Hindawi
International Journal of Intelligent Systems
Volume 2024, Article ID 7393431, 20 pages
https://doi.org/10.1155/2024/7393431

Research Article

Kernel Probabilistic Dependent-Independent Canonical Correlation Analysis

Reza Rohani Sarvestani,[1] **Ali Gholami**,[2] **and Reza Boostani**[3]

[1]*Department of Computer Engineering, Shahrekord University, Shahrekord, Iran*
[2]*Department of Electrical Engineering, Faculty of Technology and Engineering, Tehran Branch, Islamic Azad University, Tehran, Iran*
[3]*CSE and IT Department, ECE Faculty, Shiraz University, Shiraz, Iran*

Correspondence should be addressed to Reza Rohani Sarvestani; rrohani.cse@gmail.com

Received 11 August 2023; Revised 31 October 2023; Accepted 6 December 2023; Published 3 January 2024

Academic Editor: Mohammad R. Khosravi

There is growing interest in developing linear/nonlinear feature fusion methods that fuse the elicited features from two different sources of information for achieving a higher recognition rate. In this regard, canonical correlation analysis (CCA), cross-modal factor analysis, and probabilistic CCA (PCCA) have been introduced to better deal with data variability and uncertainty. In our previous research, we formerly developed the kernel version of PCCA (KPCCA) to capture both nonlinear and probabilistic relation between the features of two different source signals. However, KPCCA is only able to estimate latent variables, which are statistically correlated between the features of two independent modalities. To overcome this drawback, we propose a kernel version of the probabilistic dependent-independent CCA (PDICCA) method to capture the nonlinear relation between both dependent and independent latent variables. We have compared the proposed method to PDICCA, CCA, KCCA, cross-modal factor analysis (CFA), and kernel CFA methods over the eNTERFACE and RML datasets for audio-visual emotion recognition and the M2VTS dataset for audio-visual speech recognition. Empirical results on the three datasets indicate the superiority of both the PDICCA and Kernel PDICCA methods to their counterparts.

1. Introduction

It is evident that data collection from a single sensor (modality) does not capture all discriminative information from an observation. For instance, when we take a film of a person during his/her speech, voice signals or video frames solely cannot capture all the information in the input states. Therefore multisource data collection has attracted the researchers' attention because various data from an observation can cover the uncertainty, variability, and partial observation of each other. For instance, when we listen to somebody and watch him/her simultaneously, even if we cannot hear a word well, we can guess the corresponding word via observing his/her lip motions. Thus, in this research, we want to fuse audio and visual information to achieve a higher speech recognition rate. Although a few fusion techniques have been developed to fuse the elicited correlated features of different modalities, this research proposes a novel approach for fusing both dependent and independent probabilistic information of feature vectors extracted from two different modalities.

1.1. Background. It is empirically shown that the raw audio and video data passes through several neural processing stages [1], which are nonlinear and eventually integrated into each other via a complex neural process. Bimodal data collection is repeatedly used in several applications such as medical image fusion [2], multimodal interaction [3], multimodal emotion detection [4], and audio-visual speech recognition [5, 6].

Information fusion can be carried out at different levels of data processing including raw data fusion, feature fusion, model fusion, and decision fusion. Among these approaches, feature fusion [7, 8] is of concern in this study because this approach can consider the linear/nonlinear relation among elicited features. It is noteworthy to say that feature fusion is different from feature concatenation [9]. In other words, extracting feature vectors from different sources of data and arranging them into a long feature vector is not feature fusion. Feature concatenation methods highly affect the number of training parameters in several classifiers such as Bayes classifier [10] or deep learning schemes [11, 12]. In the case of a small sample size problem, the covariance of such a small dataset is underestimated. From another perspective, conventional classifiers are unable to either capture the interaction of all features or tolerate the uncertainty and variability of features [6, 13–15].

Therefore, an efficient feature fusion should not produce very high-dimensional feature vectors. To mimic the human's fusion system, representative features from each modality are elicited and then projected into a new space (processing spikes in a higher level) in a way that the projected features of these modalities have a maximum correlation or have a minimum distance, while having a proper size. Therefore, by fusing the projected features (e.g., audio and video features) in the correlation space, better recognition performance can be obtained [4].

Canonical correlation analysis (CCA) [16] is a known technique for future fusion by identifying the shared (dependent) information between two different sources of data. CCA is determined by optimizing two different linear projections of features belonging to two different modalities. These features are projected into a new space (called correlation space), in which the cross-correlation of the two projected features is maximized. These projected features are called latent variables/features. Nonetheless, CCA has some drawbacks such as a lack of understanding of the stochastic nature of the features. Moreover, CCA is unable to capture nonlinear relations between features. Cross-modal factor analysis (CFA) [17] is another feature fusion method that is similar to CCA, projects the input features of two different modalities into a new space in such a way that the distance (Frobenius norm) between the projected features is minimized. CCA and CFA have been adopted in practical multimodal recognition systems such as face recognition [18], signal processing [19], monitoring and fault detection [20], audio-visual speaker detection [21], fusion of multimodal medical imaging [22, 23], and audio-visual synchronization [24]. To enable both CCA and CFA to capture the input nonlinearities, their kernel versions, called KCCA [25] and KCFA [26] were developed. They have been applied to various data fusion applications like specific radar emitter identification [27],audio-visual emotion recognition [26, 28], and feature selection [20, 29]. Nevertheless, during the recording of audio-video signals, a few undesired disturbing factors occur such as the slight movement of the recording camera or getting close and far from the recording microphone. For instance, in a bimodal recognition system [26], the KCFA scheme is deployed to elicit the latent variables of audio and video data but the achieved results are not convincing. This is because ignoring the variability factors during the recording affects the quality of the recorded data and declines the performance of the recognizer.

In practice, each set of recorded data has a degree of randomness due to several reasons, such as the movement of sources during data acquisition, power line noise, and the induction noise of other equipment. To capture the stochastic nature as well as the variability of recorded data from two modalities, Bach and Jorden [30] have proposed the probabilistic CCA (PCCA) model. Although PCCA is linear, we propose its kernel version in our previous study to capture the nonlinearity of elicited features from two modalities [6].

Although correlated features can cover the lack of each other, independent features do not suffer from redundancy and reveal a new perspective from an input observation. It is interesting that some fusion methods just estimate the dependent (shared) features between two modalities of data while a few of them use both dependent and independent features. Thus, employing both dependent and independent features in the PCCA framework, which is termed as PDICCA in the literature [31], allows us to move from partial description toward a wider observability of inputs.

Since PDICCA is a linear method and cannot capture nonlinear relations between the elicited features, the main contribution of this study is devoted to kernelizing the PDICCA method, which we call KPDICCA hereafter. The proposed method is able to capture different aspects of inputs such as dependencies, independencies, uncertainty, variability, and nonlinearities. As we see, in the case of encountering a limited number of samples, kernel methods are able to provide convincing results because the size of the kernel depends on the number of feature vectors. In contrast, to get a convincing result from a classifier, the number of training samples should be high. Hence, there is a trade-off between applying a kernel to input features and well training a classifier.

The rest of this paper is structured as follows: In Section 2, the PDICCA method and the proposed method are introduced. Section 3 introduces the deployed datasets and their feature extraction techniques. Section 4 presents the experimental results obtained from the proposed method along with state-of-the-art methods, and their achievements are compared and discussed. Finally, the paper is concluded in Section 5.

2. Methodology

In this section, first PDICCA is briefly explained, and then the proposed method (KPDICCA) is introduced.

2.1. PDICCA. To overcome the lack of capturing uncertainty in both CCA and KCCA models optimized by the maximum likelihood (ML) method [30], one solution is to consider a linear projection between observations of sources and dependencies among latent variables. In addition to the Gaussian distribution assumption, the probabilistic CCA (PCCA) method is able to model the variability of data and outlines a solution for the CCA problem. PCCA has been

extended [31] by incorporating a dependent variable Z similar to CCA and two other independent latent variables Z_x and Z_y, which are not dependent and exclusively belonged to the two modalities of x and y, as described in Figure 1. This method is termed probabilistic dependent-independent CCA (PDICCA) categorizing as a generative data model. PDICCA captures both dependence and independence of latent variables as follows:

$$\begin{aligned}\mathbf{x} &= \mathbf{f}(\mathbf{z}|\mathbf{W_x}) + \mathbf{g}(\mathbf{z_x}|\mathbf{B_x}) + \epsilon_{\mathbf{x}},\\ \mathbf{y} &= \mathbf{f}(\mathbf{z}|\mathbf{W_y}) + \mathbf{g}(\mathbf{z_y}|\mathbf{B_y}) + \epsilon_{\mathbf{y}},\end{aligned} \tag{1}$$

where $\mathbf{f}(.)$ and $\mathbf{g}(.)$ are two deterministic functions that transform dependent latent variables ($\mathbf{z}$) and independent latent variables ($\mathbf{z_x}, \mathbf{z_y}$) to the observation space. However, $\epsilon_{\mathbf{x}}$ and $\epsilon_{\mathbf{y}}$ denote the additive noise on $\mathbf{x}$ and $\mathbf{y}$ observations, respectively.

For the simplification, they use two linear functions, $\mathbf{f}(\mathbf{z}|\mathbf{W_i}) = \mathbf{z}\mathbf{W_i^T}$ and $\mathbf{g}(\mathbf{z_i}|\mathbf{B_i}) = \mathbf{z_i}\mathbf{B_i^T}$, and they consider an independent Gaussian distribution with equal variance for noise parameters $\epsilon_{\mathbf{i}} = \mathcal{N}(0, \sigma_i^2 I)$. Furthermore, they assume a Gaussian distribution with zero mean and unit covariance for latent variables $(\mathbf{z}, \mathbf{z_x}, \mathbf{z_y})$.

Therefore, we can write

$$\begin{aligned}\mathbf{z_x}, \mathbf{z_y}, \mathbf{z} &\sim \mathcal{N}(0, I_d),\\ \mathbf{x}|\mathbf{z}, \mathbf{z_x} &= \mathcal{N}\left(\mathbf{z}\mathbf{W_x^T} + \mathbf{z_x}\mathbf{B_x^T}, \sigma_x^2 I\right),\\ \mathbf{y}|\mathbf{z}, \mathbf{z_y} &= \mathcal{N}\left(\mathbf{z}\mathbf{W_y^T} + \mathbf{z_y}\mathbf{B_y^T}, \sigma_y^2 I\right).\end{aligned} \tag{2}$$

To solve the above equation, at first, the parameters $\mathbf{z}$ (shared latent variable), $\mathbf{W_x}$, and $\mathbf{W_y}$ (projection matrices), must be estimated while the set-specific parameters $\mathbf{z_x}$ and $\mathbf{z_y}$ should be marginally out. Afterward, the probabilistic model is marginalized over the shared latent variable $\mathbf{z}$ and then $\mathbf{B_i}$ and $\boldsymbol{\sigma}_\mathbf{i}(\mathbf{i} = \mathbf{x}$ or $\mathbf{y})$ can be optimized, accordingly. To summarize this learning scheme, its pseudo code is illustrated as follows (see Algorithm 1).

CCA, CFA, and PDICCA methods are all linear approaches, which are capable of finding linear relationship between two synchronous recording modalities. It should be noted that these models cannot digest the nonlinear correlations between two sets of features elicited from two different modalities. Herein, a nonlinear kernel is inserted into PDICCA to overcome this drawback.

2.2. The Proposed Method. Our approach is similar to KCCA [25] and KCFA [26]. To the best of the authors' knowledge, deriving the kernel version of PDICCA has not been proposed yet. This paper aims to propose the kernel version of PDICCA (KPDICCA) for considering the nonlinear relations among the observations (from audio and visual modalities). To equip the PDICCA method for

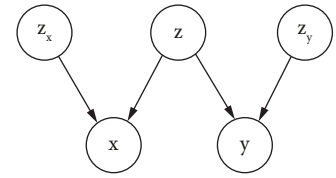

Figure 1: Graphical representation of the generative model structure used to detect dependencies and independencies (Klami and Kaski [31]).

capturing nonlinear relations between the observations and their elicited latent variables, here the kernel derivation of PDICCA is developed by implicitly mapping data from the original space to a higher dimension and then apply the Klami and Kaski method [31] to find dependent and independent latent variables, as shown in Figure 2. To derive the formula, first, we consider that all latent variables have normal distribution. Similar to the derivation of KCCA, we can write

$$\begin{aligned}\mathbf{z_x}, \mathbf{z_y}, \mathbf{z} &\sim \mathcal{N}(0, I_d),\\ \phi(\mathbf{x})|\mathbf{z}, \mathbf{z_x} &= \mathcal{N}\left(\mathbf{z}\mathbf{W_x^T} + \mathbf{z_x}\mathbf{B_x^T}, \sigma_x^2 I\right),\\ \psi(\mathbf{y})|\mathbf{z}, \mathbf{z_y} &= \mathcal{N}\left(\mathbf{z}\mathbf{W_y^T} + \mathbf{z_y}\mathbf{B_y^T}, \sigma_y^2 I\right).\end{aligned} \tag{3}$$

To simplify the above relations, similar to KCCA, we assume that $\mathbf{W}_i$ and $\mathbf{B}_i$ ($i = x$ or y), which are the transformation matrices that project $\phi(\mathbf{x})$ and $\psi(\mathbf{y})$ into $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ subspaces, can be written as follows:

$$\begin{aligned}\mathbf{W}_x &= \phi(\mathbf{x})^T \boldsymbol{\alpha}_\mathbf{x},\\ \mathbf{W}_y &= \psi(\mathbf{y})^T \boldsymbol{\alpha}_\mathbf{y},\\ \mathbf{B}_x &= \phi(\mathbf{x})^T \boldsymbol{\beta}_\mathbf{x},\\ \mathbf{B}_y &= \psi(\mathbf{y})^T \boldsymbol{\beta}_\mathbf{y}.\end{aligned} \tag{4}$$

Considering $\mathbf{W}_i$ and $\mathbf{B}_i$ ($i = x$ or y) parameters, we can derive a learning method using the expectation maximization (EM) algorithm [31] to obtain parameters $\theta = \left\{\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\delta}_\mathbf{i}^\mathbf{2}\right\}$ where $i \in \{x, y\}$ and $\boldsymbol{\delta}_\mathbf{i}^\mathbf{2}$ is the noise variance of the primary domain. Assuming $\boldsymbol{\beta}_x$ and $\boldsymbol{\beta}_y$ are fixed and marginalizing over independent factors of z_x and z_y, constructing a probabilistic model that is only dependent to $\mathbf{W}_x$ and $\mathbf{W}_y$ with the following covariances:

$$\begin{aligned}\varphi_x &= \phi(\mathbf{x})^T\left(\boldsymbol{\beta}_\mathbf{x}\boldsymbol{\beta}_\mathbf{x}^\mathbf{T} + \boldsymbol{\delta}_\mathbf{x}^\mathbf{2}\mathbf{I}\right)\phi(\mathbf{x}),\\ \varphi_y &= \psi(\mathbf{y})^T\left(\boldsymbol{\beta}_\mathbf{y}\boldsymbol{\beta}_\mathbf{y}^\mathbf{T} + \boldsymbol{\delta}_\mathbf{y}^\mathbf{2}\mathbf{I}\right)\psi(\mathbf{y}).\end{aligned} \tag{5}$$

If we consider $\mathbf{W} = [\mathbf{W}_x \quad \mathbf{W}_y]^\mathbf{T} = \varpi(\mathbf{O})^T\left[\boldsymbol{\alpha}_\mathbf{x} \quad \boldsymbol{\alpha}_\mathbf{y}\right]^\mathbf{T} = \varpi(\mathbf{O})^T\boldsymbol{\alpha}$ and covariance matrix as

$$\varphi = \varpi(\mathbf{O})^T\begin{bmatrix}\left(\boldsymbol{\beta}_\mathbf{x}\boldsymbol{\beta}_\mathbf{x}^\mathbf{T} + \boldsymbol{\delta}_\mathbf{x}^\mathbf{2}\mathbf{I}\right) & 0\\ 0 & \left(\boldsymbol{\beta}_\mathbf{y}\boldsymbol{\beta}_\mathbf{y}^\mathbf{T} + \boldsymbol{\delta}_\mathbf{y}^\mathbf{2}\mathbf{I}\right)\end{bmatrix}\varpi(\mathbf{O}) = \varpi(\mathbf{O})^T\Omega\,\varpi(\mathbf{O}), \tag{6}$$

(1) Assume that $\mathbf{B_x}$ and $\mathbf{B_y}$ are fixed and marginalized over $\mathbf{z_x}$ and $\mathbf{z_y}$ to get
$\boldsymbol{\varphi_x} = \mathbf{B_x}\mathbf{B_x^T} + \boldsymbol{\sigma_x^2}\mathbf{I}$ and $\boldsymbol{\varphi_y} = \mathbf{B_y}\mathbf{B_y^T} + \boldsymbol{\sigma_y^2}\mathbf{I}$
(i) Update the parameters $\mathbf{W} = [\mathbf{W}_x; \mathbf{W}_y]$ using
$$\mathbf{W} = \boldsymbol{\Sigma}\mathbf{A^T}(\mathbf{M} + \mathbf{A}\boldsymbol{\Sigma}\mathbf{A^T})^{-1}$$
Here $\mathbf{M} = (\mathbf{I} + \mathbf{W^T}\boldsymbol{\varphi}^{-1}\mathbf{W})^{-1}$, $\mathbf{A} = \mathbf{M}\mathbf{W^T}\boldsymbol{\varphi}^{-1}$, and φ is a block-diagonal matrix that consist of $\boldsymbol{\varphi_x}$ and $\boldsymbol{\varphi_y}$. The $\boldsymbol{\Sigma}$ is the joint sample covariance matrix.
(2) Marginalize over $\mathbf{z}$ to get $\boldsymbol{\varphi_x} = \mathbf{W_x}\mathbf{W_x^T} + \boldsymbol{\sigma_x^2}\mathbf{I}$
(i) Update $\mathbf{B_x}$ with
$$\mathbf{B_x} = \boldsymbol{\Sigma_x}\mathbf{A_x^T}(\mathbf{M_x} + \mathbf{A_x}\boldsymbol{\Sigma_x}\mathbf{A_x^T})^{-1}$$
where $\mathbf{M_x} = (\mathbf{I} + \mathbf{B_x^T}\boldsymbol{\varphi_x}^{-1}\mathbf{B_x})^{-1}$ and $\mathbf{A_x} = \mathbf{M_x}\mathbf{B_x^T}\boldsymbol{\varphi_x}^{-1}$. And $\boldsymbol{\Sigma_x}$ is the sample covariance of x.
(ii) Update $\boldsymbol{\sigma_x^2}$ using
$\boldsymbol{\sigma_x^2} = \mathbf{1}/\mathbf{d_x}\,\mathrm{trace}(\boldsymbol{\Sigma_x} - \boldsymbol{\Sigma_x}\mathbf{A_x^T}B_x^T - \mathbf{W_x}\mathbf{W_x^T})$
where $\mathbf{d_x}$ is the dimensionality of x, and $\mathbf{B_x}$ is the new value just updated.
Repeat the above two substeps for parameters related to y, replacing all subscripts x with y.

Algorithm 1: PDICCA algorithm.

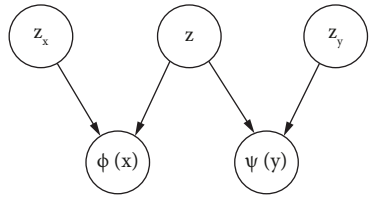

Figure 2: Graphical representation of the generative model structure used to detect dependencies and independence in the kernel domain.

where $\varpi(\mathbf{O}) = \begin{bmatrix} \phi(\mathbf{x}) & 0 \\ 0 & \psi(\mathbf{y}) \end{bmatrix}$, we can obtain updating formula for the parameter of $\mathbf{W}$ such as

$$\mathbf{W} = \Pi\mathbf{A^T}\left(\mathbf{M} + \mathbf{A}\boldsymbol{\Sigma}\mathbf{A^T}\right)^{-1}. \tag{7}$$

If $\mathbf{K}_x$ and $\mathbf{K}_y$ are invertible, we will have

$$\begin{aligned} \boldsymbol{\Pi} &= \varpi(\mathbf{O})^T\,\varpi(\mathbf{O}), \\ \mathbf{M} &= \left(I + \mathbf{W}^T\varphi^{-1}\mathbf{W}\right)^{-1} = \left(I + \boldsymbol{\alpha}^T\Omega^{-1}\boldsymbol{\alpha}\right)^{-1}, \end{aligned} \tag{8}$$

$$\begin{aligned} \mathbf{A} &= \mathbf{M}\boldsymbol{\alpha}^T\varpi(\mathbf{O})\left(\varpi(\mathbf{O})^T\Omega\,\varpi(\mathbf{O})\right)^{-1} \\ &= \mathbf{M}\boldsymbol{\alpha}^T\Omega^{-1}\varpi(\mathbf{O})^{T^{-1}}. \end{aligned} \tag{9}$$

Therefore, we can obtain the update formula for $\boldsymbol{\alpha}$ as follows:

$$\mathbf{W} = \varpi(\mathbf{O})^T\boldsymbol{\alpha} = \varpi(\mathbf{O})^T\,\Omega^{-1}\boldsymbol{\alpha}\mathbf{M}^T\left(\mathbf{M} + \left(\mathbf{M}\boldsymbol{\alpha}^T\Omega^{-1}\right)\left(\mathbf{M}\boldsymbol{\alpha}^T\Omega^{-1}\right)^{\mathbf{T}}\right)^{-1}, \tag{10}$$

$$\boldsymbol{\alpha} = \Omega^{-1}\boldsymbol{\alpha}\mathbf{M}^T\left(\mathbf{M} + \left(\mathbf{M}\boldsymbol{\alpha}^T\Omega^{-1}\right)\left(\mathbf{M}\boldsymbol{\alpha}^T\Omega^{-1}\right)^{\mathbf{T}}\right)^{-1}. \tag{11}$$

In the second step, we marginalize over z parameter and use similar method to provide an updating function for $\boldsymbol{\beta}_i$ as follows:

$$\boldsymbol{\beta}_i = \Lambda_i^{-1}\boldsymbol{\beta}_i\,\mathbf{N}_i^T\left(\mathbf{N_i} + \left(\mathbf{N_i}\,\boldsymbol{\beta}_i^T\Lambda_i^{-1}\right)\left(\mathbf{N_i}\,\beta_i^T\Lambda_i^{-1}\right)^{\mathbf{T}}\right)^{-1}, \tag{12}$$

where

$$\begin{aligned} \Lambda_i &= \boldsymbol{\alpha_i}\boldsymbol{\alpha_i^T} + \boldsymbol{\delta_i^2}\mathbf{I}, \\ \mathbf{N_i} &= \left(I + \boldsymbol{\beta}_i^T\Lambda_i^{-1}\boldsymbol{\beta}_i\right)^{-1}. \end{aligned} \tag{13}$$

Actually, the variance is recovered by the following equation:

$$\boldsymbol{\delta_i^2} = \frac{\mathbf{1}}{\mathbf{d_i}}\mathbf{trace}\left(\mathbf{I_i} - \Lambda_i^{-1}\boldsymbol{\beta}_i\mathbf{N}_i^T\boldsymbol{\beta}_i^T - \boldsymbol{\alpha_i}\boldsymbol{\alpha_i^T}\right). \tag{14}$$

The posterior expectation of shared and set-specific latent variables given observation x, latent variables can be obtained by ML estimation. Applying ML to the probabilistic model, we achieve

$$\begin{aligned} E(z|\mathbf{x}) &= \phi(\mathbf{x}).\left[\phi(\mathbf{x})^T\boldsymbol{\alpha_x}\right]M_x^T = \mathbf{K}_x\boldsymbol{\alpha_x}M_x^T, \\ E(z_x|\mathbf{x}) &= \phi(\mathbf{x}).\left[\phi(\mathbf{x})^T\boldsymbol{\beta_x}\right]N_x^T = \mathbf{K}_x\boldsymbol{\beta_x}N_x^T. \end{aligned} \tag{15}$$

Similarly, for the observation y, at the above equation, replacing all subscript x by y and ϕ by ψ. If $\mathbf{K}_x$ and $\mathbf{K}_y$ are not invertible, the above equations will not provide a solution for KPCCA. This problem can be solved using a similar

regularization approach that has been presented for KCCA. In this study, we use Golub et al. [32] and Koskinen et al. [33] methods that both consider a priori knowledge on $\mathbf{W}_i \sim \mathcal{N}(0, r^{-1}I_i)$ and $\mathbf{B}_i \sim \mathcal{N}(0, r^{-1}I_i)$, where r is a regularization parameter and apply the EM algorithm to achieve the following relations for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_i$:

$$\begin{aligned}\boldsymbol{\alpha} &= \Omega^{-1}\boldsymbol{\alpha}\mathbf{M}^T\left(\mathbf{M} + \left(\mathbf{M}\boldsymbol{\alpha}^T\Omega^{-1}\right)\left(\mathbf{M}\boldsymbol{\alpha}^T\Omega^{-1}\right)^{\mathbf{T}} + r\lambda_\varphi\lambda_{\mathbf{K}}\right)^{-1},\\ \boldsymbol{\beta}_i &= \Lambda_i^{-1}\boldsymbol{\beta}_i\mathbf{N}_i^T\left(\mathbf{N}_\mathbf{i} + \left(\mathbf{N}_\mathbf{i}\,\boldsymbol{\beta}_i^T\Lambda_i^{-1}\right)\left(\mathbf{N}_\mathbf{i}\boldsymbol{\beta}_i^T\Lambda_i^{-1}\right)^{\mathbf{T}} + r\lambda_{\Lambda_i}\lambda_{\mathbf{K}_\mathbf{i}}\right)^{-1},\end{aligned} \tag{16}$$

where $\lambda_\varphi = \text{trace}(\varphi)$ and $\lambda_{\Lambda_i} = \text{trace}(\Lambda_i)$ and $\lambda_{\mathbf{K}_\mathbf{i}} = \text{trac}(\mathbf{K}_i)$ and $\lambda_{\mathbf{K}} = (\lambda_{\mathbf{K}_\mathbf{x}} + \lambda_{\mathbf{K}_\mathbf{y}})$. To obtain parameters $\theta = \{\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\delta}_\mathbf{i}^2\}$ in the KPDICCA method, we used expectation maximization (EM) algorithm presented in [34]. In order to infer $\mathbf{z}$, we need to marginalize out the latent independent factors of $\mathbf{z}_\mathbf{x}$ and $\mathbf{z}_\mathbf{y}$. Therefore, we assume that $\boldsymbol{\beta}_x$ is fixed (then $\mathbf{B}_x$ is fixed) and marginalizing over an independent factor. Similarly, the study [34] involves an integral over $p(\tilde{\mathbf{x}}|\mathbf{z}_\mathbf{x})p(\mathbf{z}_\mathbf{x})$ where $\tilde{\mathbf{x}} = \phi(\mathbf{x}) - \mathbf{z}\boldsymbol{\alpha}_\mathbf{x}^\mathbf{T}\phi(\mathbf{x})$.

As the prior $p(\mathbf{z}_\mathbf{x})$ is Gaussian and it is multiplied with a linear term, we can integrate $\mathbf{z}_\mathbf{x}$ out analytically, obtaining $\tilde{\mathbf{x}} = \mathcal{N}(\mathbf{0}, \varphi_x)$ where $\varphi_x = \phi(\mathbf{x})^T(\boldsymbol{\beta}_\mathbf{x}\boldsymbol{\beta}_\mathbf{x}^\mathbf{T} + \boldsymbol{\delta}_\mathbf{x}^2\mathbf{I})\phi(\mathbf{x})$.

Doing the same marginalization for $\mathbf{z}_\mathbf{y}$ leads to the generative model.

$$\begin{aligned} z &\sim \mathcal{N}(0, I_d),\\ \phi(\mathbf{x})|z &= \mathcal{N}\left(\mathbf{z}\boldsymbol{\alpha}_\mathbf{x}^\mathbf{T}\phi(\mathbf{x}), \varphi_x\right),\\ \psi(\mathbf{y})|z &= \mathcal{N}\left(\mathbf{z}\boldsymbol{\alpha}_\mathbf{y}^\mathbf{T}\psi(\mathbf{y}), \varphi_y\right).\end{aligned} \tag{17}$$

If we consider $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_x\boldsymbol{\alpha}_y]^\mathbf{T}$ and $\varpi(\mathbf{O}) = \begin{bmatrix}\phi(\mathbf{x}) & 0\\ 0 & \psi(\mathbf{y})\end{bmatrix}$, we can write above model as

$$\varpi(\mathbf{O})|z = \mathcal{N}(\mathbf{z}\boldsymbol{\alpha}\,\varpi(\mathbf{O}), \varphi), \tag{18}$$

where

$$\varphi = \varpi(\mathbf{O})^T\begin{bmatrix}\left(\boldsymbol{\beta}_\mathbf{x}\boldsymbol{\beta}_\mathbf{x}^\mathbf{T} + \boldsymbol{\delta}_\mathbf{x}^2\mathbf{I}\right) & 0\\ 0 & \left(\boldsymbol{\beta}_\mathbf{y}\boldsymbol{\beta}_\mathbf{y}^\mathbf{T} + \boldsymbol{\delta}_\mathbf{y}^2\mathbf{I}\right)\end{bmatrix}\varpi(\mathbf{O}) = \varpi(\mathbf{O})^T\Omega\,\varpi(\mathbf{O}). \tag{19}$$

This is exactly the model proposed in [30] for interpreting CCA probabilistically. For obtaining the above solution, we implicitly assume that the dimensionalities of the $\mathbf{z}_\mathbf{x}$ and $\mathbf{z}_\mathbf{y}$ are sufficiently high to produce a nonconstrained covariance matrix (φ). However, Klami and Kaski [31] propose a new algorithm that does not require this assumption and propose a more general EM algorithm for linear projections. The algorithm includes an additional step that marginalizes the z out to enable estimation of the dependent matrixes ($\mathbf{B}_x$ and $\mathbf{B}_y$).

The EM algorithm for optimizing the extended probabilistic CCA is described in Figure 2 and repeats the two steps until convergence. This method is a linear approach and is capable of finding linear relationship between two modalities. For extending this approach to nonlinear relation, similar to the KCCA method, by considering $\mathbf{W}$ and $\mathbf{B}$ as (10) and (11) and substituting into the Klami's method, we can obtain the EM algorithm for updating the parameters $\theta = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}^2\}$.

By substituting equation (10) into EM algorithm (Figure 2) part (1), we have

$$\begin{aligned}\mathbf{M}_\mathbf{t} &= \left(I + \mathbf{W}_\mathbf{t}^\mathbf{T}\varphi_t^{-1}\mathbf{W}_\mathbf{t}\right)^{-1} = \left(I + \boldsymbol{\alpha}_\mathbf{t}^\mathbf{T}\Omega_\mathbf{t}^{-\mathbf{1}}\boldsymbol{\alpha}_\mathbf{t}\right)^{-1},\\ A_t &= \mathbf{M}_\mathbf{t}\boldsymbol{\alpha}_\mathbf{t}^\mathbf{T}\varpi(\mathbf{O})\left(\varpi(\mathbf{O})^T\Omega_\mathbf{t}\,\varpi(\mathbf{O})\right)^{-1} = \mathbf{M}_\mathbf{t}\boldsymbol{\alpha}_\mathbf{t}^\mathbf{T}\Omega_\mathbf{t}^{-\mathbf{1}}\varpi(\mathbf{O})^{T^{-1}},\end{aligned} \tag{20}$$

where $\varpi(\mathbf{O})\varpi(\mathbf{O})^{-1} = \begin{bmatrix}\mathbf{K}_\mathbf{x}\mathbf{K}_\mathbf{x}^{-1} & 0\\ 0 & \mathbf{K}_\mathbf{y}\mathbf{K}_\mathbf{y}^{-1}\end{bmatrix}$ and considering $\mathbf{K}_x$ and $\mathbf{K}_y$ are invertible, we will have $\varpi(\mathbf{O})\varpi(\mathbf{O})^{-1} = \mathrm{I}$. Now by the use of EM for updating the formula for $\mathbf{W}$, we can obtain the following equation for updating the $\boldsymbol{\alpha}$ parameter:

$$\boldsymbol{\alpha}_{\mathbf{t+1}} = \Omega_\mathbf{t}^{-\mathbf{1}}\boldsymbol{\alpha}_\mathbf{t}\mathbf{M}_\mathbf{t}^\mathbf{T}\left(\mathbf{M}_\mathbf{t} + \left(\mathbf{M}_\mathbf{t}\boldsymbol{\alpha}_\mathbf{t}^\mathbf{T}\Omega_\mathbf{t}^{-\mathbf{1}}\right)\left(\mathbf{M}_\mathbf{t}\boldsymbol{\alpha}_\mathbf{t}^\mathbf{T}\Omega_\mathbf{t}^{-\mathbf{1}}\right)^{\mathbf{T}}\right)^{-\mathbf{1}}. \tag{21}$$

In the second step, we marginalize over the parameter z and use similar a method to provide an updating function for $\boldsymbol{\beta}_i$ as follows:

$$\boldsymbol{\beta}_{it+1} = \Lambda_{it}^{-1}\boldsymbol{\beta}_{\mathbf{it}}\mathbf{N}_{it}^T\left(\mathbf{N}_\mathbf{it} + \left(\mathbf{N}_\mathbf{it}\,\boldsymbol{\beta}_{it}^T\Lambda_{it}^{-1}\right)\left(\mathbf{N}_\mathbf{it}\,\boldsymbol{\beta}_{it}^T\Lambda_{it}^{-1}\right)^{\mathbf{T}}\right)^{-\mathbf{1}}, \tag{22}$$

where $i \in \{x, y\}$ and

$$\begin{aligned}\Lambda_{it} &= \boldsymbol{\alpha}_{\mathbf{i}t+1}\boldsymbol{\alpha}_{\mathbf{i}t+1}^\mathbf{T} + \boldsymbol{\delta}_{\mathbf{i}t}^{\mathbf{2}}\mathbf{I},\\ \mathbf{N}_{\mathbf{i}t} &= \left(I + \boldsymbol{\beta}_{it}^T\Lambda_{it}^{-1}\boldsymbol{\beta}_{\mathbf{i}}t\right)^{-1}.\end{aligned} \tag{23}$$

Actually, the variance is recovered using the model described in equation (9) by

$$\boldsymbol{\delta}_{\mathbf{i}t+1}^{\mathbf{2}} = \frac{\mathbf{1}}{\mathbf{d_i}}\mathbf{trace}\left(\mathbf{I_i} - \Lambda_{it}^{-1}\boldsymbol{\beta}_{it+1}\mathbf{N}_{\mathbf{i}t}^T\boldsymbol{\beta}_{\mathbf{i}t+1}^T - \boldsymbol{\alpha}_{\mathbf{i}t+1}\boldsymbol{\alpha}_{\mathbf{i}t+1}^\mathbf{T}\right). \tag{24}$$

3. Application

In this paper, we assess the proposed method and its competitors on the speech recognition and emotion recognition datasets. Here, M2VTS [35] as an audio-visual database is used for speech recognition. The M2VTS dataset includes 185 recordings from 37 subjects (12 females and 25 males). Each speaker utters five shots. The subjects utter the digits from "0" to "9" within each shot, and their audio and video signals are recorded. The sampling rate of audio

signals is 48 KHz, and the frame rate of video is 25 Hz. Several features for characterizing speech signals have been proposed such as cepstral coefficients [13], discrete cosine transform (DCT), mel-frequency cepstral coefficients (MFCCs) [9], and the perceptual linear predictor (PLP) [36]. First, the background speech signal is removed, and then the cleaned speech signal is segmented into successive hamming windows with 50% overlap, where the length of each window is 512 samples. From each windowed signal, 12 MFCC coefficients are extracted.

Since our processing is simultaneous, we have to characterize the lip motion in parallel. In this regard, the lip contour should be first elicited to trace its key points in successive frames. We here use the Rohani et al. [13] method, in which we first divide a colored face image into lip and nonlip clusters. This segmentation is done by simulating a simple geometric lip model and applying spatial fuzzy C-mean clustering in order to extract the lip contour. The geometric lip model described by equation (25) is presented in Figure 3.

$$y_1 = h_1\left(\left(\frac{x - sy_1}{w}\right)^2\right)^{1+\delta^2} - h_1,$$
$$y_2 = \frac{-h_2}{\left(w - x_{\text{off}}\right)^2}\left(\left|x - sy_2\right| - x_{\text{off}}\right)^2 + h_2, \tag{25}$$

where $x \in [-w, w]$ at (0, 0).

After matching the lip model to each image, a lip contour is extracted. The lip model contains six features (two key points in the upper lip and four points in the lower lips) that need to be traced in successive frames.

The employed emotion recognition datasets are eNTERFACE and RML, both of which include six states of emotions such as anger, disgust, fear, happy, sad, and surprise [37, 38]. In eNTERFACE, 44 subjects participated whose video is recoded at 25 frame per second, and their acoustic signals are recorded at a sampling rate of 48 KHz. On the other hand, in the Ryerson database, eight subjects speak six different languages, generating three believable reactions to all the situations. Their acoustic sampling rate is 22050 Hz, with the video frame rate of 30.

For the emotion recognition dataset, similar processing stages are applied. In order to remove the speech noise, we take the wavelet transform from this signal and by applying thresholding to the energy of wavelet coefficients on different scales, we remove those scales whose energy values are less than an empirical threshold [39]. After reconstructing the signals, the first energy of the signal in the time domain within each window is determined [40] and then the first 12 MFCC features [26, 34] are added to the feature vector. In the emotion recognition system, facial expression features play a very important role. The challenging issue in the video processing is to precisely extracting the face margin. In this research, the Haar cascade technique [41] is employed to detect the face part. The image in each frame is resized to 64×64 pixels. Afterward, a Gabor wavelet filter [42] in five scales and eight orientations is applied to each image to elicit key facial features [6, 26]. Nonetheless, Gabor feature vectors are high-dimensional. To reduce the feature size, principle component analysis (PCA) is deployed.

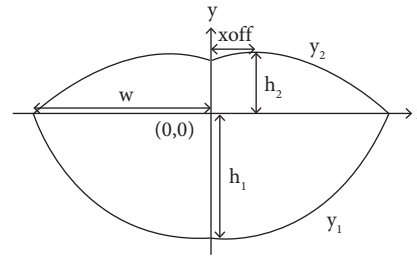

Figure 3: Geometric lip model.

4. Experimental Results

In this section, the results of applying the proposed method along with PDICCA, KPCCA, KCCA, CFA, and KCFA to the described speech processing and emotional recognition datasets are presented. As described before, the M2VTS database (https://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/) [35] for audio-visual speech recognition, eNTERFACE (https://www.enterface.net/enterface05/docs/results/databases/project2_database.zip) [43], and Ryerson (RML) databases (https://www.kaggle.com/datasets/ryersonmultimedialab/ryerson-emotion-database) [44] for audio-visual emotion recognition are employed in this research. The model parameters are estimated during the cross-validation phase. The number of hidden states and the number of Gaussian per state in the hidden Markov model (HMM), over the speech dataset, are selected at three and one, respectively. The number of HMM hidden states and the number of Gaussian per state in the emotion recognition dataset are six and three, respectively. The variance parameter of the Gaussian kernel is set to 14 for both applications.

The computational complexity of the kernel-based methods depends on the number of samples. Here, the audio-visual features' dimension is 200. In addition, among the subjects, 10 persons are selected randomly to overcome the memory shortage in each experiment. 70% of subjects are selected for the training phase, and the rest are chosen as the test set. We repeat this dividing 10 times, and the final results are determined by taking an average over the experiments. The final results are demonstrated in Figures 4–15.

Figures 4 and 5 report the audio-visual emotion recognition accuracy and F1-score of the eNTERFACE and Ryerson (RML) datasets for the CCA, CFA, and PDICCA methods, and the results are calculated for different dimension sizes.

Figure 6 reports audio-visual emotion recognition ROC curves for the best accuracy result of the eNTERFACE and Ryerson (RML) datasets for the CCA, CFA, and PDICCA methods and the ROC curves are shown for each emotion classes.

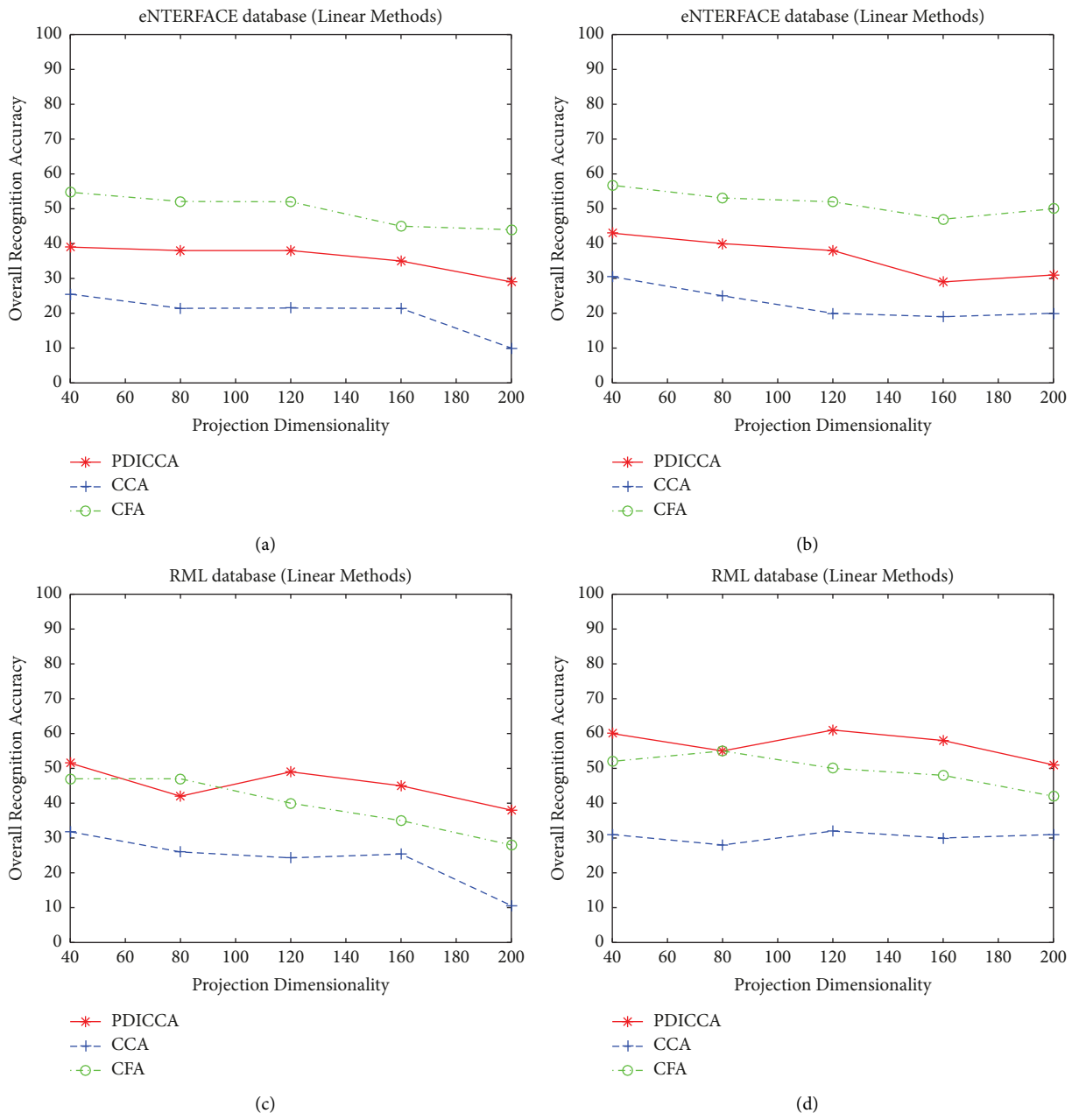

Figure 4: Experimental results of linear CCA, CFA, and PDICCA methods based on overall accuracy measure. (a) eNTERFACE feature level. (b) eNTERFACE decision level. (c) RML feature level. (d) RML decision level.

The experimental results for audio-visual speech recognition accuracy and F1-score using MFCC and PLP acoustic features for feature fusion and decision fusion are shown in Figures 7 and 8, respectively. In this proposed algorithm, the dimension sizes between one and five are of concern for independent space parts. The results obtained from the tests indicate that at the feature level and decision level for three and five dimensions, the best results are achieved. The ROC curves in each class for best accuracy results are reported in Figure 9. These results exhibit that the fusion of dependent and independent parts of some bimodal data could improve the recognition accuracy. Nevertheless, the accuracy of the linear methods is still not acceptable for a real application; therefore, we applied the kernel version of these methods to the same features in order to increase the accuracy.

Figures 10 and 11 depict the experimental results for the proposed KPDICCA and state-of-the-art KCCA and KCFA on the eNTERFACE and Ryerson (RML) datasets based on accuracy and F1-score metrics. In the proposed KPCDICA method, we set different dimension sizes for the independence latent variable dimension and the validation results. These results are reported for different regularization parameter values, and it can be demonstrated that this parameter affects the recognition performance; however, it is difficult to identify an optimum interval for this parameter.

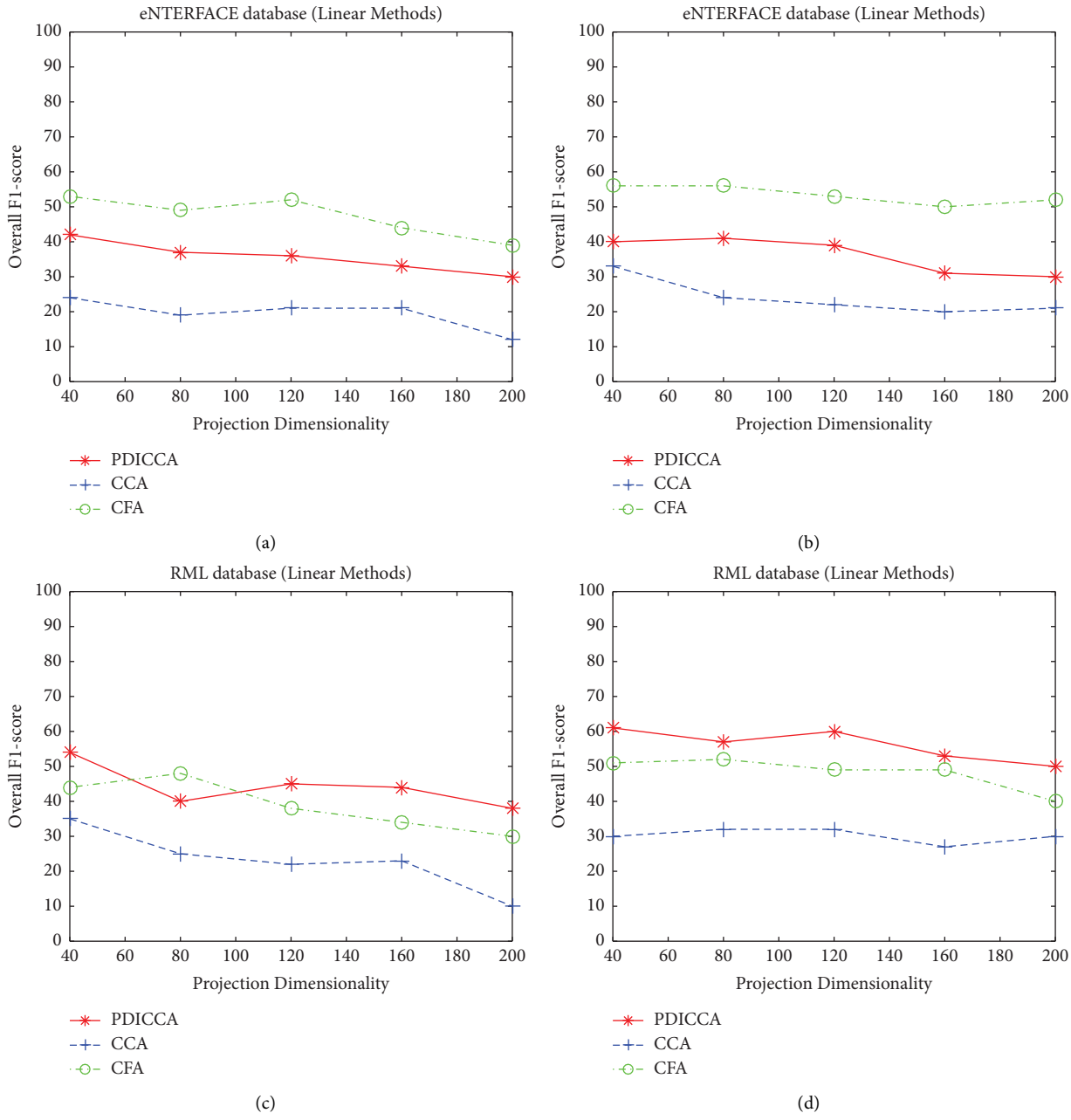

Figure 5: Experimental results of linear CCA, CFA, and PDICCA methods based on overall F1-score measure. (a) eNTERFACE feature level. (b) eNTERFACE decision level. (c) RML feature level. (d) RML decision level.

For instance, in the decision fusion case, for the eNTERFACE and RML datasets, the best results are obtained at $r = 1.0$ and $r = 0.4$, respectively.

Figure 12 reports audio-visual emotion recognition ROC curves for the best accuracy results of the eNTERFACE and Ryerson (RML) datasets for the KPDICCA, KCCA, and KCFA methods, and the ROC curves are show for each emotion class.

The audio-visual speech recognition accuracy and F1-score of the conventional methods, together with the proposed method for the M2VTS database using MFCC and PLP acoustic features are presented in Figures 13 and 14. In KPDICCA, the different dimension sizes are considered between one and six for the independence space, and the results indicate that for independence size four, the best result is achieved.

Figure 15 depicts audio-visual speech recognition ROC curves for the best accuracy result of the M2VTS database using MFCC and PLP for the KPDICCA, KCCA, and KCFA methods. The ROC curves are show for each emotion classes.

By comparing the recognition accuracy and F1-score in Figures 4 and 5, 7-8, 10-11, and 13-14 on real datasets, we can find that the relation between audio and video data are nonlinear and using kernel can handle this problem and find a suitable accuracy for the emotion and speech recognition system. This supremacy is emerged from incorporating

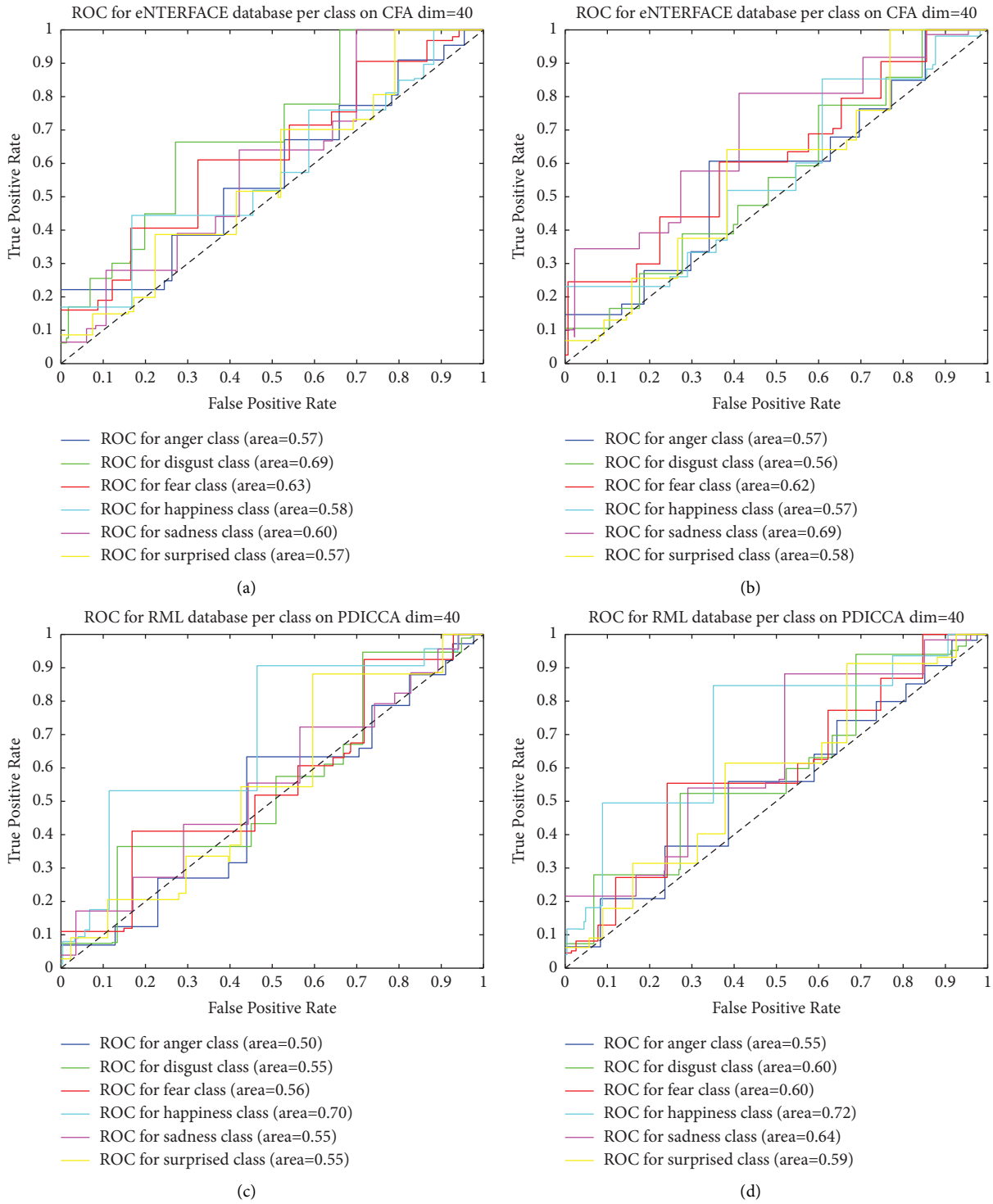

Figure 6: ROC curves for best accuracy results of linear CCA, CFA, and PDICCA methods. (a) eNTERFACE feature level. (b) eNTERFACE decision level. (c) RML feature level. (d) RML decision level.

dependent and independent variables containing nonlinear information. It can be also interpreted that the extra information that implies the superiority of KPDICCA to KCCA is the incorporation of independent latent variable information as an independent feature for each modality in the HMMs. In other words, when we consider just the common information between two modalities, some discriminative features belonging to each modality carrying unique and independent information are removed. This elimination causes a decrease in the performance of the recognition system. As we can see from the results, the performance of KPDICCA declines when the dimension of

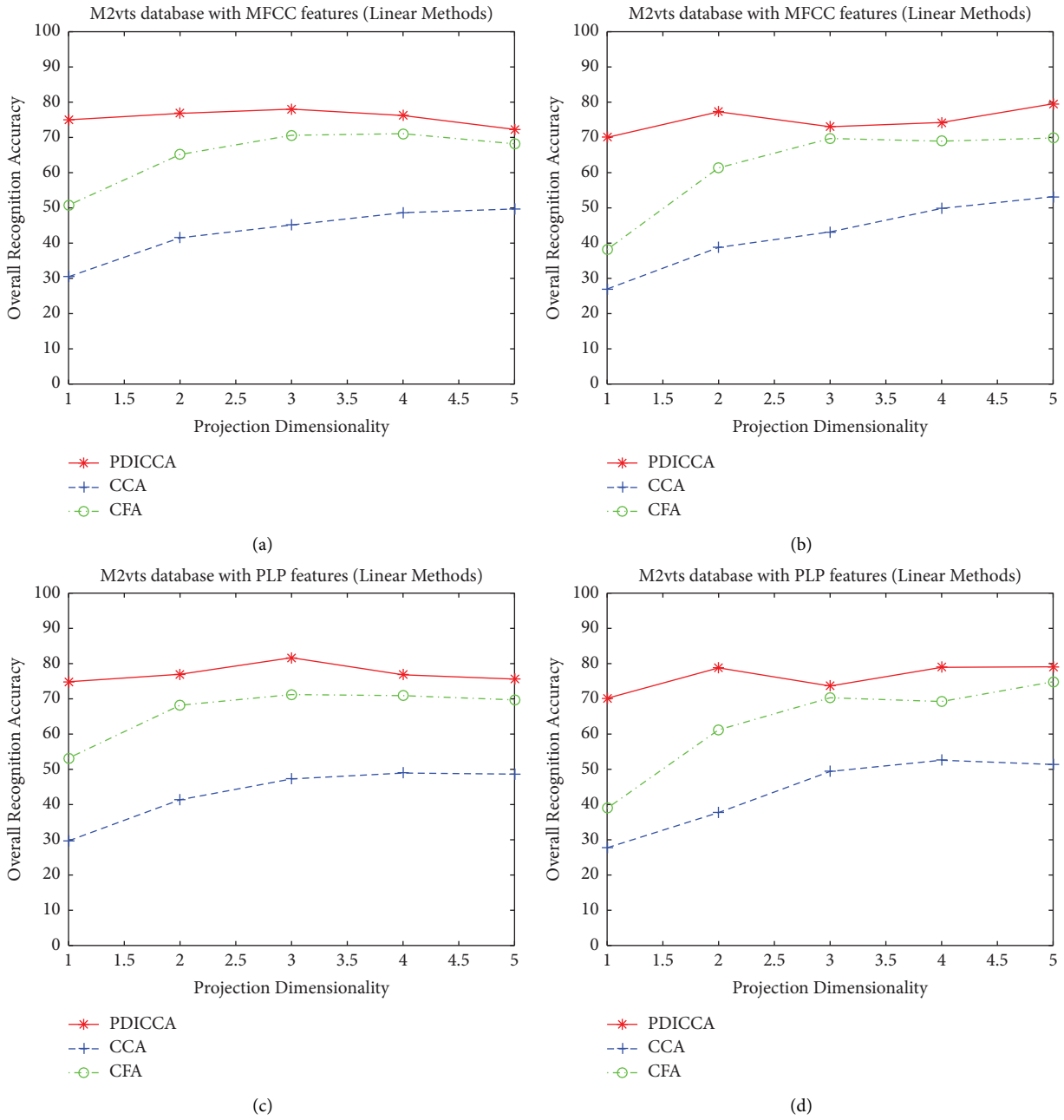

Figure 7: Experimental results of the linear CCA, CFA, and PDICCA methods based on overall accuracy measure. (a) Feature level on the M2VTS database with MFCC features. (b) Decision level on the M2VTS database with MFCC features. (c) Feature level on the M2VTS database with PLP features. (d) Decision level on the M2VTS database with PLP features.

the elicited variables is increased. On the other hand, it should be pointed out that in the RML dataset, due to the low number of samples, by increasing the number of elicited features, the performance of both the KCCA and KPDICCA methods decreases, which is caused by the curse of dimensionality.

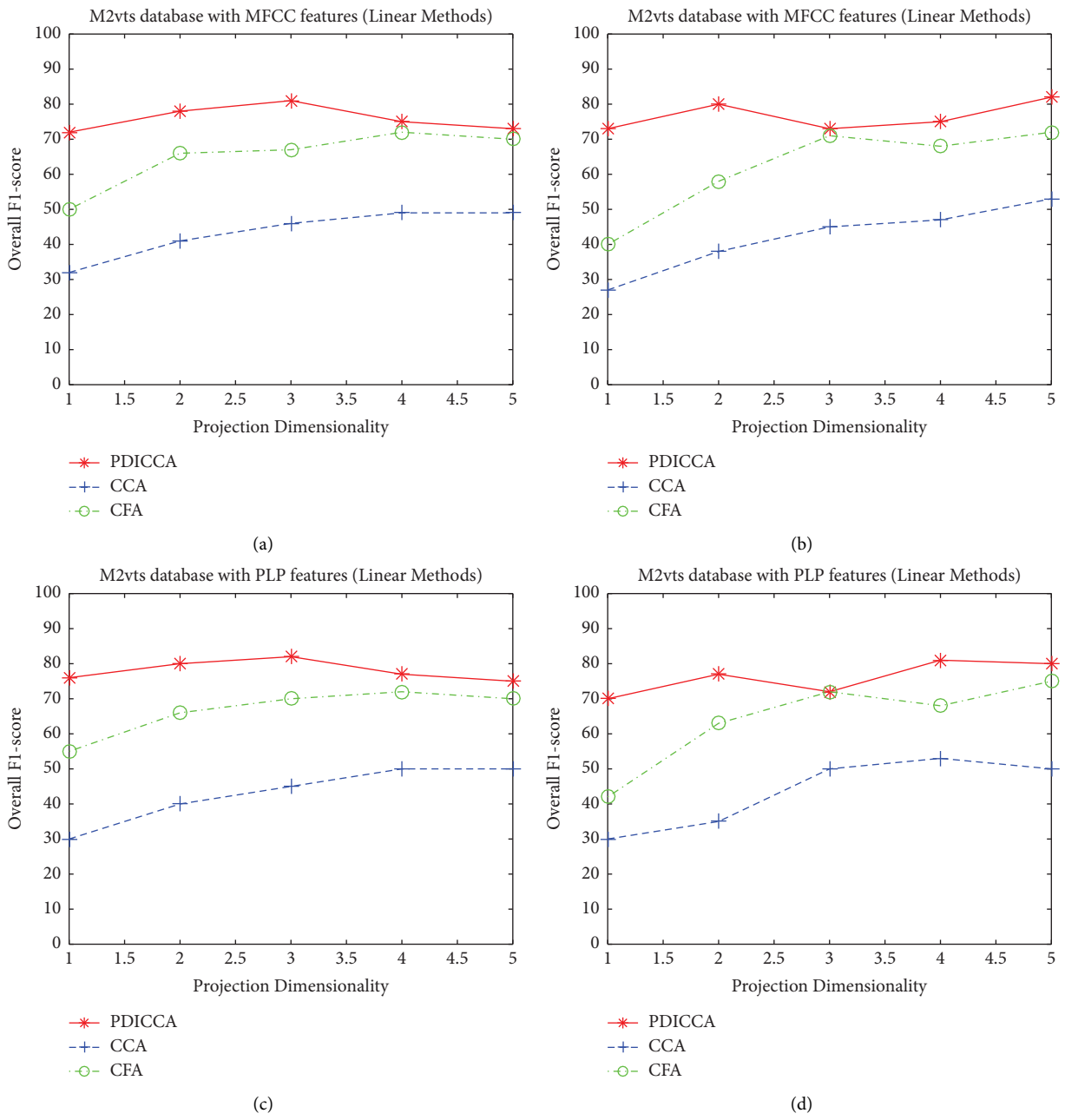

Figure 8: Experimental results of linear CCA, CFA, and PDICCA methods based on overall F1-score measure. (a) Feature level on M2VTS database with MFCC features. (b) Decision level on M2VTS database with MFCC features. (c) Feature level on M2VTS database with PLP features. (d) Decision level on M2VTS database with PLP features.

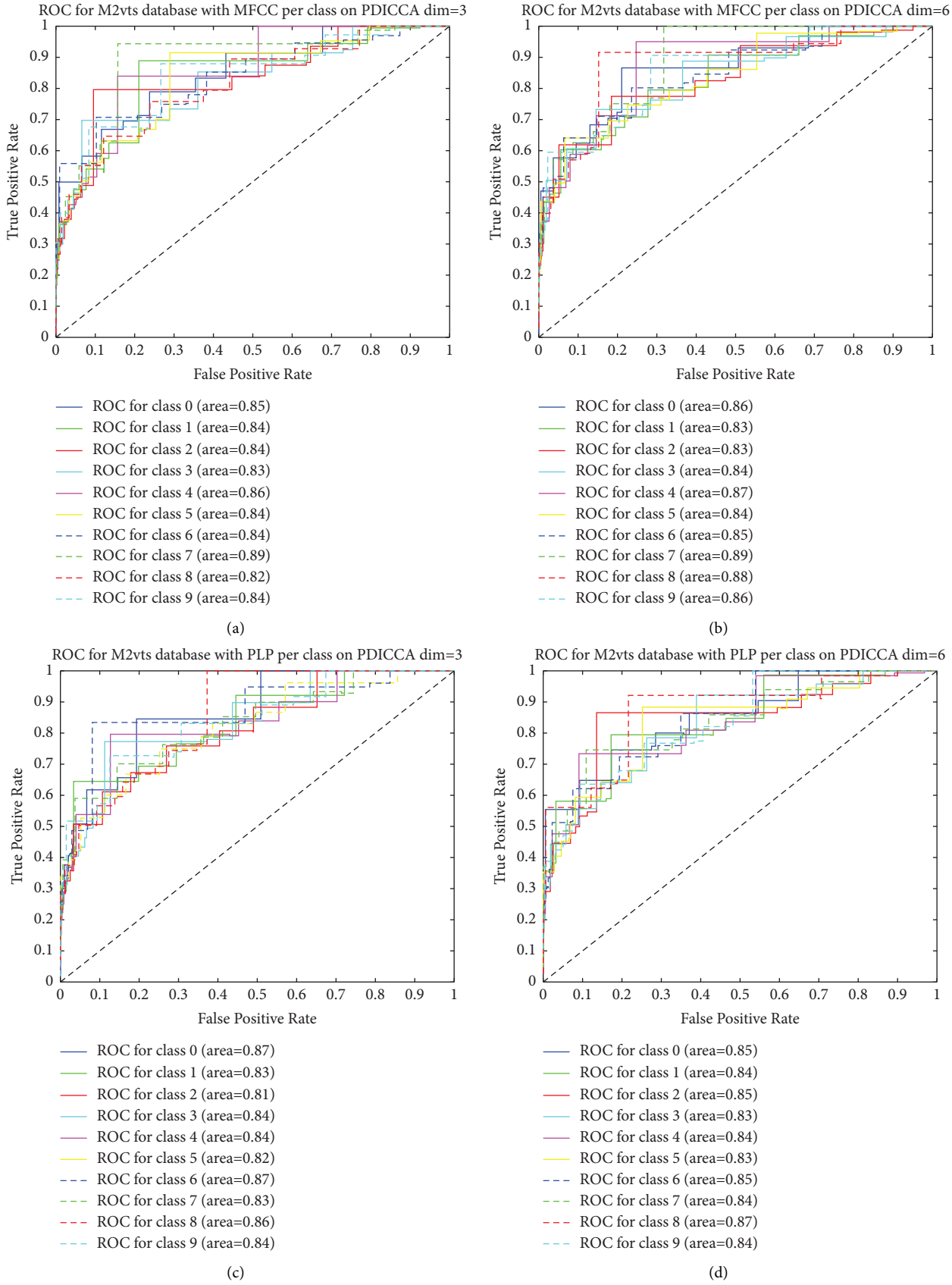

Figure 9: ROC curves for best accuracy results of linear CCA, CFA, and PDICCA methods. (a) Feature level on M2VTS database with MFCC features. (b) Decision level on M2VTS database with MFCC features. (c) Feature level on M2VTS database with PLP features. (d) Decision level on M2VTS database with PLP features.

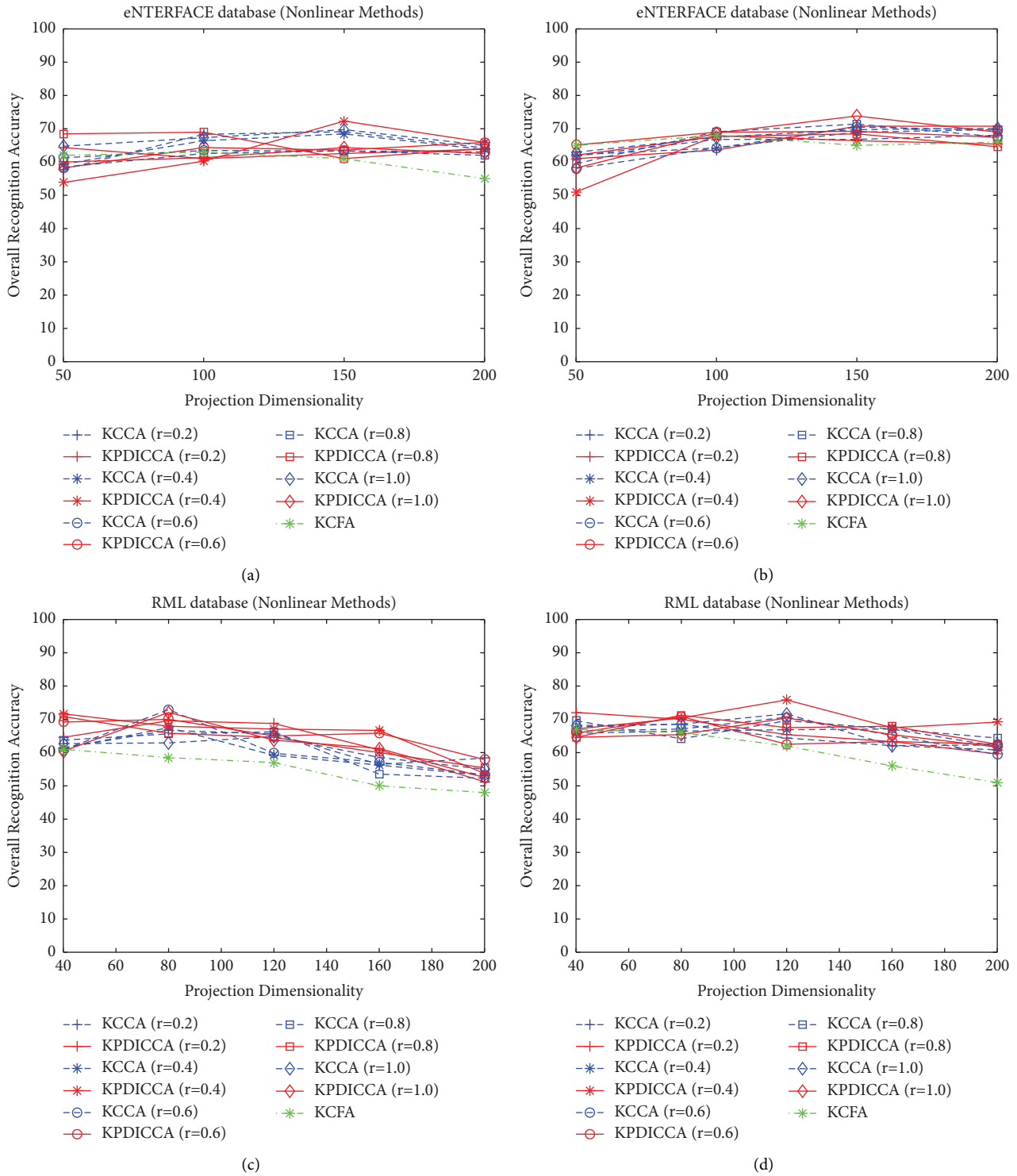

Figure 10: Experimental results of nonlinear KCCA, KCFA, and KPDICCA methods base on overall accuracy measure. (a) eNTERFACE feature level. (b) eNTERFACE decision level. (c) RML feature level. (d) RML decision level.

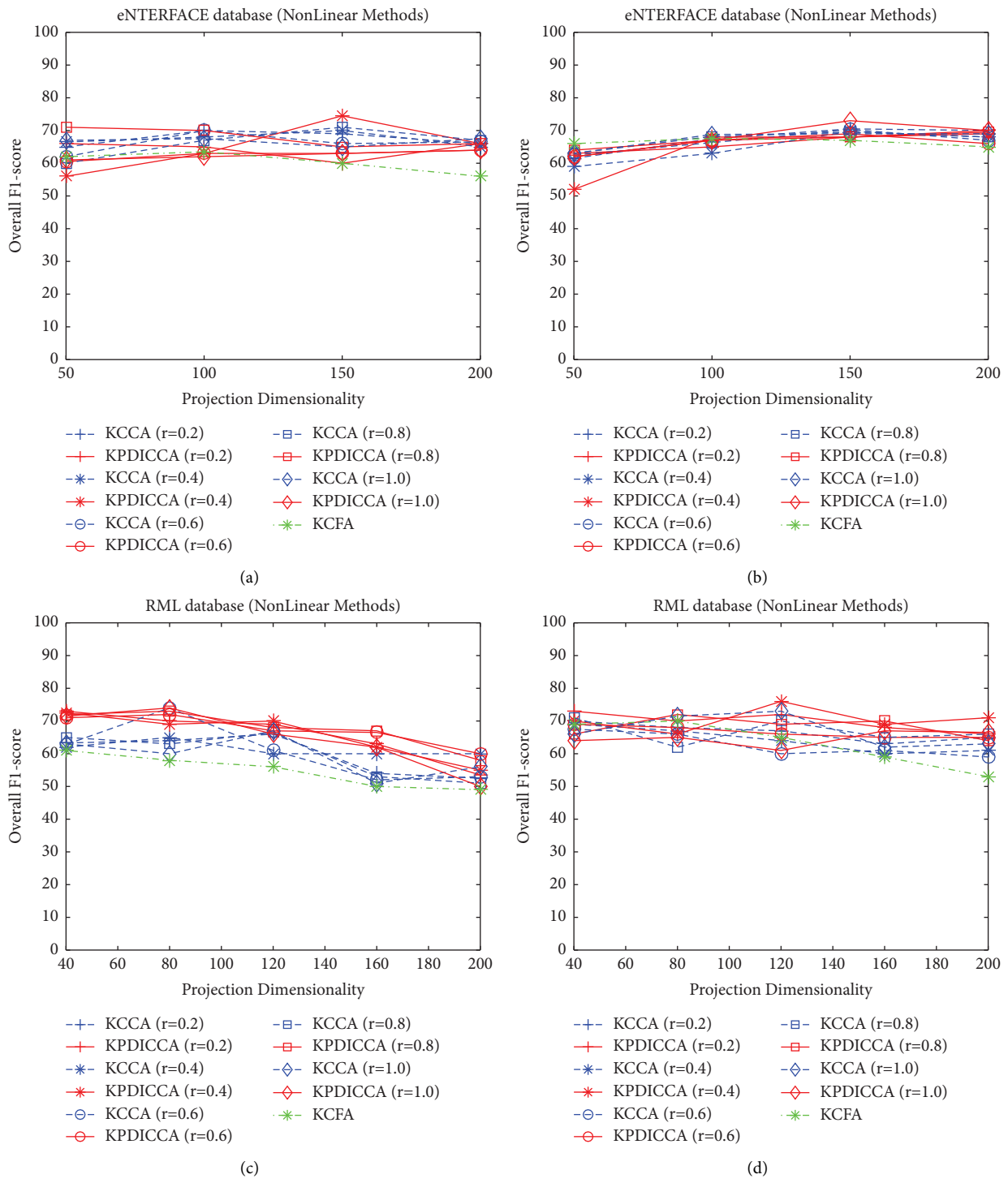

Figure 11: Experimental results of nonlinear KCCA, KCFA, and KPDICCA methods based on overall F1-score measure. (a) eNTERFACE feature level. (b) eNTERFACE decision level. (c) RML feature level. (d) RML decision level.

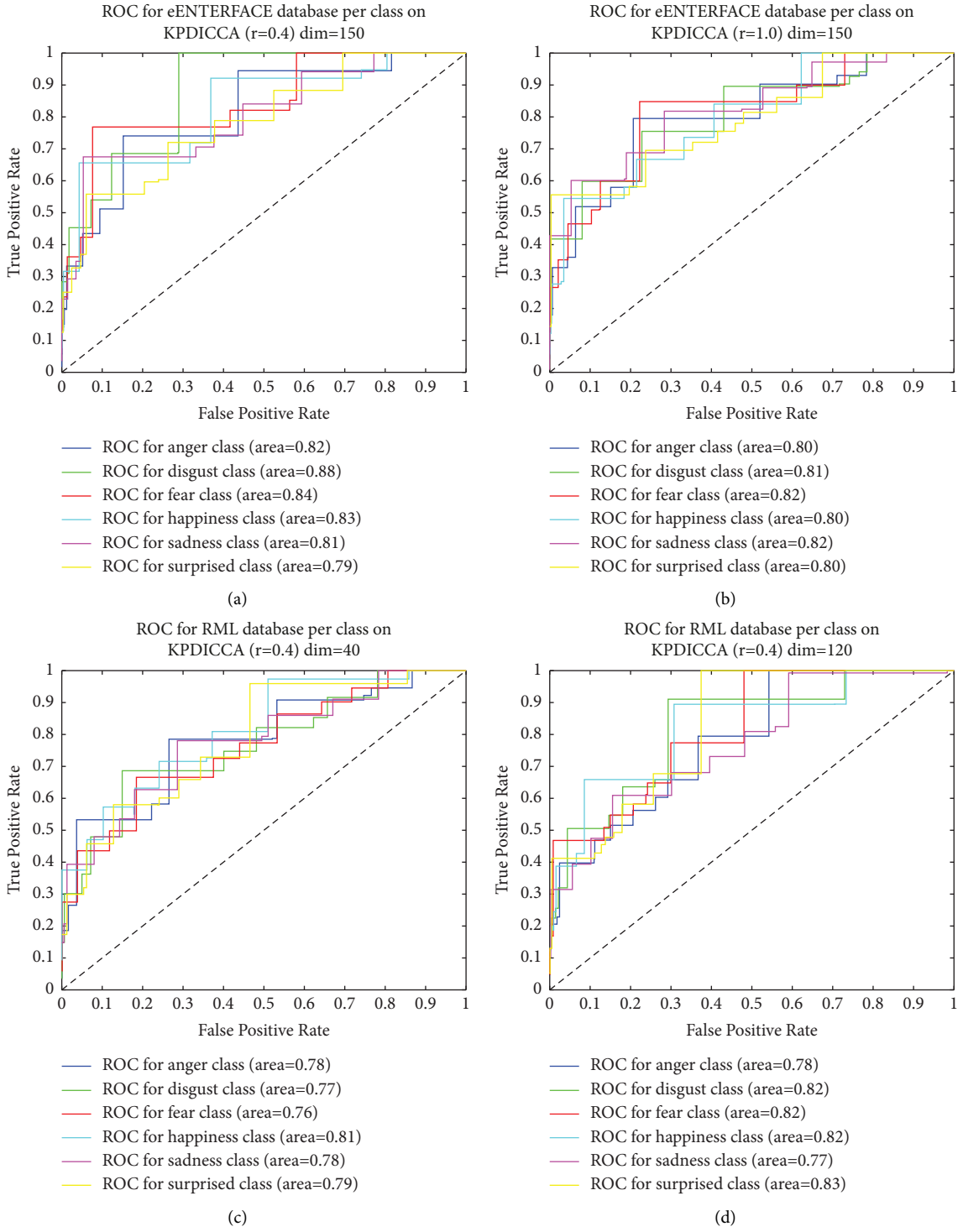

Figure 12: ROC curves for best accuracy results of nonlinear KCCA, KCFA and KPDICCA methods. (a) eNTERFACE feature level. (b) eNTERFACE decision level. (c) RML feature level. (d) RML decision level.

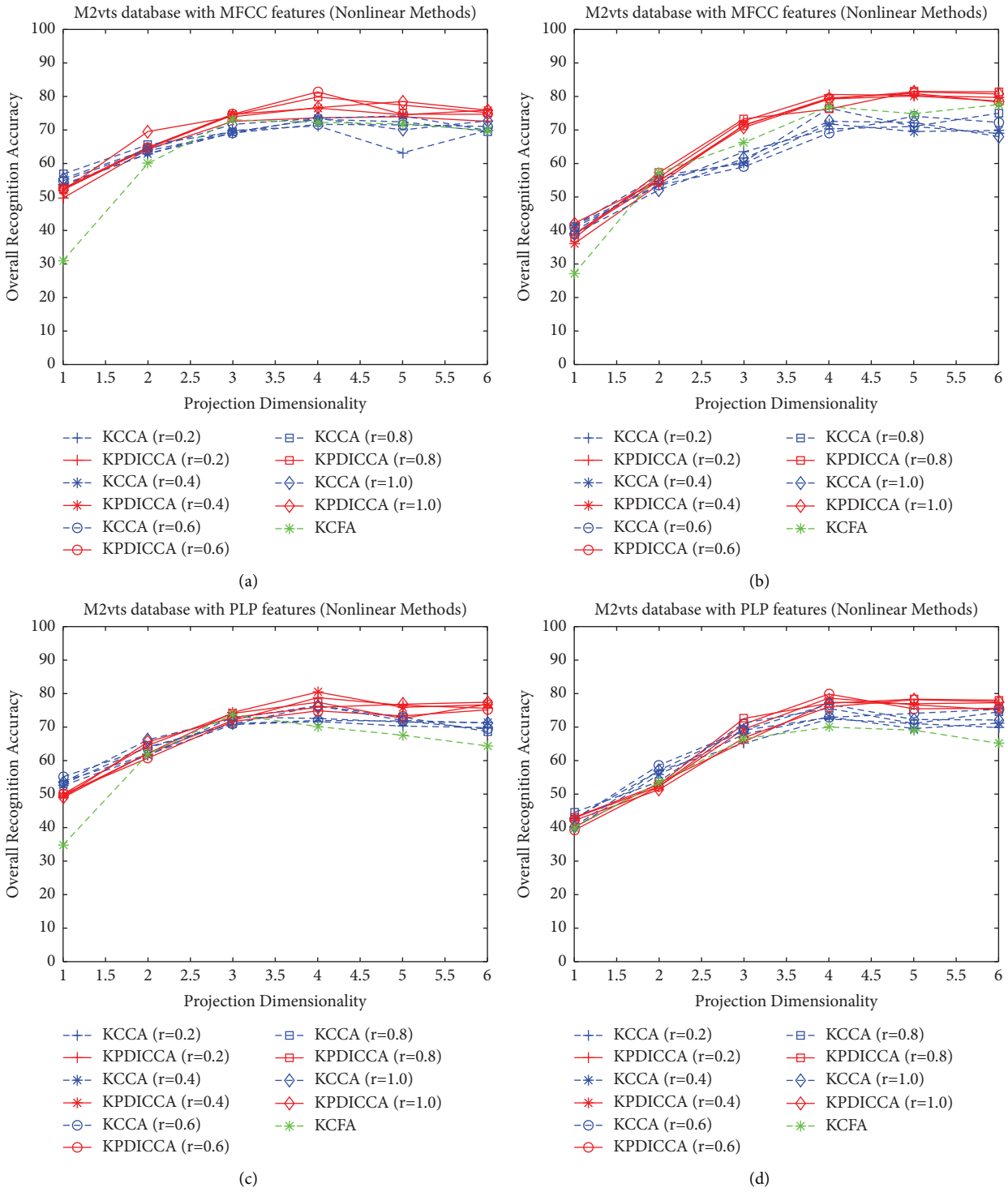

Figure 13: Experimental results of nonlinear KCCA, KCFA, and KPDICCA methods based on overall accuracy measure. (a) Feature level on the M2VTS database with MFCC features (b) Decision level on M2vts database with MFCC features (c) Feature level on the M2VTS database with PLP features. (d) Decision level on the M2VTS database with PLP features.

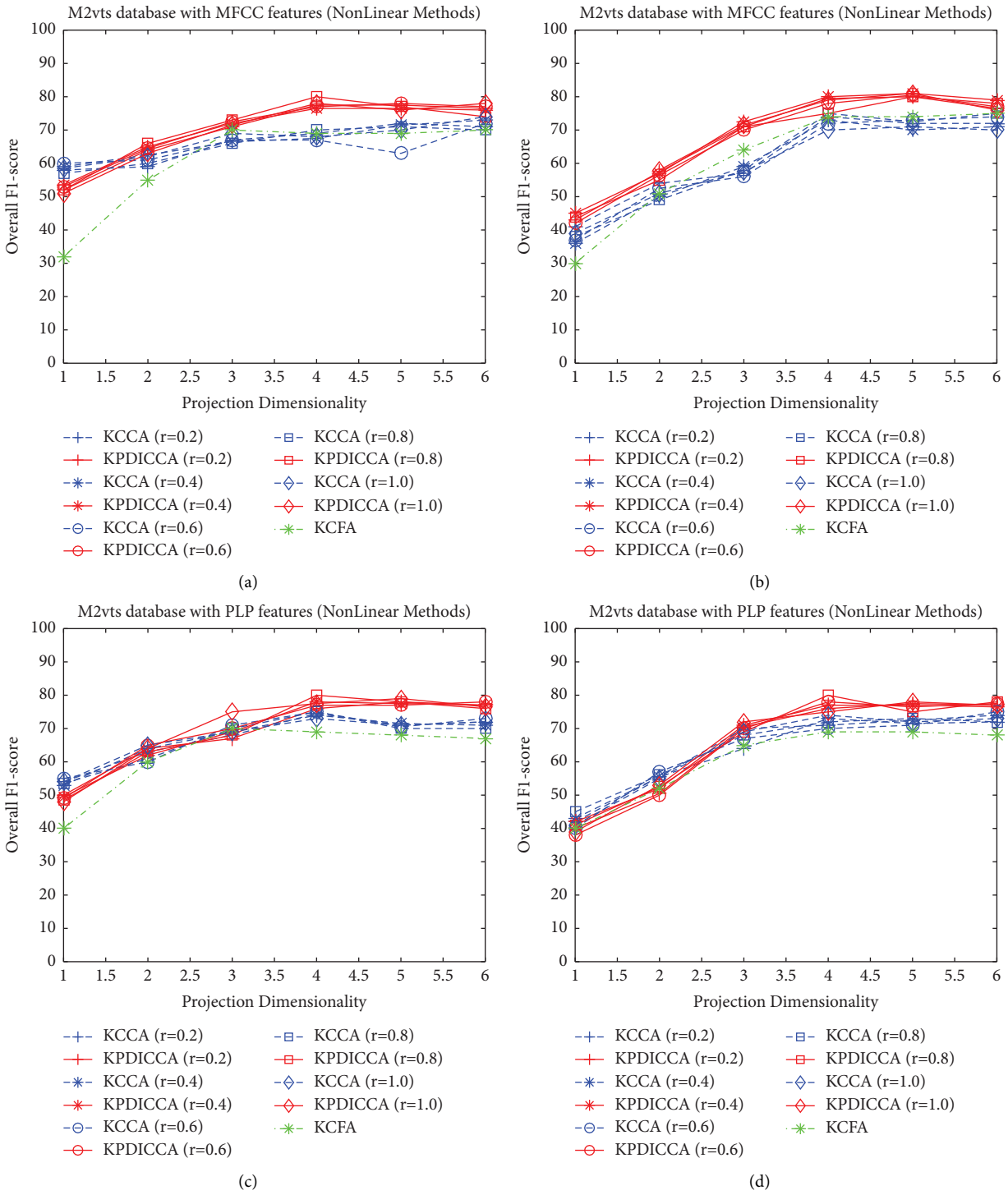

Figure 14: Experimental results of nonlinear KCCA, KCFA, and KPDICCA methods based on overall F1-score measure. (a) Feature level on the M2VTS database with MFCC features (b) Decision level on M2vts database with MFCC features (c) Feature level on M2VTS database with PLP features. (d) Decision level on M2VTS database with PLP features.

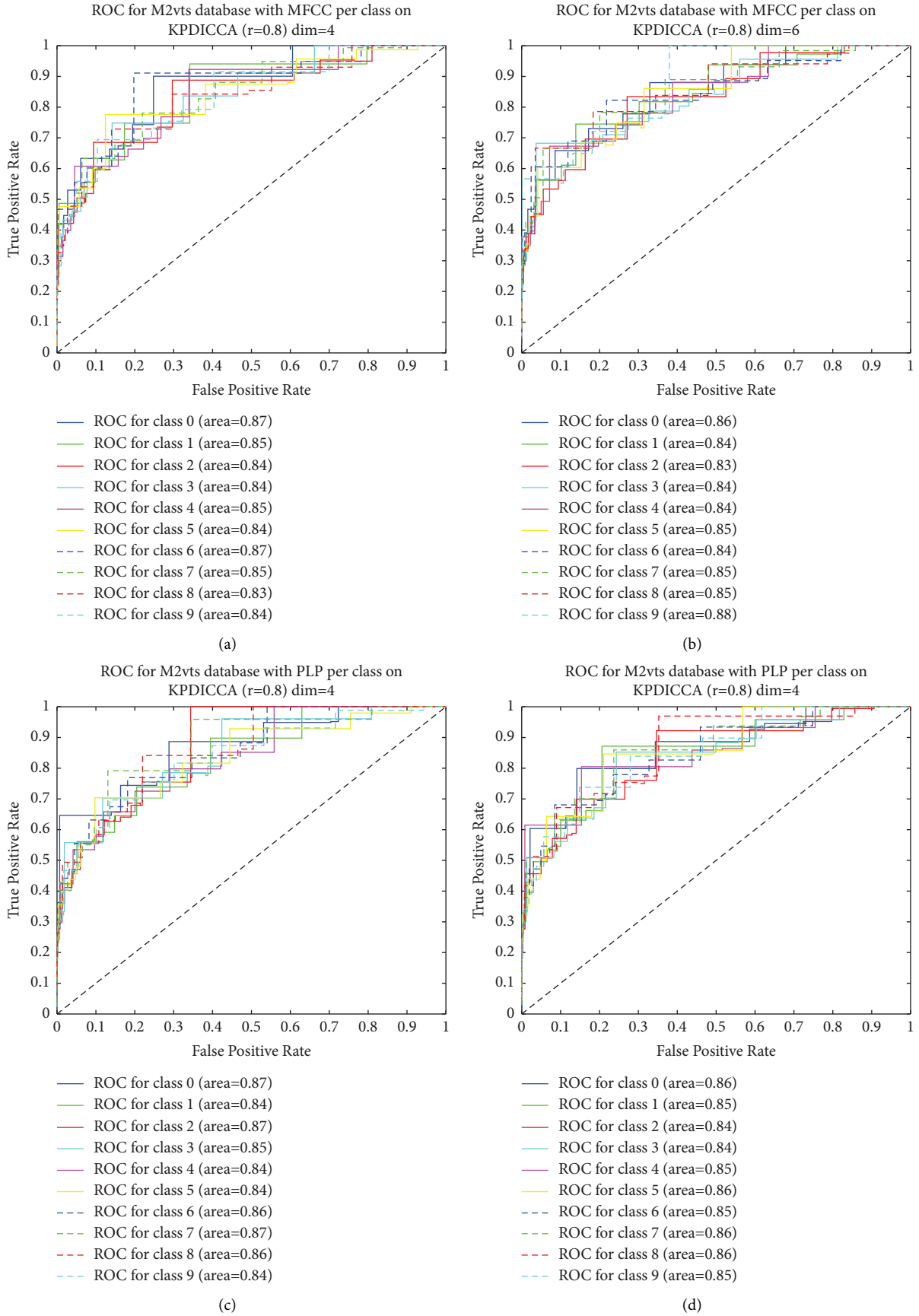

Figure 15: ROC curves for best accuracy results of nonlinear KCCA, KCFA, and KPDICCA methods. (a) Feature level on the M2VTS database with MFCC features. (b) Decision level on M2vts database with MFCC features. (c) Feature level on the M2VTS database with PLP features. (d) Decision level on the M2VTS database with PLP features.

5. Conclusion

In this paper, a novel approach for audio-visual information fusion based on probabilistic dependent and independent canonical correlation analysis (PDICCA) is proposed. Empirical results reveal that the fusing dependent and independent latent variables of bimodal inputs can increase recognition accuracy. Although a combination of nonlinear dependent latent and set-specific (independent) features provides more discriminative information than just using the dependent latent features, these dependent latent variables have high share in the final results, and nonlinear independent features can be considered auxiliary features that can slightly improve the performance of a recognition system. However, this superiority rises from the fact that KPDICCA captures the data variation in its covariance metrics while KCCA and KCFA do not consider any input tolerance in their formulas. Our experimental results confirmed the feasibility and efficiency of KPDICCA for the multimodal data fusion application. This method provides good results on low-dimensional inputs but for high-dimensional, selecting a suitable regularization factor is capable of better handling high-dimensional inputs when the covariance matrix is sparse and its results on the emotion datasets confirm this claim.

In future work, temporal information can be added to the proposed model to increase its performance. To extend this study, other types of kernels can be assessed for different applications, and also other types of regulation models can be employed. On the other hand, to achieve the best kernel map and dependent and independent latent features, the deep learning approach such as deep CCA (DCCA) and deep canonically correlated autoencoders (DCCAEs) can be used.

Data Availability

The data used to support the findings of the study are available at the following links: https://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb, https://www.enterface.net/enterface05/docs/results/databases/project2_database.zip, and https://www.kaggle.com/datasets/ryersonmultimedialab/ryerson-emotion-database.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] N. Abolpour, R. Boostani, M. A. Masnadi-Shirazi, B. Tahayori, and A. Almasi, "A chaotic multilayer LIF scheme to model the primary visual cortex," *Biomedical Engineering: Applications, Basis and Communications*, vol. 33, no. 04, Article ID 2150030, 2021.

[2] A. P. James and B. V. Dasarathy, "Medical image fusion: a survey of the state of the art," *Information Fusion*, vol. 19, pp. 4–19, 2014.

[3] R. Gupta, N. Malandrakis, B. Xiao et al., "Multimodal prediction of affective dimensions and depression in human-computer interactions," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pp. 33–40, ACM, Orlando, FL, USA, November 2014.

[4] C. Marechal, D. Mikolajewski, K. Tyburek et al., "Survey on AI-based multimodal methods for emotion detection," *High-performance modelling and simulation for big data applications*, vol. 11400, pp. 307–324, 2019.

[5] D. Ivanko, A. Karpov, D. Fedotov et al., "Multimodal speech recognition: increasing accuracy using high speed video data," *Journal on Multimodal User Interfaces*, vol. 12, no. 4, pp. 319–328, 2018.

[6] R. R. Sarvestani and R. Boostani, "FF-SKPCCA: kernel probabilistic canonical correlation analysis," *Applied Intelligence*, vol. 46, no. 2, pp. 438–454, 2017.

[7] Y. Chen, J. Yang, C. Wang, and N. Liu, "Multimodal biometrics recognition based on local fusion visual features and variational Bayesian extreme learning machine," *Expert Systems with Applications*, vol. 64, pp. 93–103, 2016.

[8] J. Cong and B. Zhang, "WITHDRAWN: multi-model feature fusion for human action recognition toward sport sceneries," *Signal Processing: Image Communication*, vol. 84, Article ID 115803, 2020.

[9] O. Rudovic, S. Petridis, and M. Pantic, "Bimodal log-linear regression for fusion of audio and visual features," in *Proceedings of the 21st ACM international conference on Multimedia*, pp. 789–792, New York, NY, USA, October 2013.

[10] H. Tajalizadeh and R. Boostani, "A novel stream clustering framework for spam detection in Twitter," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 3, pp. 525–534, 2019.

[11] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: an overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[12] S. Afshar, R. Boostani, and S. Sanei, "A combinatorial deep learning structure for precise depth of anesthesia estimation from EEG signals," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 9, pp. 3408–3415, 2021.

[13] R. Rohani, F. Sobhanmanesh, S. Alizadeh, and R. Boostani, "Lip processing and modeling based on spatial fuzzy clustering in color images," *International Journal of Fuzzy Systems*, vol. 13, no. 2, pp. 65–73, 2011.

[14] R. Rohani, S. Alizadeh, F. Sobhanmanesh, and R. Boostani, "Lip segmentation in color images," in *Proceedings of the 2008 International Conference on Innovations in Information Technology*, pp. 747–750, Al Ain, United Arab Emirates, December 2008.

[15] S. Alizadeh, R. Boostani, and V. Asadpour, "Lip feature extraction and reduction for HMM-based visual speech recognition systems," in *Proceedings of the 2008 9th International Conference on Signal Processing*, pp. 561–564, IEEE, Beijing, China, October 2008.

[16] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3-4, pp. 321–377, 1936.

[17] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proceedings of the eleventh ACM international conference on Multimedia*, pp. 604–611, New York, NY, USA, November 2003.

[18] H. T. M. Nhat and V. T. Hoang, "Feature fusion by using LBP, HOG, GIST descriptors and Canonical Correlation Analysis for face recognition," in *Proceedings of the 2019 26th International Conference on Telecommunications (ICT)*, pp. 371–375, Hanoi, Vietnam, April 2019.

[19] M. S. Ibrahim and N. D. Sidiropoulos, "Reliable detection of unknown cell-edge users via canonical correlation analysis," *IEEE Transactions on Wireless Communications*, vol. 19, no. 6, pp. 4170–4182, 2020.

[20] Z. Chen, C. Liu, S. Ding et al., "A just-in-time-learning aided canonical correlation analysis method for multimode process monitoring and fault detection," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 6, pp. 5259–5270, 2021.

[21] D. Li, C. Taskiran, N. Dimitrova, W. Wang, M. Li, and I. Sethi, "Cross-modal analysis of audio-visual programs for speaker detection," in *Proceedings of the 2005 IEEE 7th Workshop on Multimedia Signal Processing*, pp. 1–4, Shanghai, China, October 2005.

[22] Y. Li, T. Eichele, V. Calhoun, and T. Adali, "Group study of simulated driving fMRI data by multiset canonical correlation analysis," *Journal of Signal Processing Systems*, vol. 68, no. 1, pp. 31–48, 2012.

[23] C. Taouche, M. C. Batouche, M. Berkane, and A. Taleb-Ahmed, "Multimodal biometric systems," in *Proceedings of the International Conference on Multimedia Computing and Systems (ICMCS)*, pp. 301–308, Coimbatore, India, January 2014.

[24] K. Kumar, G. Potamianos, J. Navratil, E. Marcheret, and V. Libal, "Audio-visual speech synchrony detection by a family of bimodal linear prediction models," *Multibiometrics for Human Identification*, Cambridge University Press, Cambridge, UK, 2011.

[25] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *International Journal of Neural Systems*, vol. 10, no. 05, pp. 365–377, 2000.

[26] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 597–607, 2012.

[27] Y. Shi and H. Ji, "Kernel canonical correlation analysis for specific radar emitter identification," *Electronics Letters*, vol. 50, no. 18, pp. 1318–1320, 2014.

[28] B. Li, L. Qi, and L. Gao, "Multimodal emotion recognition based on kernel canonical correlation analysis," in *Proceedings of the IEEE Workshop on Electronics, Computer and Applications*, pp. 934–937, Ottawa, ON, USA, May 2014.

[29] Y. Wang, S. Cang, and H. Yu, "Mutual information inspired feature selection using kernel canonical correlation analysis," *Expert Systems with Applications X*, vol. 4, Article ID 100014, 2019.

[30] F. R. Bach and M. I. Jordan, *A Probabilistic Interpretation of Canonical Correlation Analysis*, University of California, Berkeley, CA, USA, 2005.

[31] A. Klami and S. Kaski, "Probabilistic approach to detecting dependencies between data sets," *Neurocomputing*, vol. 72, no. 1-3, pp. 39–46, 2008.

[32] G. H. Golub, P. C. Hansen, and D. P. O'Leary, "Tikhonov regularization and total least squares," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 1, pp. 185–194, 1999.

[33] M. Koskinen, J. Viinikanoja, M. Kurimo, A. Klami, S. Kaski, and R. Hari, "Identifying fragments of natural speech from the listener's MEG signals," *Human Brain Mapping*, vol. 34, no. 6, pp. 1477–1489, 2013.

[34] D. Jiang, Y. Cui, X. Zhang, P. Fan, I. Gonzalez, and H. Sahli, "Audiovisual emotion recognition based on triple-stream dynamic Bayesian network models," *Affective Computing and Intelligent Interaction*, pp. 609–618, Springer, Berlin, Heidelberg, 2011.

[35] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database (release 1.00)," in *Audio- and Video-Based Biometric Person Authentication. AVBPA 1997*, J. Bigün, G. Chollet, and G. Borgefors, Eds., Springer, Berlin, Heidelberg, 1997.

[36] H. Hermansky, B. A. Hanson, and H. Wakita, "Perceptuallybased linear predictive analysis of speech," *Proceedings IEEE ICASSP*, vol. 2, pp. 509–512, 1985.

[37] E. Avots, T. Sapiński, M. Bachmann, and D. Kamińska, "Audiovisual emotion recognition in wild," *Machine Vision and Applications*, vol. 30, no. 5, pp. 975–985, 2019.

[38] J. A. Bernabé-Díaz, M. D. C. Legaz-García, J. M. García, and J. T. Fernández-Breis, "Efficient, semantics-rich transformation and integration of large datasets," *Expert Systems with Applications*, vol. 133, pp. 198–214, 2019.

[39] A. Bartlett, V. Evans, I. Frenkel, C. Hobson, and E. Sumera, "Digital hearing aids," 2004, https://www.clear.rice.edu/elec301/Projects01/dighearaid.

[40] C. Wu, J. Lin, and W. Wei, "Two-level hierarchical alignment for semi-coupled HMM-based audiovisual emotion recognition with temporal course," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1880–1895, 2013.

[41] V. Sing, V. Shokeen, and B. Singh, "Face detection by haar cascade classifier with simple and complex backgrounds images using opencv implementation," *International Journal of Advanced Technology in Engineering and Science*, vol. 1, no. 12, pp. 33–38, 2013.

[42] M. J. Lyons, J. Budynek, A. Plante, and S. Akamatsu, "Classifying facial attributes using a 2-D Gabor wavelet representation and discriminant analysis," in *Proceedngs of the 4th International Conference Automatic Face and Gesture Recognition*, pp. 202–207, Grenoble, France, March 2000.

[43] O. Martin and I. Kotsia, "The eNTERFACE05 audiovisual emotion database," in *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)*, Atlanta, GA, USA, April 2006.

[44] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 936–946, 2008.