Hindawi
IET Signal Processing
Volume 2024, Article ID 7914185, 13 pages
https://doi.org/10.1049/2024/7914185

Research Article

Lie Detection Technology of Bimodal Feature Fusion Based on Domain Adversarial Neural Networks

Yan Zhou and Feng Bu

Suzhou Vocational University, Suzhou, China

Correspondence should be addressed to Feng Bu; 92010@jssvc.edu.cn

Received 29 August 2023; Revised 7 February 2024; Accepted 22 February 2024; Published 2 March 2024

Academic Editor: Wanli Wen

In the domain of lie detection, a common challenge arises from the dissimilar distributions of training and testing datasets. This causes a model mismatch, leading to a performance decline of the pretrained deep learning model. To solve this problem, we propose a lie detection technique based on a domain adversarial neural network employing a dual-mode state feature. First, a deep learning neural network was used as a feature extractor to isolate speech and facial expression features exhibited by the liars. The data distributions of the source and target domain signals must be aligned. Second, a domain-antagonistic transfer-learning mechanism is introduced to build a neural network. The objective is to facilitate feature migration from the training to the testing domain, that is, the migration of lie-related features from the source to the target domain. This method results in improved lie detection accuracy. Simulations conducted on two professional lying databases with different distributions show the superiority of the detection rate of the proposed method compared to an unimodal feature detection algorithm. The maximum improvement in detection rate was 23.3% compared to the traditional neural network-based detection method. Therefore, the proposed method can learn features unrelated to domain categories, effectively mitigating the problem posed by different distributions in the training and testing of lying data.

1. Introduction

Lying is a complex psychological state in which an individual deliberately misleads others through the employment of false statements, distortion of facts, or intentional omission of information.

Lie detection is an important research field [1, 2] within computer linguistics, psychology, military science, and other disciplines. The psychological phenomenon of lying is a complex sensation influenced by the interactions between emotions, cognition, and will. The processes governing the generation and alteration of the psychological state of lying can affect human physiological characteristics. These changes involve speech signal characteristics, facial expressions, and EEG signals. Therefore, in the research process, it is necessary to adopt a comprehensive approach that takes into account the influence of various factors.

In actual lie detection scenarios, the distributions of the training and testing data differ because of the factors such as inconsistent collection environments and differences in collection methods. Pretrained deep learning models may experience a sharp deterioration in detection algorithm performance and model mismatch when used in practice, and a large amount of unlabeled test data cannot be fully utilized. It is important to use the transfer learning techniques to obtain transferable knowledge on different data distributions. This can be used to explore the correlation between training and testing data to obtain multimodal information. Therefore, studies on lie detection in real scenarios are important for addressing psychological computing problems.

Currently, lie detection leverages a spectrum of signals, including speech, facial expressions, and other physiological indicators [3, 4]. In the realm of speech-based detection methods, employing psycholinguistic features based on language inquiry has been proposed as an effective means for lie detection [5]. Studies have shown that the pitch, duration, energy, and pauses during speech can provide information on lying [6–8]. In a similar vein, Elliott and Leach [9] analyzed speaking proficiency to judge whether a speaker was lying. The specific operational method focused on the

relationship between language proficiency and lying through the random testing of people with different levels of English proficiency. In a previous study by Dai et al. [10], speech spoofing detection based on big data and machine learning was proposed. Among non-speech-based detection methods based on facial physiological characteristics, thermal imaging has emerged as a valuable tool for measuring facial blood flow and skin temperature [11, 12]. Some studies have found that facial micro-expressions such as protruding lips and symbolic gestures may be signs of lying [13, 14]. Moreover, some researchers have detected lies by measuring cerebral blood flow using functional brain magnetic resonance imaging or by constructing a multichannel lie detection system based on cardiac impact signals [15, 16].

Lie detection technology based on multifeature fusion has received significant attention. In the literature [17, 18], lie detection methods not only extract basic features from audio, video, and text modalities but also use manually annotated micro-expression features. In another study by Li [19], convolutional neural networks (CNNs) and long short-term memory deep learning models were used to extract audio and video features. In another study by Mathur and Matarić [20], a multimodal model was proposed to detect lies by combining the acoustic, vision, and text modalities. A multifeature noncontact lie detection technique was developed. Although, considerable research has been dedicated to voice-based lie detection [21–24], these methods often require a large number of training samples and have significant sensitivity to data discrepancies when the training and testing datasets originate from the different distributions. In addition, improvements in lie detection model performance obtained by combining the features of different modalities must be compared and verified repeatedly. Therefore, distinguishing the most effective lie features requires further investigation.

In this study, a lie detection method is developed based on a domain adversarial neural network (DANN) and bimodal features. The detection feature is a fusion of speech signals and facial expressions. Ordinarily, training and testing data from different domains have individual characteristics and distributions, which lead to a mismatch problem in the pretrained deep learning model. The purpose of this study was to solve the problem of an actual lie detection scene. Labeled data were used in the training process, and unlabeled data were used in the testing process. Moreover, the correlation between labeled and unlabeled data from different domains should be fully investigated. This can improve the lie detection performance by extracting and fuzing bimodal features of the data. The proposed DANN-based lie detection technology utilizes the adversarial competitive relationship in a DANN. It learns lie features from data that do not contain the domain category information. This provides a good solution to the domain mismatch problem faced by lie detection models and improves their performance.

This study proposes a new lie detection model based on a multimodal domain adversarial neural network that integrates speech and facial expression features to detect lies. The experimental results demonstrated that the extracted speech and facial expression features were significant indicators for detecting lies, and the accuracy of lie detection significantly improved through the combination of multiple features. Compared to the existing methods, the method proposed in this study has significant advantages and outstanding contributions in the following two aspects:

First, as concerns model selection, the mismatch between the pretrained deep learning models was considered. The pretrained system has the problem of sharp degradation in performance owing to the distribution differences between the training and testing data. This study uses the DANN model, which introduces the idea of adversarial learning in transfer learning. This deep learning model focuses on the selection of transferable features between the different domains and achieves good classification. The model in this study is more suitable for the actual requirements of lie detection scenarios.

Second, in terms of feature selection, the proposed lie detection model effectively addresses the problems of low accuracy and poor model robustness reported in previous studies. This method integrates speech and facial expressions to comprehensively determine the lying state of a person from multiple perspectives. For speech signals, this study designs a deep separable convolutional neural network (DSCNN) that can learn speech features well and has absolute advantages in lightweight aspects. A sparse CNN (SCNN) was used for facial expression signals. It can comprehensively learn facial features at each stage. In addition, this study introduces a multihead attention mechanism to search for the credibility of various modal features. Decision fusion can be achieved through weight allocation. Therefore, the proposed method yielded more comprehensive and accurate classification results.

In summary, the proposed multifeature lying psychological-state detection method based on DANNs represents an advancement in psychological computing. The DANN provides a good solution for alleviating domain mismatch problems. The core concept is to learn the discriminative features of lying states from source and target domain data. These features do not contain domain category information.

2. Principle of Domain Adversarial Neural Networks

A DANN alleviates the problem of domain mismatch by making feature vectors contain no domain-specific information [25–28]. Typically, domain adversarial transfer learning consists of a feature extractor G_f, a domain discriminator G_d, and a tag predictor G_y. The purpose of G_f is to learn the domain-invariant feature representations to confuse G_d. However, G_d attempts to distinguish the characteristics of source domain samples from those of target domain samples. The construction of G_y is designed to classify the objects into different categories. The purpose of domain adversarial training is to reduce the distribution differences between the source and target domains using a zero-sum game process. The network diagram is shown in Figure 1.

The structure consists of a feature extractor, a domain discriminator, and two label classifiers for adversarial

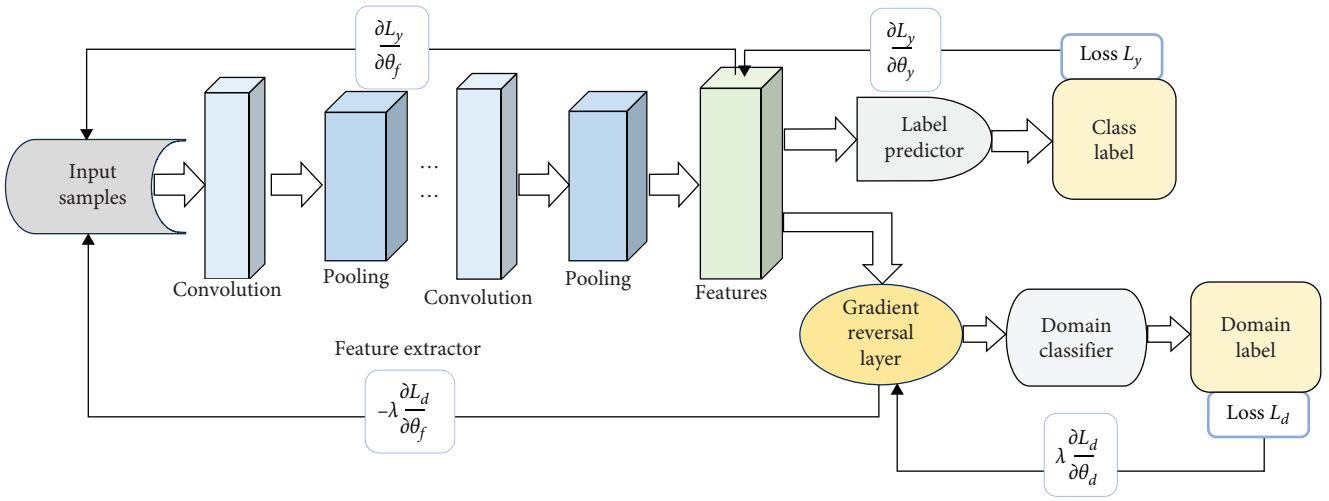

FIGURE 1: Network diagram of DANN.

interactions. The network comprises two flow directions. The source domain data, labeled as lying data, are used as the input data for the first flow. The other flow handles target domain data, which contain unlabeled data. The source- and target-domain data flow through the feature extractor. The source-domain data flow into the label classifier and are used to calculate the source-domain label classification loss. The source and target domain data flow jointly into the domain classifier and the domain classification loss is calculated. The optimization goal of the network is to minimize the source domain label classification loss while maximizing the domain classification loss. The goal is to determine domain-independent feature parameters.

In the training phase, the weight parameters of the feature extractor and domain classifier are optimized and updated according to the loss function. This phase ceases when the domain classifier cannot determine whether the input data are from the source or the target domain datasets. To ensure that the learned features do not contain domain class information, the DANN model introduces a gradient reversal layer (GRL) between the domain classifier and the feature generation network. The GRL is located between the feature layer and domain classifier. The error gradient of the domain classifier is transmitted to the feature generation network through backpropagation. It is multiplied by a negative value before propagating to the feature generation network. The purpose is to prevent the domain classifier from distinguishing the feature vectors generated by the input data. This ensures that the feature distributions of different domain data in the feature space tend to be consistent. The mathematical definition of DANN can be expressed as follows:

$$\mathscr{E}(\theta_f,\theta_y,\theta_d)=\frac{1}{n}\sum_{i=1}^{n}\mathscr{L}_y^i(\theta_f,\theta_y)-\lambda\left(\frac{1}{n}\sum_{i=1}^{n}\mathscr{L}_d^i(\theta_f,\theta_d)+\frac{1}{m}\sum_{i=1}^{m}\mathscr{L}_d^i(\theta_f,\theta_d)\right), \tag{1}$$

where

$$\mathscr{L}_y^i(\theta_f,\theta_y)=\mathscr{L}_y\left(G_y\left(G_f\left(x_i;\theta_f\right);\theta_y,y_i\right)\right), \tag{2}$$

$$\mathscr{L}_d^i(\theta_f,\theta_d)=\mathscr{L}_d\left(G_d\left(G_f\left(x_i;\theta_f\right);\theta_d,d_i\right)\right). \tag{3}$$

Here, x_i denotes the classification label of the ith sample, and y_i denotes the classification label of the i-th sample. G_f, G_y, and G_d, respectively, represent the weight parameters of the feature extraction, label classification, and domain discrimination layers, and $\mathscr{L}_y^i$ and $\mathscr{L}_d^i$, respectively, represent the loss functions of the i-th sample passing through the label classifier and domain discriminator, respectively. The hyperparameter λ is a weighting factor used to balance the contributions of the label and domain discriminators to the target loss function. The weight parameters of the feature extractor G_f and label classifier G_y are updated to minimize the loss function of the DANN. The dataset contains n labeled source domain datasets D_s and m unlabeled target domain datasets D_t.

$$\left(\widehat{\theta}_f,\widehat{\theta}_y\right)=\underset{\theta_f,\theta_y}{\operatorname{argmin}}\ \mathscr{E}\left(\theta_f,\theta_y,\widehat{\theta}_d\right). \tag{4}$$

The training of parameter θ_d is performed in a manner exactly the opposite of that of the feature extraction and label classification layers. It is updated alternately with the parameters of the feature extractor θ_f and the label classifier θ_y. Given the parameters θ_f and θ_y, the objective function of the DANN is maximized. This makes the distributions of the source and target domains close to each other, rendering distinguishing them using a domain classifier difficult. The data are defined as follows:

$$\widehat{\theta}_d=\underset{\theta_d}{\operatorname{argmax}}\ \mathscr{E}\left(\widehat{\theta}_f,\widehat{\theta}_y,\theta_d\right). \tag{5}$$

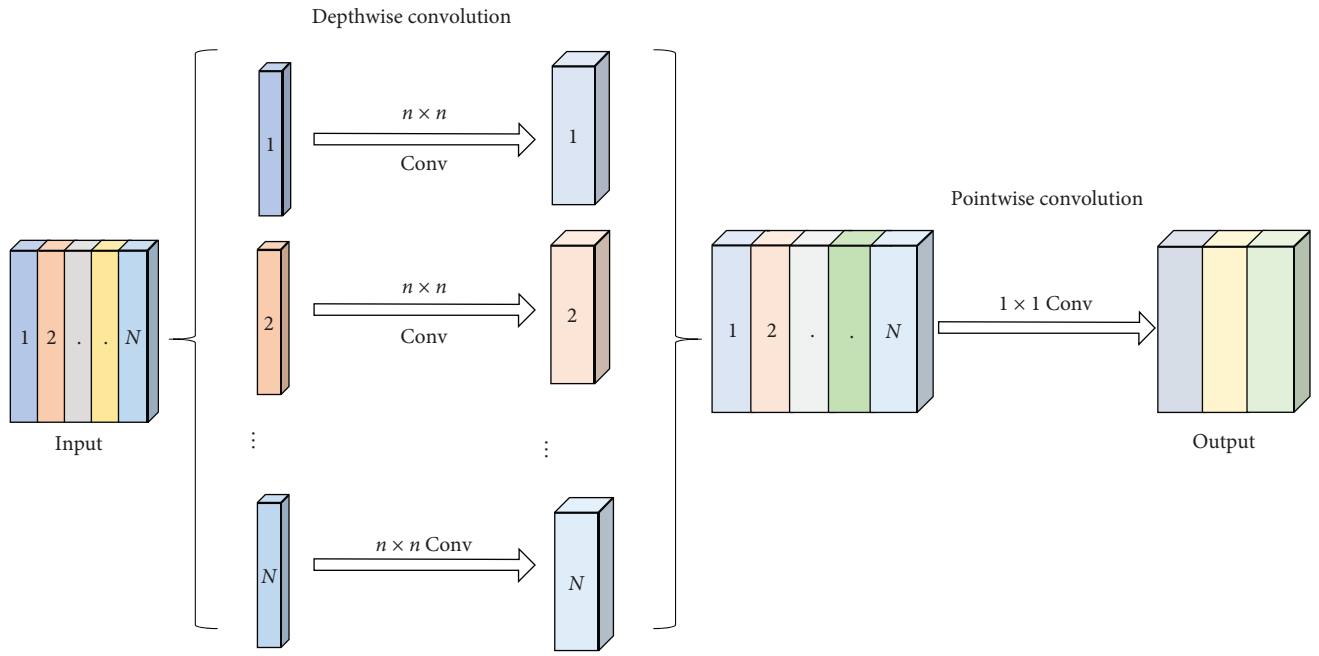

Figure 2: Schematic diagram of DSCNN.

During the training phase, the model utilizes the maximum–minimum strategy to update the parameters in the model alternately according to Equations (4) and (5). The training process continues until the model produces discriminative features that do not contain domain category information. This alleviates the problem of mismatched features between the training and testing domains.

3. Bimodal Lie Feature Fusion

A lie detection technique based on DANNs using speech and facial expressions with bimodal features has been developed. The detection model learns speech and facial expression signals using the DANN method, and the domain spatial representation features are obtained. The purpose is to realize domain-independent features of lie detection knowledge to improve the success rate of lie detection.

3.1. Feature Extraction

(1) Speech Lying Feature Extraction. The speech feature extractor uses a DSCNN as the feature extraction network. This ensures classification accuracy by minimizing the network parameters to the extent possible. In contrast to traditional convolution, deep separable convolution first performs a separate convolution operation on each channel, which is called "depthwise convolution". Subsequently, a 1×1 convolution operation is performed to combine several outputs, which is called "pointwise convolution." A schematic representation of the DSCNN is shown in Figure 2.

In the DSCNN, the calculation of depthwise convolution uses a convolutional kernel for each channel of the input feature. Subsequently, the outputs from all these convolutional kernels are concatenated to obtain the final output. In contrast, pointwise convolution is a 1×1 convolution that assumes a dual role in the DSCNN. First, it enables the DSCNN to freely change the number of output channels. Second, it performs channel fusion on the features, facilitating the mapping of the output through the depthwise convolution. The computational complexity of the DSCNN is approximately $1/C$ that of traditional convolution, where C is the number of convolution kernels. Therefore, the utilization of deep separable convolutions leads to a substantial reduction in computational complexity compared to that of a traditional CNN.

It is necessary to extract the lying speech feature sets from the source and target domains using a DSCNN. The extraction process unfolds in several steps. First, 16 low-level descriptions (LLDs) are extracted from the speech signals: zero-crossing rate, root-mean square, pitch frequency (normalized to 500 Hz), harmonic-to-noise ratio, and mel-frequency cepstral coefficients in the range 1–12. Second, the first-order Δ coefficients for these 16 LLD features are calculated to obtain 16 new coefficient features. Finally, a set of statistical functions containing 12 categories is obtained: mean, standard deviation, kurtosis, skewness, minimum, maximum, relative position, range, two linear regression coefficients, and their mean square errors. By performing these computations across 32 features, spanning the 12 categories, a final feature set that included $16 \times 2 \times 12 = 384$ attributes is obtained.

(2) Lying Facial Expression Features Extraction. Facial expression features are extracted using a SCNN. These features specifically pertain to lying and are extracted from the source and target domains. A total of 68 key facial expression points are detected. Three different features of the facial expression images are extracted: texture, shape, and spatial relationships. Texture features are used to analyze local texture information in the expression images. Shape features are used to generate the corresponding models based on the external shape information of the trained expression image

and to match them with the measured image. Spatial relationship features are used to extract features based on the spatial position relationships of the most important parts of an image.

A set of lying facial expression training samples is assumed to be given as $(x^1, y^1), (x^2, y^2), \ldots, (x^m, y^m)$, where y^i $(i = 1, 2, \ldots, m)$ is the training sample with labels and the objective function of the SCNN is expressed as follows:

$$J = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}\frac{1}{2}\left\|\gamma_{w,b}(x^i)_k - y_k^i\right\|^2, \tag{6}$$

where w is the weight, b is the corresponding bias, k is the number of expected classifications, and $\gamma_{w,b}(x_k^i)$ is an important k-dimensional vector. The cost function of the SCNN is based on the l_1 norm sparse constraint. The objective function of the SCNN is expressed as follows:

$$J_\theta = J(w, b) + \mu L(w, b). \tag{7}$$

Here, the sparse constraint function can be expressed as follows:

$$L(w, b) = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}\frac{1}{2}\ln\left(1 + \frac{\left(\gamma_{w,b}(x^i)_k - y_k^i\right)^2}{f^2}\right). \tag{8}$$

From Equations (6)–(8), the objective function can be expressed as follows:

$$J_\theta = J(w, b) + \mu\sum_{i=1}^{m}\sum_{k=1}^{K}\frac{1}{2m}\ln\left(1 + \frac{\left(\gamma_{w,b}(x^i)_k - y_k^i\right)^2}{f^2}\right), \tag{9}$$

$$J_\theta = \frac{1}{m}\sum_{i=1k=1}^{m}\sum^{K}\frac{1}{2}\left\|\gamma_{w,b}(x^i)_k - y_k^i\right\|^2 + \mu\sum_{i=1k=1}^{m}\sum^{K}\frac{1}{2m}\ln\left(1 + \frac{\left(\gamma_{w,b}(x^i)_k - y_k^i\right)^2}{f^2}\right). \tag{10}$$

It also can be expressed as follows:

$$J_\theta = \frac{1}{2m}\sum_{i=1}^{m}\sum_{k=1}^{K}\left(\left\|\gamma_{w,b}(x^i)_k - y_k^i\right\|^2 + \mu\ln\left(1 + \frac{\left(\gamma_{w,b}(x^i)_k - y_k^i\right)^2}{f^2}\right)\right), \tag{11}$$

where $\mu\ln(1 + \frac{(\gamma_{w,b}(x^i)_k - y_k^i)^2}{f^2})$ in Equation (11) is part of the sparse constraint term, and μ is the regularization coefficient. The values of μ and f generally affect the computational performance of the SCNN. The output of the fully connected layer is the extracted feature.

3.2. Feature Fusion Mechanism. The DSCNN and SCNN are used to obtain two modal data features for the speech and facial expression signals. The multihead attention mechanism is then deployed to achieve the interactive learning of attention features within and between speech and facial expressions. Finally, multimodal fusion is performed on the learning features, culminating in the generation of an inference prediction output. Using this method, features closely related to lying can be enhanced, while simultaneously the significance of features unrelated to deception can be diminished. This results in more accurate multimodal features. The calculation process for the multihead self-attention (MHSA) is as follows:

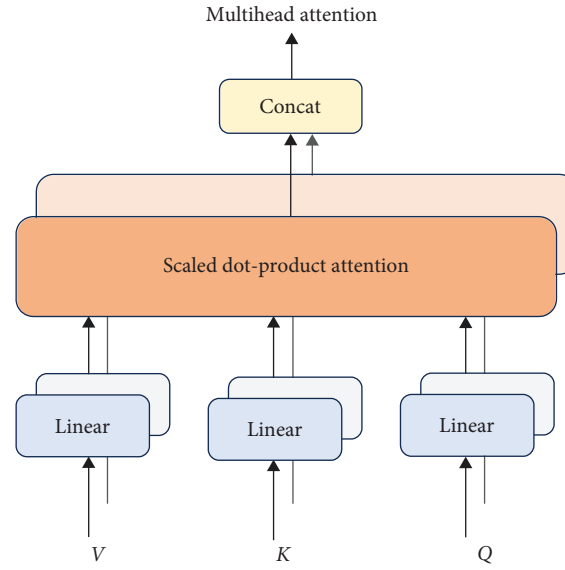

FIGURE 3: Structure of MHSA mechanism.

$$\text{Multihead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^o, \tag{12}$$

$$\text{where head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right), \tag{13}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{14}$$

where Q, K, and V represent the input matrices, $\sqrt{d_k}$ represents the dimension of the K matrix, and $Q = qw_{q1}, K = kw_{k1}, V = vw_{v1}$, where $q_1 = F_S, F_s$ is the lying speech feature, F_E is the facial expression feature, $F_E = k_1 = v_1$,, and w_{q1}, w_{k1}, w_{v1} express the training parameters. Matrices Q and K are transformed by multiplication and then multiplied by matrix V using the softmax function. By adopting this method, the MHSA can search for commonalities between speech and facial expression features. The structure of the MHSA is shown in Figure 3.

3.3. Lie Detection Based on Bimodal Features. A DANN that simultaneously uses bimodal feature information with speech and facial expression signals is used in this study. The algorithm aligns the data distribution of the signals in the source and target domains to improve the lie detection performance. In this study, the MHSA feature-fusion method is adopted to

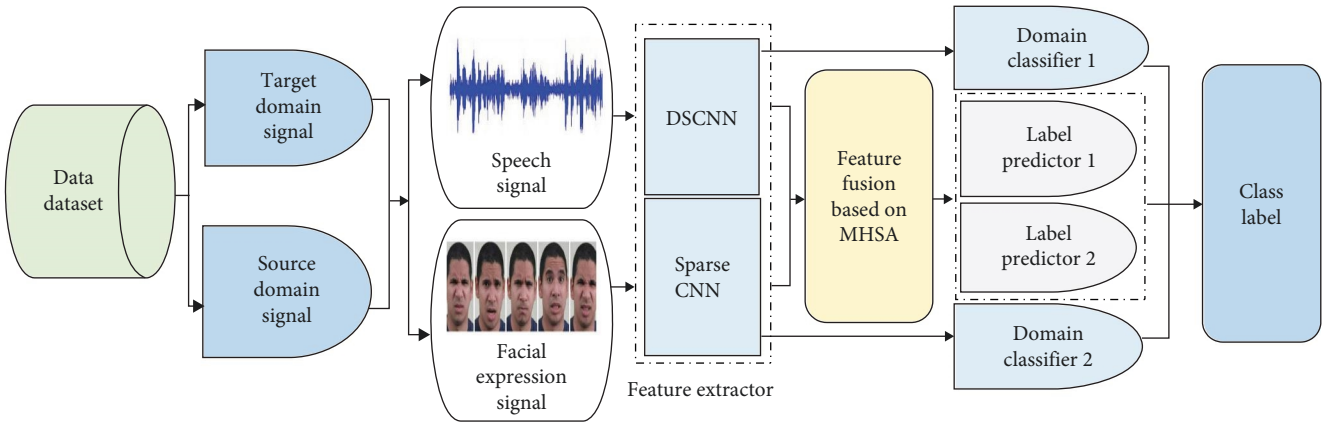

Figure 4: Diagram of lie detection model based on bimodal features.

fuze the features of various modalities. This approach enables the creation of a detection methodology that seamlessly integrates information from the different modalities. The proposed dual-modal DANN lie detection model is shown in Figure 4.

The input data for the detection model are F_S and F_E, representing the features derived from speech and facial expressions, respectively. The source domain contains labeled data, whereas the target domain contains unlabeled data. The training process of the detection algorithm is as follows:

Step 1: The multimodal information from the source domain data and target domain pertaining to speech signal features is generated using feature extractor 1. In total, 300 speech-feature dimensions are generated in the hidden layer.

Step 2: The modal information from the source domain data and the target domain of the facial expression features is generated using feature extractor 2. A total of 300 facial expression feature dimensions are generated in the hidden layer.

Step 3: Speech features are input into domain classifier 1 to classify the domain and calculate the domain classification loss function. Simultaneously, the facial expression features are input into domain classifier 2 to perform domain classification and calculate the domain classification loss function.

Step 4: In the feature fusion technology, an MHSA mechanism is used to fuze speech and facial expression features. The fuzed feature length has 300 dimensions. The feature is input into the label classifier for source domain label classification, and the loss is calculated.

Step 5: The two parts of the domain classification are combined to obtain the total loss for gradient backpropagation. Label classification loss, speech feature domain classification loss, and facial expression domain classification loss are combined to obtain the total loss.

Step 6: After training, the target domain test data are input into the feature extractor for deep feature extraction. The classifier then detects the liar. The final optimization goal is as follows:

$$\begin{aligned} E\left(\theta_{f_1}, \theta_{f_2}, \theta_y, \theta_{d_1}, \theta_{d_2}\right) = & \sum_{i=1}^{n_s} \mathscr{L}_y\left(R_y\left(R_{f_1}\left(x_i^s\right) + R_{f_2}\left(x_i^s\right), y_i^s\right)\right) \\ & - \lambda_1 \sum_{i=1}^{n_s+n_t} \mathscr{L}_{d_1}\left(R_{d_1}\left(R_{f_1}\left(X_j\right)\right), d_i\right) \\ & - \lambda_2 \sum_{i=1}^{n_s+n_t} \mathscr{L}_{d_2}\left(R_{d_2}\left(R_{f_2}\left(X_j\right)\right), d_i\right) \end{aligned} \tag{15}$$

where $R_{f_1}, R_{f_2}, R_y, R_{d_1}$, and R_{d_2} are, respectively, feature extractor 1, feature extractor 2, the source domain label classifier, domain discriminator 1, and domain discriminator 2. The parameters $\theta_{f_1}, \theta_{f_2}, \theta_y, \theta_{d_1}, \theta_{d_2}$ belong to feature extractor 1, feature extractor 2, the source domain label classifier, domain discriminator 1, and domain discriminator 2, respectively. Where n_s and n_t are the numbers of samples in the source and target domains, respectively. The parameters y_i and d_i are category and domain labels, respectively, and λ_1, λ_2 are the weight coefficients. Domain discriminator 1, domain discriminator 2, and the source domain label classifier use a cross-entropy function to minimize the total loss and optimize the model through iterative training. Target domain recognition and classification use trained feature classifiers and source-domain label classifiers. The DSCNN and SCNN are used as feature extractors. The structures of domain classifiers 1 and 2 are identical.

4. Simulation Experiments and Analysis

4.1. Datasets. In this study, the model environment was a TensorFlow deep learning framework. The algorithm runs on the Windows 10 operating system, the graphics card is a GTX1050, and the processor was trained using a graphics processing unit. The experimental dataset used was the open-source real-life trial dataset (RLTD) [29] comprised real court trial videos. It consisted of 121 court trial video clips, including 61 videos depicting deceptive statements and 60 videos with truthful statements. The average durations of the deceptive and truthful videos were 27.7 and 28.3 s,

Table 1: Description of RLTD and SULD datasets.

Corpus	Language	Text type	Acquisition method	Data type	Total data
RLTD	English	Fixed	Natural	Audio and video	121 Court trial video clips
SULD	Chinese	Unfixed	Natural and intentional	Audio and video	300 Voice segments

respectively. MP4 format video data were converted into WAV-format audio data to obtain an audio dataset. Meanwhile, facial expression images of the speaker were collected from MP4 video data, and the collection time was synchronized with the audio data. Additionally, it is necessary to ensure that the target domain data are distributed differently from the source domain data. The self-developed Suzhou University Lying Database (SULD) [30, 31], which includes three parts: induced lying speech, deliberately imitative lying speech, and natural lying speech, was also used. In this study, the induced-lying speech component was used. The SULD database was recorded in a quiet environment through conversations. The content of the corpus included student dormitory conflicts and students' opinions of teachers during the recording process. Each participant recorded five different types of language materials, including students' cheating on exams and their emotional status. Each participant recorded five segments, resulting in 300 audio and video signals. All data were used for training and testing. The speakers' facial expression data were collected using the method described above. The database descriptions are presented in Table 1.

4.2. Simulation Parameter Settings. The structure of the lie detection model based on the multimodal DANN used in this study refers to [32], which mainly includes three parts: a feature extractor, domain classifier, and label predictor. The parameter λ in the domain adversarial network in this paper is set to 0.01, epoch is set to 100, and batch size is set to 32. First, the DSCNN parameter settings for extracting speech features were improved based on [33], using three convolutional layers with a convolutional kernel size of 3×3. The input is a Mel-level filter bank feature, calculated every 10 ms within a 25 ms window, and the final linear layer produces an output. The gradient norm, learning rate, dropout, label smoothing rate, and random sampling rate are set to 15, 0.05, 0.1, 0.05, and 0.01, respectively. Second, the sparse CNN model used for extracting facial expression features was improved on the basis of [34], with four convolutional layers: the number of convolutional kernels was (256,512,512,512), kernel size was set to 3×3, and step size was set to 2×1. Pooling operations were performed in each layer with a pooling size of 2×2 and step size of 1×1. The activation function for each convolutional layer was a leaky rectified linear unit, and the dropout layer probability was set to 0.25. A sparse representation layer was added before the convolutional layer. The K-singular value decomposition sparse representation algorithm [35, 36] was implemented during the sparse transformation process.

4.3. Analysis of Experimental Results

4.3.1. Analysis of Feature Dimensions across Different Domains. In general, high-dimensional features can capture more information, yet they may also introduce redundancy and dimensional complexity. Conversely, the lower dimensions inherently convey reduced information. The DANN model, renowned for its capability to acquire feature vectors that are devoid of domain-specific categories, exhibits strong generalization. Nevertheless, the dimensions of the learned feature vectors may affect their ability to represent input data in the feature space. In this experiment, the impact of the feature dimensions on the model recognition performance was evaluated by setting different dimension values.

This experiment verifies the detection performance of the DANN model with feature dimensions of 50, 100, 150, 200, 250, 300, 350, 400, 450, and 500 pixels. The detection performance for different vector dimensions was verified using a tenfold cross-validation method. The recognition performances of the model for the different vector dimensions are shown in Figure 5. In the figure, "SD" and "TD" represent the source and target domains, respectively. The unweighted average recall (UAR) rate is used to measure the performance of the system. The UAR is calculated as follows:

$$\mathrm{UAR} = \frac{\sum_{i=1}^{N_c} \mathrm{Recall}_i}{N_c}, \tag{16}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{17}$$

where N_c denotes the number of classes. "TP" indicates that the true value is positive, and the model judges it to be positive. "FN" indicates that the true value is positive, but the model judges it to be negative. Recall refers to the original sample, which indicates the probability of being predicted as a positive sample in the actual positive sample. Recall_i represents the recall rate for each type of data sample.

The experimental results indicate that different data distributions have different optimal feature dimensions. The feature vector dimensions affect the recognition performance of the model. The performance of the DANN model changed with the size of the feature dimensions. When the RLTD was used as the source and target domain datasets, the model achieved optimal performance when the feature vector dimension was set to 150. When SULD was used for the source and target domain datasets, the model achieved optimal performance when the feature vector dimension size was set to 200. Interestingly, when RLTD was used as the source domain dataset, and SULD was used as the target domain dataset, the model achieved optimal performance when the feature vector dimension size was set to 300. When SULD was used as the source domain dataset, and RLTD was used as the target domain dataset, the model achieved optimal performance when the feature vector dimension size was set to 250.

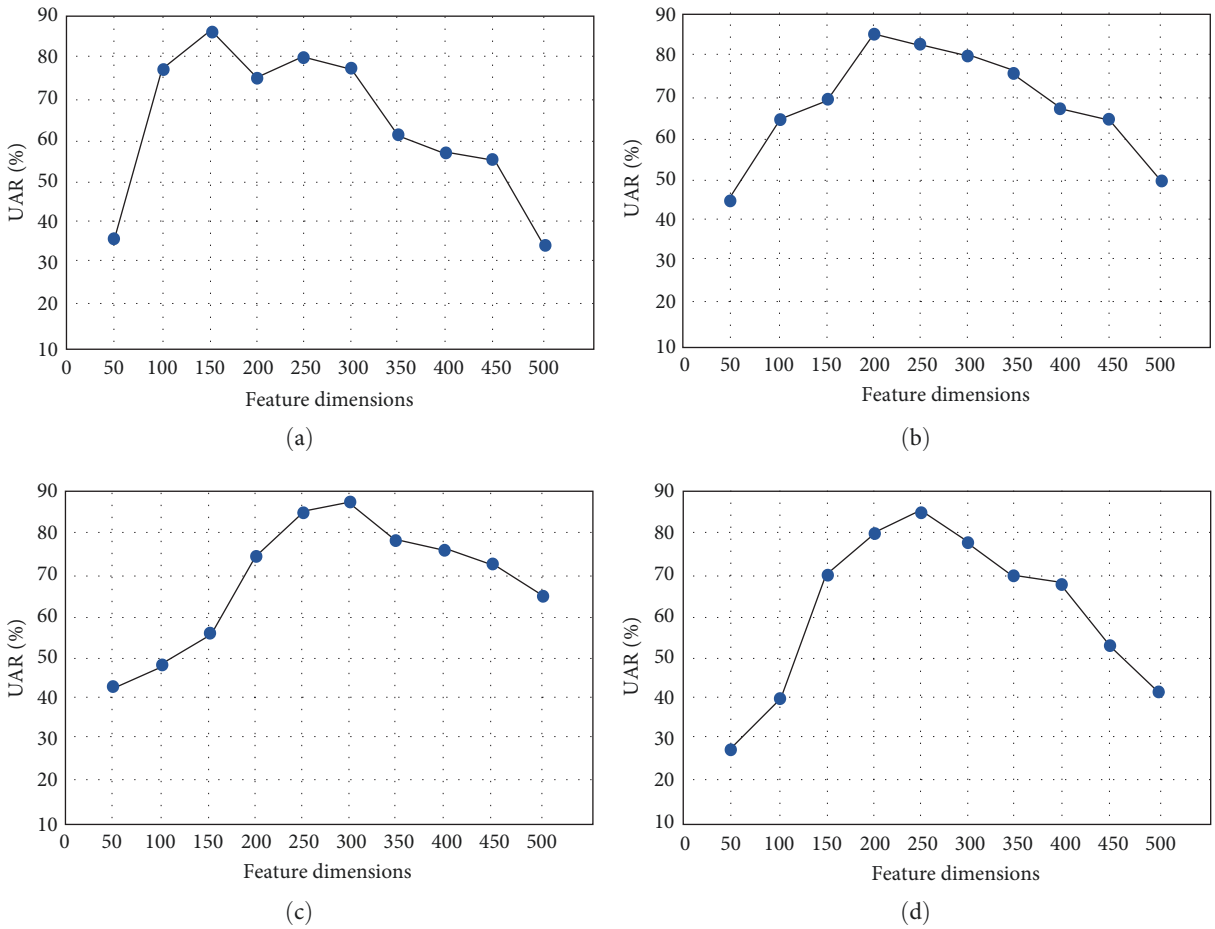

Figure 5: Detection performance of DANN with different feature dimensions. (a) RLTD is used for SD and TD. (b) SULD is used for SD and TD. (c) RLTD is used for SD and SULD is used for TD. (d) SULD is used for SD and RLTD is used for TD.

The results showed that when the source and target domain samples were consistent, the required feature dimensions were low. However, when the source and target domain samples were inconsistent, the required feature dimensions were high. Moreover, these results indicate that neither high nor low feature vector dimensions can achieve optimal performance for the model. Therefore, a reasonable feature vector dimension achieves optimal performance.

4.3.2. Analysis of Feature Fusion. The purpose of this experiment was to determine the differences between features extracted from lying and nonlying samples using the proposed lie detection method based on a DANN. The t distribution stochastic neighbor embedding algorithm was used to display the visualized features. This method was proposed by van der Maaten and Hinton. In this experiment, RLTD was used as the source domain dataset and SULD was used as the target domain dataset. All samples were labeled. The feature dimensions of the last fully connected layer of the deep CNN were reduced to two and represented in the form of a scatter plot. Figure 6 shows the visualization-fused feature results for lying and nonlying samples.

As shown in Figure 6(a), for the nonlying samples, the feature distributions of the source and target domain data samples were almost indistinguishable. This indicates that the features learned by the model do not contain any lying-specific information. However, the feature vector distribution shown in Figure 6(b) exhibits an evident variance. Owing to the better generalization of the DANN, the features of the lying samples can be clearly distinguished in the feature space, which contains the lying information. This experiment demonstrates that the DANN model can effectively extract the lying feature vectors.

4.3.3. Analysis of Feature Performance. A lie detection model based on DANN was developed. In this experiment, the goal was to analyze the performance of lie detection systems based on features from the different modalities while ensuring their independence from both the source and target domains. Both the source and target domain samples were set to 100. The samples were preprocessed with zero mean and unit variance normalization. To evaluate the improvement in the lie detection model performance by feature combinations of different modalities, experimental comparisons were conducted using single and bimodal feature combinations. The feature vectors for each sample in the speech and facial expression modalities were set as the best-dimensional vectors. The feature fusion algorithm used in this study adopts the method described in Section 3.2.

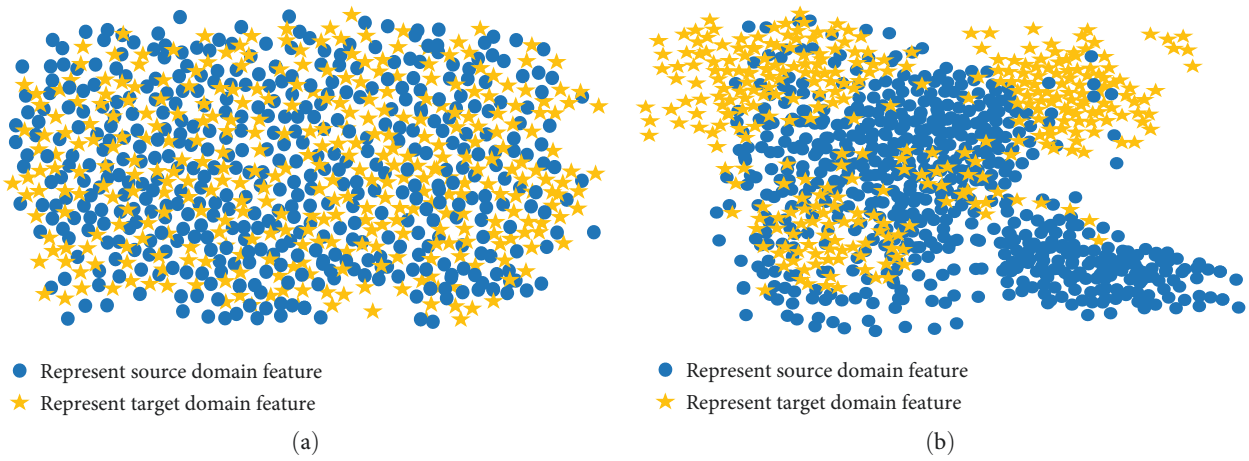

Figure 6: Visualized features. (a) Visualized fuzed features of nonlying samples. (b) Visualized fuzed features of lying samples.

Table 2: Comparison of lie detection accuracy based on different features.

Features	Accuracy (%) (Source domain dataset)	Accuracy (%) (Target domain dataset)
Facial expression features	58.3	53.2
Speech features	72.5	68.4
Feature fusion based on Hadamard product method	82.1	80.3
Feature fusion based on MHSA mechanism method	88.7	85.5

The experiment was conducted using a tenfold cross-validation method. In this model, nine–tenths of the dataset were used as the training sample, and one-tenth of the sample was used as the test sample. Ten experiments were conducted, and the average of the 10 experimental results was considered as the final result. The Hadamard product feature fusion method was used for comparison. Accuracy is defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{18}$$

where "FP" indicates that the true value is negative, but the model determines it to be positive; "TN" indicates that the true value is negative, and the model determines it to be negative.

A comparison of the results is presented in Table 2. As shown in the table, in contrast to the unimodal feature, the proposed bimodal lie detection model has a relatively high-classification accuracy.

The test results presented in the table show that the bimodal model employed in this study has a significantly improved detection accuracy when compared to the single-mode model. For the source domain dataset, the testing accuracy of the lie detection model using the Hadamard product fusion of bimodal features reached 88.7%, which was 30.4% higher than that of the single-mode model based on facial expression features and 16.2% higher than that of the single-mode model based on speech features. For the target domain dataset, the accuracy of bimodal combination detection reached approximately 85.5%, which was 32.3% higher than that of the single-mode model based on facial expression features and 17.1% higher than that of the single-mode model based on speech features. The test results show that the bimodal feature combination significantly improves the performance of the lie detection model and achieves higher detection accuracy.

4.3.4. Ablation Experiments. To verify the effectiveness of each submodule in the proposed method, ablation experiments were conducted using various datasets. The datasets included male trial videos from RLTD, female trial videos from RLTD, and male and female videos from SURD. Four modules were used in the experiments. The details are as follows:

(1) Basic DANN: This submodule uses one-dimensional convolution and speech modal features.

(2) A multimodal fusion module was added to the DANN. A feature fusion module was added to the basic DANN. Two convolutional networks were used to extract speech and facial expression features. This submodule is multimodal fusion DANN (MF-DANN).

(3) A multihead attention module was added to the DANN. An attention mechanism was added to the DANN. Subsequently, a multihead attention mechanism based on DANN (MHSA DANN) was established. This submodule considers only the speech modal features of the data.

(4) Final model: This is the final model proposed in this study. It considers the multimodal features of speech and facial expressions based on a DANN, while

Table 3: Detection accuracy of the ablation experiment (%).

Methods	Dataset			
	Male trial videos in RLTD	Female trial videos in RLTD	Male videos in SULD	Female videos in SULD
DANN	73.4	75.2	70.3	72.7
MF-DANN	80.6	79.3	78.4	79.8
MHSA-DANN	81.5	83.4	79.3	78.4
MF-MHSA-DANN	87.7	88.5	88.9	86.2

incorporating an attention module. This final model is a multimodal multihead attention mechanism DANN (MF-MHSA-DANN).

The experimental results for the four models are listed in Table 3. From the table, it is observed that the model with the addition of the multimodal fusion module improved the accuracy index for all four datasets compared with the basic DANN model. This indicates that the multimodal fusion module can effectively extract features from time-series signals and provide more effective information for the lie detection tasks than a single modality. The model with the added multihead attention mechanism exhibited significant improvements on different datasets. This indicates that it can effectively extract the temporal and spatial dependencies of time-series signals. The final model exhibited a significant improvement in accuracy compared to using the multimodal fusion module or the multihead attention module alone. The ablation experiments prove that adding the submodules of multimodal feature fusion and the multihead attention mechanism can help to learn the distribution characteristics of the data and improve the performance of the lie detection model.

4.3.5. Comparison of Different Detection Models. This experiment aimed to validate the performance of the proposed domain adversarial transfer CNN for lie detection and demonstrate its superiority. RLTD was used in this study. The area under the curve (AUC) was used as an objective indicator.

This study considers multimodal feature-based methods as a baseline. All experiments were evaluated using the same setup and dataset as those used in our model. The proposed method was compared with three neural network models [37–39] using the same experimental setup as that used in our method. These methods use features extracted from videos and multimodal features, including speech and facial expression features, as described. In [37], the GhostNet model was used to extract recognition features, combined with the design ideas of island and circle loss functions, and a loss function was designed and adopted based on cosine similarity to guide the learning of neural networks. In [38], deep separable convolution (DSConv) was used to design a lightweight fully CNN, including Part 1, which is a parallel convolutional structure comprising three parallel convolutional layers, and Part 2, which adopts the residual structure concept. The main edge contained two convolutional layers, each with 64 convolutional layers of size 3×3. The convolutional kernel of three comprises four consecutive convolutional layers, each with a kernel size of 3×3. The numbers of cores were 128, 160, 256, and 300, respectively.

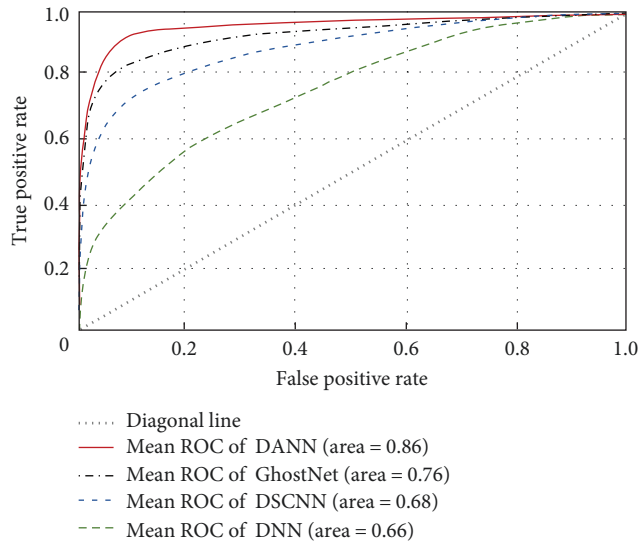

Figure 7: Average accuracies under different recognition modes.

In [39], a deep pretrained neural network was constructed to measure the human ability to detect deceptive utterances. This model was used to obtain text representations using a universal sentence encoder. The model performs better for typed utterances than for the spoken utterances.

The receiver operating characteristic (ROC) curves of the different models are shown in Figure 7.

The experimental results showed strong performance across all models. Particularly, the AUC values of the DANN model used in this study surpassed those of the DNN, DSCNN, and GhostNet network models by margins of 0.2, 0.18, and 0.1, respectively. The ROC curve in Figure 7 also shows that the bimodal feature combination of the DANN model significantly improved the model performance.

Moreover, to compare the performances of the various models under different data distributions, experiments were conducted using different databases, specifically the RLTD and SULD databases. The performance metric employed was accuracy. DNN, GhostNet, and DSCNN models were selected for comparison. The experiment was performed 10 times to obtain the average value and reduce the impact of random errors during network initialization. In the first test, the sample was a mixture of the RLTD and SULD data. In the second test, data were obtained only from the RLTD, and in the final test, data were obtained from the SULD.

As shown in the detection accuracy results in Figure 8, the DANN model has superior lie detection ability compared

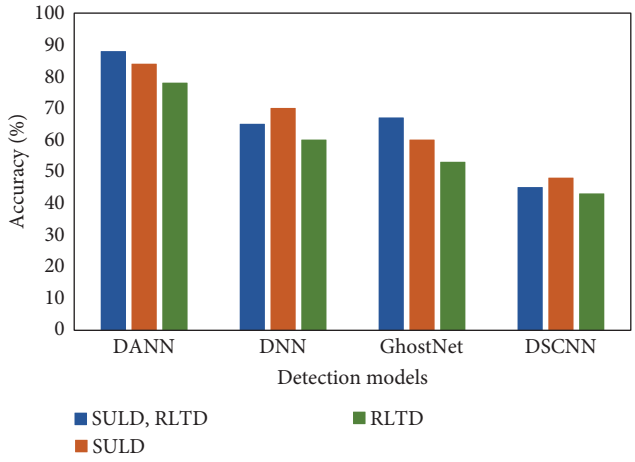

Figure 8: Average accuracies under different recognition models.

to the other three models, particularly when the database of the sample is inconsistent. In the case of mixed test data of RLTD and SULD, where the training and testing data exhibited different data distributions, the accuracy values achieved by the DANN model surpassed those of the other three models. Specifically, these values were 23%, 21%, and 43% higher than those of the other three models, respectively. This can be attributed to the DANN model's capacity to learn features devoid of domain class information. The model's parameters were updated and optimized using the joint objective functions of these tasks. Consequently, the shared feature vectors learned by the model have the characteristics of discriminability, generalization, and domain-class independence. In conclusion, the proposed method, based on bimodal feature fusion and a DANN, is significantly superior and more suitable for lie detection, particularly in scenarios involving inconsistent data distributions.

5. Summary and Outlook

The mismatch problem of a pretrained deep-learning model when the lie training data and test data originate from different data distributions was addressed. The proposed lie detection model, based on DANNs, can learn common feature vectors from both source and target domain data. The proposed method was validated using open-source datasets and a high correct lie detection rate was achieved. The main conclusions of this study are as follows:

First, the use of a DANN to construct feature extractors not only improves the detection accuracy in the source domain, but also significantly improves detection accuracy in the target domain. The experimental results indicated that DANNs can extract invariant features from lying samples and provide strong support for subsequent lie detection tasks. Second, a bimodal lie detection model was developed to detect lies by fuzing speech and facial expression features. The experimental results indicate that integrating different modal features to detect lies can significantly improve detection performance and achieve high accuracy. Third, the DANN detection model resolves the impact of inconsistent data distribution on the performance of machine learning models, ensuring robust performance across different scenarios.

However, the proposed method has limitations similar to those of the other data-sensitive detection models, such as regression problems and the sacrifice of discriminant features. Future research on lie detection should focus on searching for more effective detection features and expanding the database of liar scenarios to improve detection accuracy and the overall generalization ability of the detection model. This direction represents planned future research efforts in the field of lie detection.

Data Availability

Data for this research article are available upon corresponding author's request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Natural Science Foundation of Jiangsu Higher Education Institutions of China (grant 21KJB510022), the Seventh Batch of Science and Technology Development Plan (Agriculture) Project of Suzhou (SNG2023007), the Youth Natural Science Foundation of Jiangsu Province of China (grant BK20160361), and the Research Project on Higher Education Teaching Reform in Jiangsu Province (grant 2021JSJG176). The authors acknowledge the Intelligent Computing and Knowledge Learning Research Platform Construction Project of Suzhou Vocational University, 3C-Product Intelligent Manufacturing Engineering Technology Research and Development Center of Jiangsu Province, and QingLan Project of Colleges and Universities in Jiangsu Province.

References

[1] Y. Zhi and W. Hong, "Research and analysis of speech lie detection technology," *Guangdong Public Security Technology*, vol. 29, pp. 48–50, 2021.

[2] Z. Li, L. Ruiyu, X. Yue, and Z. Dongze, "Research status and prospects of speech lie detection technology," *Data Collection and Processing*, vol. 32, pp. 246–257, 2017.

[3] A. Derakhshan, M. Mikaeili, T. Gedeon, and A. M. Nasrabadi, "Identifying the optimal features in multimodal deception detection," *Multimodal Technologies and Interaction*, vol. 4, no. 2, Article ID 25, 2020.

[4] Y. Fang, H. Fu, H. Tao, R. Liang, and L. Zhao, "A novel hybrid network model based on attentional multi-feature fusion for deception detection," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E104.A, no. 3, pp. 622–626, 2021.

[5] D. P. Jayathunga, R. M. I. S. Ranasinghe, and R. Murugiah, "A comparative study of supervised machine learning techniques for deceptive review identification using linguistic inquiry and word count," in *Computational Intelligence in Information Systems. CIIS 2021*, W. S. H. Suhaili, N. Z. Siau, S. Omar, and S. Phon-Amuaisuk, Eds., vol. 1321 of *Advances*

in Intelligent Systems and Computing, pp. 97–105, Springer, Cham, 2021.

[6] P. Taylor, N. Griffiths, V. Hall, Z. Xu, and A. Mouzakitis, "Feature selection for supervised learning and compression," *Applied Artificial Intelligence*, vol. 36, no. 1, Article ID 2034293, 2022.

[7] Y. Zhou and F. Bu, "An overview of advancements in lie detection technology in speech," *International Journal of Information Technologies and Systems Approach*, vol. 16, no. 2, pp. 1–24, 2023.

[8] J. F. George, D. P. Biros, M. Adkins, and J. K. Burgoon, "Testing various modes of computer-based training for deception detection," in *Intelligence and Security Informatics. ISI 2004*, H. Chen, R. Moore, D. D. Zeng, and J. Leavitt, Eds., vol. 3073 of *Lecture Notes in Computer Science*, pp. 411–417, Springer, Berlin, Heidelberg, 2004.

[9] E. Elliott and A.-M. Leach, "You must be lying because I don't understand you: language proficiency and lie detection," *Journal of Experimental Psychology: Applied*, vol. 22, no. 4, pp. 488–499, 2016.

[10] J. Dai, L. Sun, and X. Shen, "Research on speech spoofing detection based on big data and machine learning," in *2021 2nd International Conference on Artificial Intelligence and Education (ICAIE)*, pp. 137–140, IEEE, Dali, China, June 2021.

[11] Z. Ren and H. Ning, "A review of microexpression recognition research," *Computer Engineering and Application*, vol. 57, pp. 38–47, 2021.

[12] D. A. Curtis, "Deception detection and emotion recognition: investigating F.A.C.E. software," *Psychotherapy Research*, vol. 31, no. 6, pp. 802–816, 2021.

[13] H. Karimi, J. Tang, and Y. Li, "Toward end-to-end deception detection in videos," in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 1278–1283, IEEE, Seattle, WA, USA, December 2018.

[14] E. P. Lloyd, K. Hugenberg, A. R. McConnell, J. W. Kunstman, and J. C. Deska, "Black and white lies: race-based biases in deception judgments," *Psychological Science*, vol. 28, no. 8, pp. 1125–1136, 2017.

[15] R. W. Picard, *Affective Computing, Linlin, L., Translated by*, Beijing University Press of Technology Press, 2005.

[16] S. Kumar, C. Bai, V. S. Subrahmanian, and J. Leskovec, "Deception detection in group video conversations using dynamic interaction networks," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, pp. 339–350, 2021.

[17] X. Li, C. Deng, Q. Wu, R. Cui, J. Tang, and Y. Zhang, "Research on polygraph technology based on ballistocardiogram signal," in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, pp. 92–97, IEEE, Chongqing, China, June 2020.

[18] G. Krishnamurthy, N. Majumder, S. Poria, and E. Cambria, "A deep learning approach for multimodal deception detection," in *Computational Linguistics and Intelligent Text Processing. CICLing 2018*, A. Gelbukh, Ed., vol. 13396 of *Lecture Notes in Computer Science*, pp. 87–96, Springer, Cham, 2023.

[19] X. Li, *Research on Lie Detection Technology Based on Speech and Radar Dual Sensors*, Nanjing University of Science and Technology, 2021.

[20] L. Mathur and M. J. Matarić, "Affect-aware deep belief network representations for multimodal unsupervised deception detection," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pp. 1–8, IEEE, December 2021.

[21] H. Fu and P. Lei, "Speech deception detection algorithm based on denoising auto-encoder and long short-term memory network," *Journal of Computer Applications*, vol. 40, pp. 589–594, 2020.

[22] N. Srivastava and S. Dubey, "Deception detection using artificial neural network and support vector machine," in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 1205–1208, IEEE, Coimbatore, India, March 2018.

[23] Y. Xie, R. Liang, and Y. Bao, "Deception detection with spectral features based on deep belief network," *Journal of Acoustics*, vol. 44, pp. 214–220, 2019.

[24] W. Jiangping, L. Jiajun, and C. Ning, "Multi feature noncontact lie detection technology," *Jounal of East China University of Science Technology*, vol. 46, pp. 556–563, 2020.

[25] C. Yadong, W. Huapeng, L. En, N. Lingge, and L. Yuanzhou, "Speech pressure lie detection based on speech sentiment analysis system," *Criminal Technology*, vol. 45, pp. 155–159, 2020.

[26] W. Yanxin, Y. Jing, W. Jianhua, G. Yingsan, and L. Zhiyuan, "Intelligent diagnosis method for insulation defects in small sample GIS based on domain adversarial transfer convolutional neural network," *Journal of Electrical Engineering-Elektrotechnicky Casopis*, vol. 37, pp. 2150–2160, 2022.

[27] X. Qiang, L. Baoguo, W. Xiang, D. Wen, and D. Zhiyi, "Modulation recognition algorithm based on multimodal domain adversarial neural networks," *Aerosp Electron Countermeasures*, vol. 37, pp. 32–37, 2021.

[28] S. Yunhao, X. Hua, and D. Junjie, "Modulation classification method based on domain adaptive neural networks," *Journal of Air Force Engineering University (Natural Science Edition)*, vol. 21, pp. 69–75, 2020.

[29] Zhihao, *Research on single channel speech enhancement method based on prior information at different semantic levels*, Ph.D. Dissertation, Harbin Institute of Technology, 2020.

[30] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 59–66, Association for Computing Machinery, November 2015.

[31] Y. Zhou, *Research on Lie Detection Based on Speech Sparse Representation*, Soochow University, China, 2017.

[32] J. Wei, J. Lin, and N. Chen, "Multi feature non-contact lie detection technology," *Journal of East China University of Science and Technology: Natural Science Edition*, vol. 46, no. 4, pp. 556–563, 2020.

[33] Y. Ganin, E. Ustinova, H. Ajakan et al., "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[34] A. Hannun, A. Lee, Q. Xu, and R. Collobert, "Sequence-to-sequence speech recognition with time-depth separable convolutions," 2019.

[35] G. Yang, J. Yang, Z. Lu, and D. Liu, "A convolutional neural network with sparse representation," *Knowledge-Based Systems*, vol. 209, Article ID 106419, 2020.

[36] R. Rubinstein, T. Faktor, and M. Elad, "K-SVD dictionary-learning for the analysis sparse model," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5405–5408, IEEE, Kyoto, Japan, March 2012.

[37] W. Fang, Z. Zhang, and H. Wang, "Multimodal emotion recognition integrating speech, EEG, and facial expressions," *Computer System Applications*, vol. 32, no. 1, pp. 337–347, 2023.

[38] A. Karatzoglou, N. Schnell, and M. Beigl, "Applying depthwise separable and multi-channel convolutional neural networks of varied kernel size on semantic trajectories," *Neural Computing and Applications*, vol. 32, no. 11, pp. 6685–6698, 2020.
[39] A. Wawer and J. Sarzyńska-Wawer, "Detecting deceptive utterances using deep pre-trained neural networks," *Applied Sciences*, vol. 12, no. 12, Article ID 5878, 2022.