.0 | 33.3 | 66.6 | 60.0 | 73.3 | 60.0 | 73.3 | 60.0 | 66.6 | 80.0 | 93.3 |

(b): Accuracy of Neutral expression using FSTL

| **Model** | **1** | **5** | **10** | **20** | **30** | **40** | **50** | **60** | **70** | **80** | **90** | **100** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | 20.0 | 60.0 | 73.3 | 46.6 | 40.0 | 33.3 | 46.6 | 53.3 | 40.0 | 46.6 | 40.0 | 46.6 |
| Single SS | 33.3 | 53.3 | 20.0 | 26.6 | 53.3 | 60.0 | 26.6 | 93.3 | 46.6 | 46.6 | 20.0 | 46.6 |
| Iterative SS | 46.6 | 73.3 | 6.66 | 40.0 | 46.6 | 46.6 | 13.3 | 33.3 | 20.0 | 33.3 | 33.3 | 33.3 |

(c): Accuracy of sad expression using FSTL

| **Model** | **1** | **5** | **10** | **20** | **30** | **40** | **50** | **60** | **70** | **80** | **90** | **100** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | 20.0 | 66.6 | 40.0 | 53.3 | 66.6 | 73.3 | 40.0 | 53.3 | 80.0 | 66.6 | 66.6 | 73.3 |
| Single SS | 33.3 | 26.6 | 46.6 | 73.3 | 46.6 | 53.3 | 80.0 | 66.6 | 86.6 | 80.0 | 80.0 | 66.6 |
| Iterative SS | 53.3 | 40.0 | 66.6 | 53.3 | 73.3 | 53.3 | 86.6 | 73.3 | 66.6 | 93.3 | 60.0 | 73.3 |

**Table 7** Accuracy of the Selected Pseudo-labelled data

| Model | 1 | 5 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Single SS | 33.3 | 33.3 | 33.3 | 53.3 | 45.8 | 66.6 | 48.7 | 60.0 | 62.9 | 56.6 | 60.8 | 53.3 |
| Iterative SS | 33.3 | 33.3 | 66.6 | 61.1 | 55.5 | 41.6 | 80.0 | 57.4 | 71.42 | 68.0 | 79.0 | 76.6 |

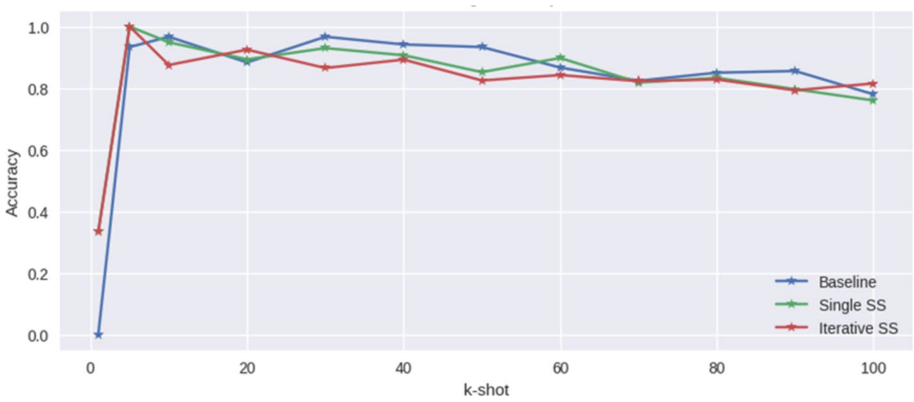


**Fig.12** Accuracy of Baseline, Single SS, and Iterative SS on Train Data

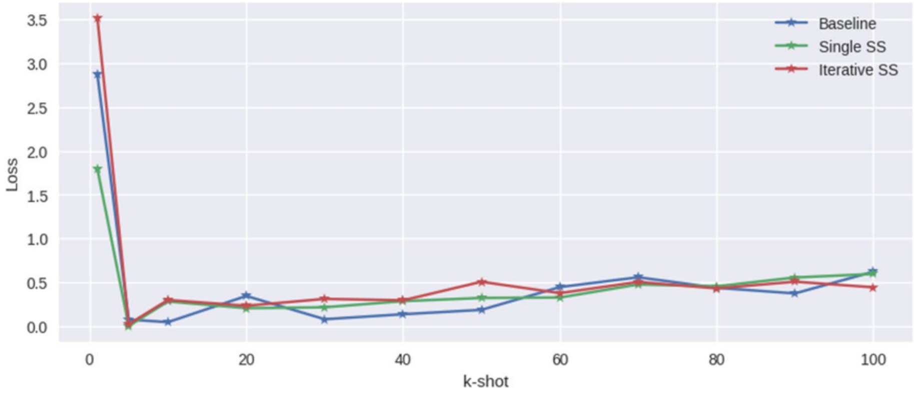


**Fig.13** Loss of Baseline, Single SS, and Iterative SS on Train Data

## 4.6 Results of adaptive selection of pseudo-labeled samples

This part will discuss how the proposed semi-supervised few-shot methods rely on selecting pseudo-labeled samples. The number of adaptively selected pseudo-labeled samples under different k-shots is shown in Fig. 16. A significant correlation was found between the number of k-shots and the selected pseudo-labeled samples. With increased k-shot,

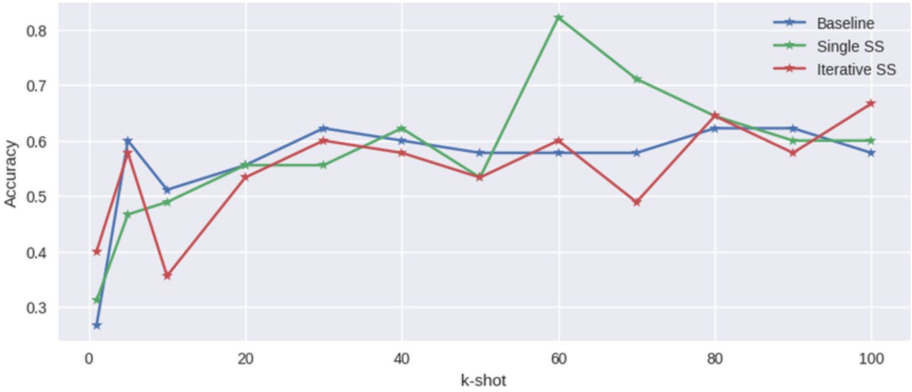


**Fig.14** Accuracy of Baseline, Single SS, and Iterative SS on test data

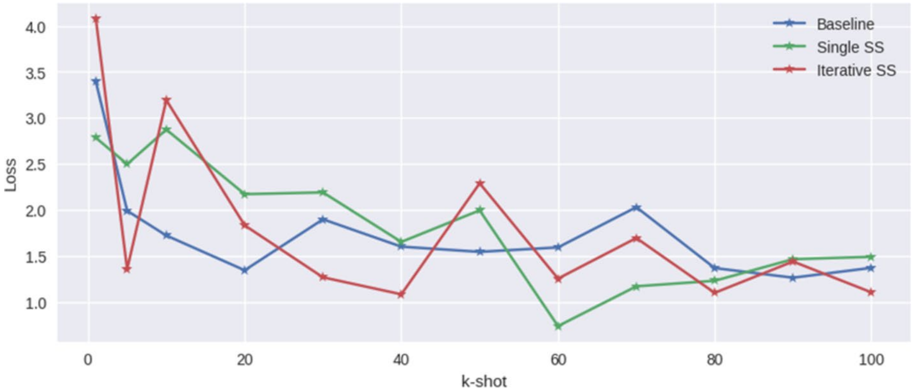


**Fig.15** Loss of Baseline, Single SS, and Iterative SS on test data

the model gets access to more training data for finetuning. Put another way, the model is stronger. Therefore, such unlabeled data may provide a more secure basis for prediction. A pseudo-labeled sample is selected if one has more than 99% predicted confidence.

Compared to the single SS strategy, more pseudo-labeled samples were selected using the iterative SS method. This is because, as opposed to the single SS technique, the iterative SS methodology includes a single additional stage for finetuning. Consequently, iterative SS provides a better grasp of the model's performance. More accurate predictions of unlabeled data were made using the analyzed categories.

## 4.7 Discussion

Automatic visual sentiment analysis categorization based on a small number of labeled samples is crucial for ensuring minimal data costs while maintaining high yield and quality. This study introduced a semi-supervised few-shot learning technique that adaptively selects pseudo-labeled samples to aid in model tuning, thereby increasing the average

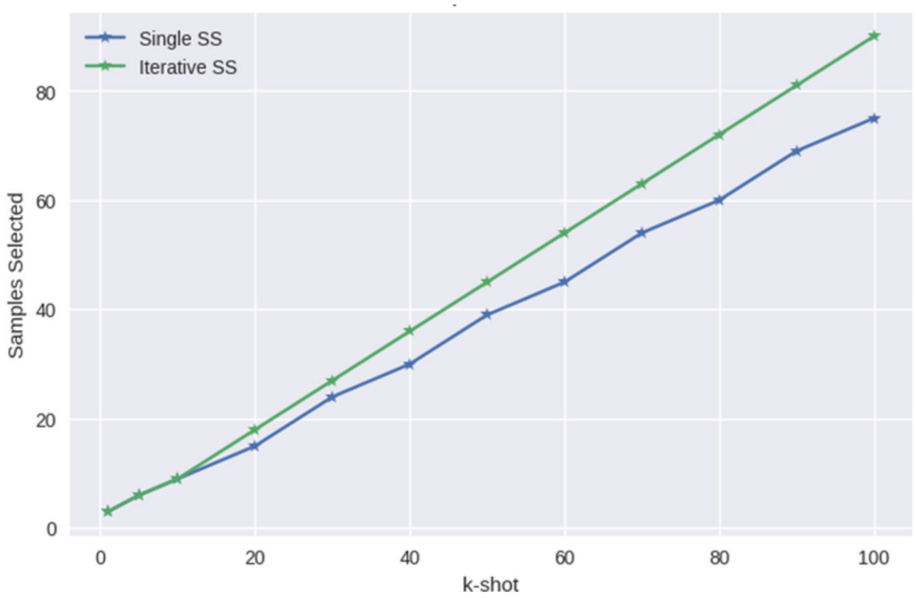


**Fig.16** The number of pseudo-labeled samples

accuracy of few-shot classification. To our knowledge, we conducted the first semi-supervised study in visual sentiment analysis through an extensive literature review. To demonstrate the accuracy and generalizability of the proposed approaches, several comparative experiments were carried out using FER2013, CK+, and AffectNet datasets.

The proposed framework is suitable for domains where labeled data is scarce but unlabeled data, such as facial images, is readily available online. The unlabeled data can be pseudo-labeled with a high prediction threshold, making the pseudo-labels nearly identical to the true labels. By utilizing an iterative semi-supervised method, the processes of pseudo-labeling and finetuning can be repeated indefinitely, leading to high-quality pseudo-labeled data that helps mitigate the scarcity of labeled data. However, the framework does not address specific issues, such as mislabeled data, which fall under the broader research area of robustness. If incorrect labels are assigned to the labeled data, it is advisable to clean them first, as this can mislead the learning process. Conversely, suppose incorrect labels correspond to the pseudo-labeled data. In that case, it is recommended to adjust the confidence interval to tighten the screening criteria and increase the number of iterations to enhance the model's filtering performance.

This study has primarily been evaluated within visual sentiment analysis and may require adaptation for effective performance in other few-shot learning tasks. Additionally, the approach may not be well-suited for domains where unlabeled data is scarce, in contrast to visual sentiment analysis, which typically benefits from abundant such data. The framework is also adaptable for in-the-wild unlabeled data, as AffectNet, used in this study, is an in-the-wild dataset with images from the internet.

## 5 Conclusion

Deep learning is currently the primary foundation of visual sentiment analysis. Despite its many successes, the disadvantages of deep learning, such as the significant expense of gathering and labeling large-scale datasets, cannot be disregarded. Few-shot learning is a crucial addition to deep learning, integrating a small number of samples and expertise focusing on model learning. The supervised paradigm has been the main emphasis of the few-shot experiments conducted in visual sentiment analysis. Therefore, we have aimed to investigate the semi-supervised paradigm to enhance the impact of few-shot classification and motivate this community. To address few-shot visual sentiments for images, we introduced FSTL-SA, a single semi-supervised and iterative semi-supervised technique. Overall, our proposed strategy demonstrates superior performance across k-shot scenarios. The DCNN achieved an accuracy of 75.33% in the source domain. The FSTL-SA method attained an average accuracy of 61% for 100-shot scenarios and 82% for the single semi-supervised approach at 60 shots. These results indicate that our methods can deliver better outcomes with fewer samples under similar conditions.

In the future, we plan to conduct few-shot classification under significant cross-domain conditions, utilizing an in-the-wild dataset as the source domain and laboratory images as the target domain to adopt a broader and more practical perspective. Additionally, since few-shot learning aims to learn from limited samples and facilitate convenient application deployment, we intend to reduce the model size to create a smaller, more efficient, and more intelligent model for easier deployment.

**Funding** NA.

**Data availability** The corresponding author can provide the datasets created and/or analyzed during the current work upon reasonable request.

## Declarations

**Ethical approval** None of the authors' investigations involving humans or animals are included in this article.

**Conflict of interest** The authors claim they have no competing interests.

## References

1. Tambo E, Al-Nazawi AM (2022) Combating the global spread of poverty-related Monkeypox outbreaks and beyond. Infect Dis Poverty 11(1):80
2. McDuff D, El Kaliouby R, Cohn JF, Picard RW (2014) Predicting ad liking and purchase intent: large-scale analysis of facial responses to ads. IEEE Trans Affect Comput 6(3):223–235
3. Ain QT, Ali M, Riaz A, Noureen A, Kamran M, Hayat B, Rehman A (2017) Sentiment analysis using deep learning techniques: a review. Int J Adv Comp Sci App 8(6)
4. Anand K, Urolagin S, Mishra RK (2021) How does hand gestures in videos impact social media engagement-Insights based on deep learning. Int J Inform Manage Data Insights 1(2):100036
5. Sahu M, Dash R (2021) A survey on deep learning: convolution neural network (CNN). In intelligent and cloud computing: proceedings of ICICC 2019, Volume 2 (pp. 317–325). Springer Singapore
6. Fe-Fei L (2003) A Bayesian approach to unsupervised one-shot learning of object categories. In proceedings ninth IEEE international conference on computer vision (pp. 1134-1141). IEEE, Nice, France, 13-16 October 2003. https://doi.org/10.1109/ICCV.2003.1238476

7. Kumari A, Tanwar S, Tyagi S, Kumar N (2018) Fog computing for Healthcare 4.0 environment: opportunities and challenges. Comput Electr Eng 72:1–13
8. Hathaliya J, Sharma P, Tanwar S, Gupta R (2019, December) Blockchain-based remote patient monitoring in healthcare 4.0. In: 2019 IEEE 9th international conference on advanced computing (IACC). IEEE, pp 87–91. https://doi.org/10.1109/IACC48062.2019.8971593
9. Vora J, DevMurari P, Tanwar S, Tyagi S, Kumar N, Obaidat MS (2018) Blind signatures based secured e-healthcare system. In 2018 International conference on computer, information and telecommunication systems (CITS) (pp. 1-5). IEEE, Alsace, Colmar, France, 11-13 July 2018. https://doi.org/10.1109/CITS.2018.8440186
10. Zhang L, Verma B, Tjondronegoro D, Chandran V (2018) Facial expression analysis under partial occlusion: a survey. ACM Comp Surveys (CSUR) 51(2):1–49
11. Do HH, Prasad PW, Maag A, Alsadoon A (2019) Deep learning for aspect-based sentiment analysis: a comparative review. Expert Syst Appl 118:272–299
12. Zhang L, Wang S, Liu B (2018) Deep learning for sentiment analysis: a survey. Wiley Interdisciplinary Rev: Data Mining Knowledge Disc 8(4):e1253
13. Huang Y, Xu H (2021) Fully convolutional network with attention modules for semantic segmentation. SIViP 15:1031–1039
14. You Q, Luo J, Jin H, Yang J (2015) Robust image sentiment analysis using progressively trained and domain transferred deep networks. In Proceedings of the AAAI conference on Artificial Intelligence (Vol. 29, No. 1). https://doi.org/10.1609/aaai.v29i1.9179
15. Montoyo A, Martínez-Barco P, Balahur A (2012) Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments. Decis Support Syst 53(4):675–679
16. Maynard D, Funk A (2012) Automatic detection of political opinions in tweets. In: The Semantic Web: ESWC 2011 Workshops: ESWC 2011 Workshops, Heraklion, Greece, May 29–30, 2011, Revised Selected Papers 8. Springer, Berlin Heidelberg, pp 88–99
17. Sufi FK (2022) Identifying the drivers of negative news with sentiment, entity and regression analysis. Int J Inform Manage Data Insights 2(1):100074
18. Kalyani A, Premalatha B, Kiran KR (2018) Real time emotion recognition from facial images using raspberry Pi. Int J Adv Technol
19. Alom, M. Z., Taha, T. M., Yakopcic C, Westberg S, Sidike P, Nasrin MS ... Asari VK (2019) A state-of-the-art survey on deep learning theory and architectures. Electronics 8(3), 292
20. Ko BC (2018) A brief review of facial emotion recognition based on visual information. Sensors 18(2):401
21. Ensafi Y, Amin SH, Zhang G, Shah B (2022) Time-series forecasting of seasonal items sales using machine learning–A comparative analysis. Int J Inf Manag Data Insights 2(1):100058
22. Siersdorfer S, Minack E, Deng F, Hare J (2010) Analyzing and predicting sentiment of images on the social web. In Proceedings of the 18th ACM international conference on Multimedia (pp. 715-718)
23. Goodfellow, I. (2016). Deep learning.
24. Lampert CH, Nickisch H, Harmeling S (2009) Learning to detect unseen object classes by between-class attribute transfer. In 2009 IEEE conference on computer vision and pattern recognition (pp. 951-958). IEEE, Miami, FL, USA, 20-25 June 2009. https://doi.org/10.1109/CVPR.2009.5206594
25. Zhang Z, Saligrama V (2015) Zero-shot learning via semantic similarity embedding. In: Proceedings of the IEEE international conference on computer vision. pp. 4166–4174
26. Jetley S, Romera-Paredes B, Jayasumana S, Torr P (2015) Prototypical priors: From improving classification to zero-shot learning. Preprint at https://arxiv.org/abs/1512.01192
27. Guo Y, Ding G, Han J, Gao Y (2017) Zero-shot learning with transferred samples. IEEE Trans Image Process 26(7):3277–3290
28. Xiao-Xu QI, Wei J (2007) Application of wavelet energy feature in facial expression recognition. In 2007 international workshop on anti-counterfeiting, security and identification (ASID) (pp. 169-174). IEEE, Xizmen, China, 16-18 April 2007.
29. Zhao L, Zhuang G, Xu X (2008) Facial expression recognition based on PCA and NMF. In: 2008 7th world congress on intelligent control and automation. IEEE. pp 6826–6829
30. Feng X, Pietikäinen M, Hadid A (2007) Facial expression recognition based on local binary patterns. Pattern Recognit Image Anal 17:592–598
31. Lee CC, Shih CY, Lai WP, Lin PC (2012) An improved boosting algorithm and its application to facial emotion recognition. J Ambient Intell Humaniz Comput 3:11–17
32. Chang CY, Huang YC (2010) Personalized facial expression recognition in indoor environments. In The 2010 international joint conference on neural networks (IJCNN) (pp. 1-8). IEEE, Barcelona, Spain, 18-23 July 2010. https://doi.org/10.1109/IJCNN.2010.5596316

33. Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis Comput 27(6):803–816
34. Jabid T, Kabir MH, Chae O (2010) Robust facial expression recognition based on local directional pattern. ETRI J 32(5):784–794
35. Alshamsi H, Kepuska V, Meng H (2017) Real time automated facial expression recognition app development on smart phones. In 2017 8th IEEE annual information technology, electronics and mobile communication conference (IEMCON) (pp. 384-392). IEEE, Vancouver, BC, Canada, 03-05 October 2017. https://doi.org/10.1109/IEMCON.2017.8117150
36. Joseph A, Geetha P (2020) Facial emotion detection using modified eyemap–mouthmap algorithm on an enhanced image and classification with tensorflow. Vis Comput 36(3):529–539
37. Fallahzadeh MR, Farokhi F, Harimi A, Sabbaghi-Nadooshan R (2021) Facial expression recognition based on image gradient and deep convolutional neural network. J AI Data Mining 9(2):259–268
38. Mohammed SB, Abdulazeez AM (2021) Deep convolution neural network for facial expression recognition. PalArch's J Archaeol Egypt/Egyptol 18(4):3578–3586
39. Hung JC, Lin KC, Lai NX (2019) Recognizing learning emotion based on convolutional neural networks and transfer learning. Appl Soft Comput 84:105724
40. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M ... Sánchez CI (2017) A survey on deep learning in medical image analysis. Med Image Anal 42, 60-88
41. Kaur H, Ahsaan SU, Alankar B, Chang V (2021) A proposed sentiment analysis deep learning algorithm for analyzing COVID-19 tweets. Inf Syst Front 23(6):1417–1429
42. Porușniuc GC, Leon F, Timofte R, Miron C (2019) Convolutional neural networks architectures for facial expression recognition. In 2019 E-health and bioengineering conference (EHB) (pp. 1-6). IEEE, Iasi, Romania, 21-23 November 2019. https://doi.org/10.1109/EHB47216.2019.8969930
43. Zhao X, Shi X, Zhang S (2015) Facial expression recognition via deep learning. IETE Tech Rev 32(5):347–355
44. Pranav E, Kamal S, Chandran CS, Supriya MH (2020) Facial emotion recognition using deep convolutional neural network. In 2020 6th international conference on advanced computing and communication systems (ICACCS) (pp. 317-320). IEEE, Coimbatore, India, 06-07 March 2020. https://doi.org/10.1109/ICACCS48705.2020.9074302
45. Wen G, Hou Z, Li H, Li D, Jiang L, Xun E (2017) Ensemble of deep neural networks with probability-based fusion for facial expression recognition. Cogn Comput 9(5):597–610
46. Ruiz-Garcia A, Elshaw M, Altahhan A, Palade V (2017) Stacked deep convolutional auto-encoders for emotion recognition from facial expressions. In 2017 international joint conference on neural networks (IJCNN) (pp. 1586-1593). IEEE, Anchorage, AK, USA, 14-19 May 2017. https://doi.org/10.1109/IJCNN.2017.7966040
47. Jain N, Kumar S, Kumar A, Shamsolmoali P, Zareapoor M (2018) Hybrid deep neural networks for face emotion recognition. Pattern Recogn Lett 115:101–106
48. Shaees S, Naeem H, Arslan M, Naeem MR, Ali SH, Aldabbas H (2020) Facial emotion recognition using transfer learning. In 2020 international conference on computing and information technology (ICCIT-1441) (pp. 1-5). IEEE, Tabuk, Saudi Arabia, 09-10 September 2020. https://doi.org/10.1109/ICCIT-144147971.2020.9213757
49. Shi M, Xu L, Chen X (2020) A novel facial expression intelligent recognition method using improved convolutional neural network. IEEE Access 8:57606–57614
50. Porcu S, Floris A, Atzori L (2020) Evaluation of data augmentation techniques for facial expression recognition systems. Electronics 9(11):1892
51. Finn C, Abbeel P, Levine S (2017) Model-agnostic meta-learning for fast adaptation of deep networks. In International conference on machine learning (pp. 1126–1135). PMLR
52. Gidaris S, Bursuc A, Komodakis N, Pérez P, Cord M (2019) Boosting few-shot visual learning with self-supervision. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 8059–8068)
53. Vinyals O, Blundell C, Lillicrap T, Wierstra D (2016) Matching networks for one shot learning. Advances in neural information processing systems, 29
54. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G (2023) Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv 55(9):1–35
55. Gao T, Fisch A, Chen D (2020) Making pre-trained language models better few-shot learners. Preprint at https://arxiv.org/abs/2012.15723
56. Jian Y, Gao C, Vosoughi S (2022) Contrastive learning for prompt-based few-shot language learners. Preprint at arXiv preprint https://arxiv.org/abs/2205.01308
57. Hosseini-Asl E, Liu W (2023) U.S. Patent No. 11,853,706. Washington, DC: U.S. Patent and Trademark Office

58. Tsimpoukelli M, Menick JL, Cabi S, Eslami SM, Vinyals O, Hill F (2021) Multimodal few-shot learning with frozen language models. Adv Neural Inf Process Syst 34:200–212
59. Yu Y, Zhang D (2022) Few-shot multi-modal sentiment analysis with prompt-based vision-aware language modeling. In 2022 IEEE international conference on multimedia and expo (ICME) (pp. 1-6). IEEE, Taipei, Taiwan, 18-22 July 2022. https://doi.org/10.1109/ICME52920.2022.9859654
60. Yu Y, Zhang D, Li S (2022) Unified multimodal pre-training for few-shot sentiment analysis with prompt-based learning. In Proceedings of the 30th ACM international conference on multimedia (pp. 189–198)
61. Qiao S, Liu C, Shen W, Yuille AL (2018) Few-shot image recognition by predicting parameters from activations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 7229–7238
62. Shaban A, Bansal S, Liu Z, Essa I, Boots B (2017) One-shot learning for semantic segmentation. Preprint at https://arxiv.org/abs/1709.03410
63. Liu B, Yu X, Yu A, Zhang P, Wan G, Wang R (2018) Deep few-shot learning for hyperspectral image classification. IEEE Trans Geosci Remote Sens 57(4):2290–2304
64. Ashrafi I, Mohammad M, Mauree AS, Habibullah KM (2019) Attention guided relation network for few-shot image classification. In: Proceedings of the 7th international conference on computer and communications management. pp 177–180
65. Wang YX, Gui L, Hebert M (2017) Few-shot hash learning for image retrieval. In proceedings of the IEEE international conference on computer vision workshops (pp. 1228–1237)
66. Yu Y, Bian N (2020) An intrusion detection method using few-shot learning. IEEE Access 8:49730–49740
67. Singh R, Bharti V, Purohit V, Kumar A, Singh AK, Singh SK (2021) MetaMed: few-shot medical image classification using gradient-based meta-learning. Pattern Recogn 120:108111
68. Zhang X, Wang C, Tang Y, Zhou Z, Lu X (2021) A survey of few-shot learning and its application in industrial object detection tasks. In: international workshop of advanced manufacturing and automation. Springer Singapore, Singapore, pp 637–647
69. Wang Y, Yao Q, Kwok JT, Ni LM (2020) Generalizing from a few examples: a survey on few-shot learning. ACM Comp Surv (csur) 53(3):1–34
70. Ying L, Yan-Bo L, Jiu-Lun F, Fu-Ping W, Yan-Chao G, Qi T (2021) Survey on image classification technology based on small sample learning. Acta Automatica Sinica 47(2):297–315
71. Du Y, Xie R, Zhang B, Yin Z (2024) FMCF: few-shot multimodal aspect-based sentiment analysis framework based on contrastive finetuning. Appl Intell 1–15
72. Cen J, Qing C, Ou H, Xu X, Tan J (2024) MASANet: multi-aspect semantic auxiliary network for visual sentiment analysis. IEEE Trans Affect Comput
73. Fei-Fei L, Fergus R, Perona P (2006) One-shot learning of object categories. IEEE Trans Pattern Anal Mach Intell 28(4):594–611
74. Mahajan K, Sharma M, Vig L (2020) Meta-dermdiagnosis: few-shot skin disease identification using meta-learning. In proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 730–731)
75. Frikha A, Krompaß D, Köpken HG, Tresp V (2021) Few-shot one-class classification via meta-learning. In proceedings of the AAAI conference on artificial intelligence, Vol. 35, No. 8, pp 7448–7456)
76. Deng S, Hua W, Wang B, Wang G, Zhou X (2020). Few-shot human activity recognition on noisy wearable sensor data. In: database systems for advanced applications: 25th International Conference, DASFAA 2020, Jeju, South Korea, September 24–27, 2020, Proceedings, Part II 25. Springer International Publishing. pp 54–72
77. Nie L, Li X, Gong T, Zhan D (2022) Few shot learning-based fast adaptation for human activity recognition. Pattern Recogn Lett 159:100–107
78. Feng S, Duarte MF (2019) Few-shot learning-based human activity recognition. Expert Syst Appl 138:112782
79. Li Y, Chao X (2021) Semi-supervised few-shot learning approach for plant diseases recognition. Plant Methods 17:1–10
80. Liu Y, Zhang H, Zhang W, Lu G, Tian Q, Ling N (2022) Few-shot image classification: current status and research trends. Electronics 11(11):1752
81. Ganesha HS, Gupta R, Gupta SH, Rajan S (2024) Few-shot transfer learning for wearable IMU-based human activity recognition. Neural Comput Appl 36(18):10811–10823
82. Ge Y, Guo Y, Das S, Al-Garadi MA, Sarker A (2023) Few-shot learning for medical text: A review of advances, trends, and opportunities. J Biomed Inform 104458. https://doi.org/10.1016/j.jbi.2023.104458

83. Chen Y, Xu X, Liu C (2024) Few-shot meta transfer learning-based damage detection of composite structures. Smart Mater Struct 33(2):025027
84. Goodfellow IJ, Erhan D, Carrier PL, 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. Springer Berlin Heidelberg. pp 117–124
85. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) 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. IEEE. pp 94–101
86. Mollahosseini A, Hasani B, Mahoor MH (2017) Affectnet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans Affect Comput 10(1):18–31
87. Meena G, Mohbey KK, Indian A, Khan MZ, Kumar S (2024) Identifying emotions from facial expressions using a deep convolutional neural network-based approach. Multimed Tools Appl 83(6):15711–15732
88. Meena G, Mohbey KK, Kumar S, Lokesh K (2023) A hybrid deep learning approach for detecting sentiment polarities and knowledge graph representation on monkeypox tweets. Decis Anal J 7:100243

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.