A Multi-label Multimodal Deep Learning Framework for Imbalanced Data Classification

Samira Pouyanfar, Tianyi Wang, Shu-Ching Chen
School of Computing and Information Sciences
Florida International University
Miami, FL 33199, USA
Email: {spouy001,wtian002,chens}@cs.fiu.edu

***Abstract*—Social media and Web services have provided a notable number of multimedia content. Due to such explosion of multimedia data, the multimedia community has been facing new challenges and exciting opportunities these days. This paper presents a new multimedia framework to address some of the main challenges in this area. In particular, it presents a multi-label multimodal framework for imbalanced data classification. For this purpose, it utilizes audio, visual, and textual data modalities and automatically generates static and temporal features using spatio-temporal deep neural networks. It also manages data with non-uniform distributions using a weighted multi-label classifier. To evaluate this framework, a video dataset containing natural disasters is used for multi-label classification. The supremacy of the proposed framework compared to the existing work is revealed with extensive experiments on this dataset.**

***Keywords*-Multi-label, Multimodal classification, Imbalanced data, Deep learning**

I. Introduction

With the advances and proliferation of social networks and mobile technologies, the world has witnessed the explosion of multimedia data. Multimedia data usually contains various types of modalities such as image, audio, and text. These data modalities are usually complementary, which can be integrated to enhance the final decisions. However, many existing studies only focus on one or two data modalities due to the complexity and difficulty of multimodal data collection, analysis, and fusion [1], [2].

In addition, many real-world data samples can be represented with multiple labels. For example, an image may contain multiple objects or a video may contain various events. In such cases, the data samples cannot be easily categorized by a single class. Therefore, Multi-Label Classification (MLC) is a necessity to solve these problems. In MLC, different from single-label classification, each instance is assigned to multiple labels simultaneously. Due to the high dimensionality of the data, the enormous number of label combinations, and the complex correlation between the labels, MLC is more challenging than a single-label classification problem. Besides, for a multimedia dataset containing multiple data types, it is essential to discover the correlation between both labels and data modalities.

In recent years, deep learning has shown promising results in various applications including image classification, language translation, voice search, cancer detection, and finance [3], [4], [5], [6]. Despite the great success of deep learning in the processing of single data modalities, there are still a few research studies focusing on multimodal deep learning frameworks [7], [8]. This problem is mainly due to the limited available datasets that contain multiple data modalities including text, audio, video, etc. For this purpose, we utilize a new multimodal dataset for natural disaster information management which is originally introduced in [8] and later used in [2]. However, in this work, this dataset is further modified to also serve for multi-label data classification.

Another important challenge in multimedia data is how to handle the non-uniform distribution of the data [9]. This problem (also known as imbalanced data) is common in many real-world scenarios. Although it has been deeply investigated for binary classification or even multi-class classification [10], [11], very few studies can be found to address this issue for MLC.

Considering all these challenges, in this paper, we present a new framework for multi-label multimodal data classification using advanced deep neural networks. In addition, we consider the imbalanced data problem to further enhance the detection performance for both minority and majority classes. This framework is specifically evaluated on a multimodal dataset designed for natural disaster information retrieval and management. However, it can be easily extended for other multimodal multi-label datasets. The contributions of this work include: (1) deep feature extraction using spatio-temporal deep learning models for each modality (text, audio, and image); (2) a new fusion technique which considers the relation between both labels and data modalities while considering the imbalanced data problem; (3) a modified disaster-based video dataset which is designed for multi-label multimodal video classification.

The remaining of this paper is organized as follows. In Section II, the literature in the multi-label multimodal deep learning and imbalanced data is briefly discussed. Section III presents the proposed framework in details. Section IV discusses the experimental results on the disaster dataset.

Finally, Section V provides conclusion and future work.

II. Related Work

As mentioned earlier, MLC is more complex than binary or multi-class classification problems. This is mainly due to the generality of multi-label problems [12]. This issue becomes more daunting when it combines with multimodal data problem [1]. Popular MLC algorithms can be categorized into two groups: 1) problem transformation methods, and 2) algorithm adaption methods [13]. The former transforms MLC into well-studied single-label problems. The latter adapts the existing classifier to handle the MLC problems directly. In this study, we adopt the Label Powerset (LP) method [14]. LP transforms the existing multi-label problem into a traditional single-label multi-class one by treating each combination of the labels as a new class. Therefore, LP preserves the correlation between different labels.

Deep learning has brought unprecedented advances in natural language processing, computer vision, and speech processing [3], [4], [5]. In particular, multimodal deep learning is a new trend which has attracted an increasing interest in recent few years [2], [8]. Suk et al [15] proposed a multimodal feature representation and fusion with deep learning for Alzheimer disease diagnosis. Huang [16] presented a multi-label conditional restricted Boltzmann machine to handle multimodal data with missing modalities and fuse them to obtain the shared representation between the modalities. In a recent work [7], a Multi-modal Multi-instance Multi-label (M3) framework is proposed for complex object classification. That work also used independent deep learning models for each modality and imposed the consistency of data modalities on bag-level prediction. Different from the existing multi-label multimodal deep learning frameworks, in this paper, we focus on three different modalities (image, audio, and text), and also consider both spatial and temporal information in these modalities.

Imbalanced data classification is another important challenge in multimedia research that has been widely studied in multimedia data classification tasks such as fraud detection, disease diagnosis, and interesting event detection [9], [10]. The general solutions include data resampling (e.g., oversampling or undersampling) [11] and cost-sensitive learning [17]. The former changes the data distribution in a way to have similar numbers of samples in the minority and majority classes, while the latter penalizes the misclassification of the minority classes more than the majority ones. In the deep learning literature, the challenges of imbalanced data classification have not been thoroughly investigated. Few recent studies have focused on this problem by generating synthetic data [18] or changing the loss function to improve the detection performance of minority classes [19]. To the best of our knowledge, this work is the first framework for multi-label multimodal imbalanced data classification using deep neural networks.

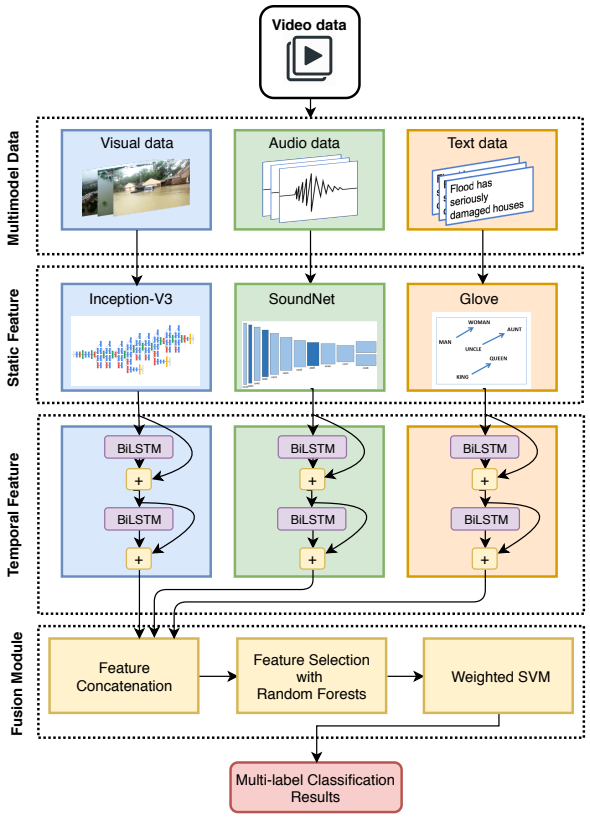

Figure 1. The proposed multi-label multimodal deep learning framework.

III. The Proposed Model

The proposed multi-label multimodal deep learning framework is shown in Figure 1. The input of our framework includes disaster videos which contain visual and audio clips as well as text descriptions. For each data modality, we extract static features using the state-of-the-art pre-trained deep learning models. In the next step, temporal features are extracted using the advanced Recurrent Neural Networks (RNNs). Then, in the fusion module, we concatenate the features from each modality and apply a Random Forest feature selection to remove the irrelevant features. Finally, the selected features are used as the input of the multi-label multimodal weighted Support Vector Machine (SVM) to generate the final classification results.

A. Static Feature Extraction Module

Static feature sets include visual, audio, and text features as explained below.

Visual Feature Extraction: In a video classification, visual data play an important role in detecting various concepts. In this paper, following our previous work [2], we used a well-known pre-trained deep learning model

called Inception-V3 [3] to extract the visual features from video clips. To do so, each video is subsampled to a fixed number of frames (40 frames in this case). Then, the features are automatically extracted from each frame using transfer learning. We used the last average pooling layer of Inception-V3 for feature extraction.

Audio Feature Extraction: Similar to our previous work [2], we utilized SoundNet [5] to extract the audio features. SoundNet is a pre-trained deep learning model that leverages the natural synchronization between audio and visual data. It is originally trained on two-million unlabeled videos by transferring the knowledge from vision to sound. The SoundNet network includes a series of one-dimensional convolutional networks followed by nonlinear activations such as ReLU. In our framework, we used the last convolutional layer (conv7) for audio static feature extraction.

Textual Feature Extraction: In comparison to visual and audio data, text data is capable of providing rich information which precisely describes various situations. By adding the knowledge learned by the textual model, the multimodal framework could capture important semantic information [4]. We extend our previous work by integrating textual data into our fusion framework. The data is extracted from the video description from all the videos used by the original dataset [2]. Preprocessing is performed to clean and format the textual data, which includes stop words and punctuation removal and tokenization. Then, the textual data is transformed into the vector space by using a pre-trained word embedding model called GloVe [20]. GloVe first learns a word co-occurrence counts matrix and generates the vector space representation based on the co-occurrence of each pair of words with a soft constraint:

$$\gamma_i^T \gamma_j + b_i + b_j = log(X_{ij})$$

where X_{ij} is the word pair i and j, γ_i and γ_j are the word vectors for words i and j, b_i and b_j are the biases term for words i and j. Then, the co-occurrence matrix is reduced to generate the final word vector. The objective of the cost function J is to penalize rare word pairs which carry less information:

$$f(X_{ij}) = \begin{cases} (\frac{X_{ij}}{X_{max}})^{\alpha} & \text{if } X_{ij} < X_{max} \\ 1 & \text{otherwise} \end{cases}$$

$$J = \sum_{i=1}^{V} \sum_{j=1}^{V} f(X_{ij})(\gamma_i^T \gamma_j + b_i + b_j - logX_{ij})^2$$

where V is the total number of words, $f(x)$ is the weighting function, X_{max} is the cutoff threshold, and α is the tunable parameter.

B. Temporal Feature Extraction Module

Video data includes a series of frames and there is valuable temporal information between the frames' sequences. This temporal information can be seen among the visual and audio frames as well as the video textual data. After we extract the static features, we extract the temporal features from each modality using deep RNNs. Specifically, we extend our previously proposed model called Residual-Bidirectional Long short-term memory (ResBiLSTM) [21] to extract the temporal features from not only the visual data but also audio and text data. Different from conventional RNNs, LSTM networks are capable of learning long-term dependencies of sequences of data and reducing the issue of vanishing or exploding gradients. However, one directional LSTM only considers the previous information in a sequence of data. This problem can be solved by Bidirectional LSTM which looks at both former and subsequent information using forward and backward paths. Finally, we used the idea of residual connections, which is originally proposed for enhancing the convolutional neural networks in image classification [22]. We add the residual connection to each LSTM layer to directly transfer the information from the upper layers and add them to the output of LSTM. This shortcut path will help transfer the spatial information through the temporal layers. The network parameters in ResBiLSTM are updated as follows.

$$h_l = \sigma(w_l h_{l-1} x + b_l) + h_{l-1}$$

where σ is the non-linearity function, h_l is the hidden layer at layer l, x is the input, and w_l and b_l refer to the weight and bias in this layer, respectively. As can be inferred from the above equation, the output of the previous layer (h_{l-1}) is added to the non-linearity results to generate the final output of the current layer (h_l). In this study, we only used two ResBiLSTM layers for each modality.

C. Fusion Module

The output from the ResBiLSTM network consists of segments of temporal features that contain the relevant information for various concepts. By incorporating the early outputs from the temporal networks, the semantic correlation from different modalities could be preserved and utilized. The overall fusion model is illustrated in Algorithm 1. The unimodal vector representation $\overrightarrow{v_i}$, $\overrightarrow{a_i}$ $\overrightarrow{t_i}$ from visual, audio, and text models are concatenated to form a single vector representation $\overrightarrow{f_i}$. Then, all the vectors $\overrightarrow{f_i}$ are grouped and formed the new dataset F based on the original ordering of the instances. The new vector has 384 dimensions that may cause various problems such as overfitting and slower training time. Therefore, dimensional reduction and feature selection techniques are applied. Random Forest (RF) is a tree-based ensemble learning algorithm that constructs multiple decision trees through the training phase and produces the final prediction score based on the majority vote of each classifier [23]. We use F as the input of RF classifier and calculate the mean decrease of Gini Impurity (GI) of each

feature. The GI is defined as:

$$GI = \sum_{j=1}^{|C|} P(j) * (1 - P(j))$$

where $|C|$ is the size of the concepts, $P(j)$ is the probability of an input be classified as class j. While training, the total decrease of Gini impurity for each feature is computed on the decision tree level. Then, the impurity decrease from each feature is averaged on the whole forest. Based on the mean decrease of Gini impurity, feature ranking R is generated. In real-world data, the distribution of the number of instances for different concepts may be heavily skewed. This imbalanced class problem could negatively impact the performance of the classifier since most of the machine learning models assume the classes' distribution are uniform. Thus, the cost function of SVM is modified to penalize the misclassification of instances that belong to the minority classes. The new cost function is defined as:

$$J = \frac{1}{\sum_j^{|C|} \delta_j} \sum_i^N \delta_j \cdot max(0, 1 - y_i(w_i^T \cdot x_i + b_i))$$

where δ_j is the inverse frequency of the number of instances containing class c_j, $|C|$ is the size of the concepts, N is the total number of instances, y_i is the label of i^{th} instance, x_i is the input instance, w_i and b_i are the learned weight and bias terms. The original multi-label ground truth L is transformed into the single-label form $\hat{L}$ using the label powerset algorithm. The weighted SVM is trained with the new ground truth label setup using the recursive feature elimination approach. This approach recursively drops the lowest ranked feature r_k in all the instances from input F based on the feature ranking R. During each iteration, the prediction result will be recorded and compared with the previous score. If the latest score is not improved then the previous best result (S) will be returned.

Algorithm 1 The proposed fusion algorithm

Input: Audio feature A, Video feature V, text feature T and ground truth label L
Output: Final prediction score S

$F \leftarrow \{\}$
for $\overrightarrow{a_i} \in A, \overrightarrow{v_i} \in V, \overrightarrow{t_i} \in T$ **do**
  $\overrightarrow{f_i} \leftarrow concatenate(\overrightarrow{a}_i, \overrightarrow{v}_i, \overrightarrow{t}_i)$
  $F \leftarrow F \cup \overrightarrow{f_i}$
end for
$R \leftarrow RandomForest(F)$
$IF \leftarrow \{\}$
$\hat{L} \leftarrow LabelPowerSet(L)$
for $c_j \in C, j = 1, 2, ..., |C|$ **do**
  $\delta_j \leftarrow \frac{1}{|F \in c_j|}$
  $IF \leftarrow IF \cup \delta_j$
end for
for $r_k \in R, k = |R-1|, |R-2|, ..., 1$ **do**
  $F \leftarrow F - r_k$
  $s_k \leftarrow WeightedSVM(IF, F, \hat{L})$
  if $s_k < s_{k-1}$ **then**
    $S \leftarrow s_k$
    return S
  end if
  $S \leftarrow s_k$
end for
return S

Table I
THE STATISTICAL INFORMATION OF THE DISASTER DATASET

No.	Concepts	# of Instances	P/N Ratio
1	Demo	150	0.047
2	Emergency Response	338	0.105
3	Flood/Storm	971	0.301
4	Human Relief	273	0.085
5	Damage	371	0.115
6	Victim	311	0.096
7	Speak/Briefing/Interview	811	0.251
	Total	3,225	

IV. Experiments and Analysis

A. Dataset Description

The data used in this paper is based on the dataset collected and used in our previous work [2]. The original dataset contains 1,540 video and audio clips that are extracted from 419 Youtube videos related to 2017 hurricane Harvey and Irma. We extend the original dataset by 1) adding text (extracted from the video descriptions) as a new modality, and 2) transforming the original single label problem into a multi-label problem. The statistics information of the disaster dataset is shown in Table I.

B. Experimental setup

Different metrics are required to evaluate the performance of MLC compared to those used in the single label classification. In the literature, several metrics have been adopted [24]. The evaluation metrics applied for our proposed framework include Hamming Loss, micro-averaged F-measure and mean average precision.

The Hamming Loss (HL) represents the proportion of the misclassified labels to the total number of labels.

$$HL = \frac{1}{|N|} \sum_{i=1}^{N} \frac{Y_i \oplus \Theta_i}{|C|}$$

where N is the total number of samples, $|C|$ is the total number of concepts, Y_i is the ground truth label, Θ_i is the prediction results, and $\oplus$ is the binary logical "exclusive or" operator. Micro averaged F-measure (MicroF1) calculates the micro-averaged F1-score of all classes by counting the global True Positives (TP), False Negatives (FN) and False Positives (FP) across all classes. Mean Average Precision

(MAP) calculates the average of the Average Precision (AP) over all the instances.

The dataset is randomly split into 60% training, 20% validation and 20% testing. In addition, we keep the distribution of classes almost similar between training, validation, and testing datasets. All model parameters are tuned using the validation dataset. The total numbers of static features for visual, audio, and text are 2048, 1024, and 1000, respectively. The temporal feature extraction model is composed of two bidirectional residual LSTM layers with 10% dropout, one dense layer using the ReLu activation function with 50% dropout and the final dense layer using Sigmoid activation function. The binary cross entropy is used as the cost function for the network training. For the weighted SVM classifier in the fusion model, a linear kernel is applied, a 0.9 penalty parameter for the error term is used and the shrinking heuristic is enabled.

C. Experimental Results

To demonstrate the effectiveness of the proposed multi-label multimodal deep learning framework, it is compared with several baselines as follows. Single visual, audio, and textual models including static features from Inception-V3, SoundNet, and Glove, respectively, each combined with a dense layer for classification. The second group of baselines includes the combinations of two different modalities (e.g., visual+audio, visual+text, text+audio). We also compared the proposed framework with two different fusion techniques including early fusion and late fusion. In early fusion, the static features are concatenated and then we apply LSTM to generate the temporal features followed by dense layers to generate the final scores. On the other hand, the late fusion concatenates the temporal features from each modality and apply the dense layers for classification.

Table II shows the detailed performance results of the baselines and the proposed framework. It can be seen from the table that the single text models perform better than the visual and audio models. Specifically, text model achieves 0.78 and 0.69 micro F1 and MAP, respectively. The visual model also achieves a reasonable performance which is significantly higher than the audio model. These results illustrate the importance of textual and visual data in event detection and disaster information management applications. In the next step, every two various modalities are combined to generate the classification results. Surprisingly, the audio+text model achieves the highest performance (micro F1 of 0.86) among all these three combinations. This is mainly due to the fact that audio and text can complete each other better than visual+audio or visual+text. For example, audio can easily detect concepts "speak" and "flood", but it cannot perform well for "damage" or "human relief" concepts, while text performs well in such concepts.

Finally, we used all the three modalities to further improve the results. It can be inferred from the table that simply

Table II
PERFORMANCE EVALUATION RESULTS ON THE DISASTER DATASET

Approach	Features	Micro F1	HL	MAP
Single modal	visual	0.6767	0.1586	0.6015
Single modal	audio	0.5022	0.2565	0.4197
Single modal	text	0.7789	0.1187	0.6945
Two modalities	visual+audio	0.6667	0.1652	0.5928
Two modalities	visual+text	0.823	0.0969	0.7472
Two modalities	text+audio	0.8586	0.078	0.7882
Three modalities (early fusion)	visual+audio +text	0.812	0.102	0.7351
Three modalities (late fusion)	visual+audio +text	0.9022	0.0575	0.8409
Proposed framework	visual+audio +text	0.9414	0.0348	0.8993

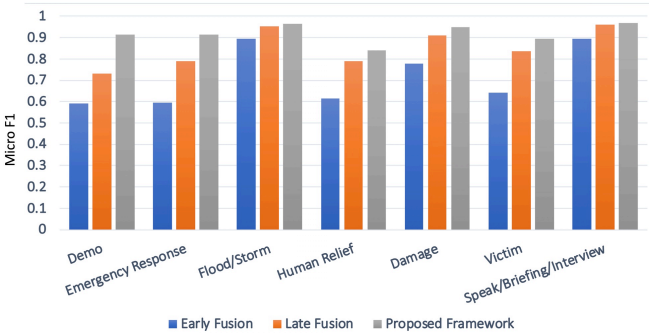

Figure 2. Performance comparison between the fusion models

concatenating the static features using early fusion cannot improve the classification performance compared to the two modalities models. This is mainly due to the different nature of the feature sets that cannot be easily combined in the early levels. However, if the features are fused in the final levels (after applying the temporal module), we can achieve a significant improvement in the final performance (e.g., 0.90 micro F1). Finally, we further improve the performance by applying our proposed fusion technique which includes late fusion followed by RF feature selection and a weighted SVM for imbalanced data classification. As a result, we could beat all the benchmarks. Specifically, the micro F1, HL, and MAP reach 0.94, 0.03, and 0.90, respectively. In other words, the F1 score is improved by 4% and MAP is improved by almost 6% compared to the best result (late fusion).

We further demonstrate the effectiveness of the proposed framework in Figure 2, in which our framework is compared with the other two fusion techniques (early and late fusions). This figure visualizes the micro F1 results for each concept in the disaster dataset. It can be observed from the figure that the proposed framework beats early and late fusions in all the concepts. For a few concepts such as "speak" and "flood", the late fusion's performance is very close to the ones from our method. However, in other concepts such as "demo" and "emergency response", there is a big gap between our performance and other fusion techniques. As shown in Table I, these concepts have lower P/N ratios

compared to "speak" and "flood". Therefore, the proposed framework can successfully enhance the performance of the minority classes without scarifying the majority ones. In summary, the proposed multi-label multimodal imbalanced data classification framework achieves an outstanding performance for a very challenging and complex dataset. The best performance for this dataset was 0.715 micro F1 for a single-label classification task that is reported in [2].

V. Conclusion

This paper presents a new multi-label multimodal framework based on deep neural networks for imbalanced data classification. The proposed framework includes static feature extraction for each modality using transfer learning, temporal feature analysis using ResBiLSTM, and a new fusion module which considers the correlation between both data modalities and labels. The proposed framework also handles the imbalanced data problem by automatically assigning a weight to each class during the classification. This framework is evaluated using a new dataset containing natural disaster videos. It will be also extended in the future to be trained end-to-end for all the modalities and also be evaluated on several larger multimodal datasets.

Acknowledgment

This project is supported in part by NSF CNS-1461926 and the Dissertation Year Fellowship award from Florida International University's Graduate School.

References

[1] V. Ranjan, N. Rasiwasia, and C. Jawahar, "Multi-label cross-modal retrieval," in *IEEE International Conference on Computer Vision*, 2015, pp. 4094–4102.

[2] S. Pouyanfar, Y. Tao, H. Tian, S.-C. Chen, and M.-L. Shyu, "Multimodal deep learning based on multiple correspondence analysis for disaster management," *World Wide Web*, 2018.

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[4] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5005–5013.

[5] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems*, 2016, pp. 892–900.

[6] H. E. Manoochehri and M. Nourani, "Predicting drug-target interaction using deep matrix factorization," in *IEEE Biomedical Circuits and Systems Conference*, 2018, pp. 1–4.

[7] Y. Yang, Y.-F. Wu, D.-C. Zhan, Z.-B. Liu, and Y. Jiang, "Complex object classification: A multi-modal multi-instance multi-label deep network with optimal transport," in *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2594–2603.

[8] H. Tian, Y. Tao, S. Pouyanfar, S.-C. Chen, and M.-L. Shyu, "Multimodal deep representation learning for video classification," *World Wide Web*, 2018.

[9] S. Pouyanfar, Y. Tao, A. Mohan, H. Tian, A. S. Kaseb, K. Gauen, R. Dailey, S. Aghajanzadeh, Y.-H. Lu, S.-C. Chen *et al.*, "Dynamic sampling in convolutional neural networks for imbalanced data classification," in *IEEE Conference on Multimedia Information Processing and Retrieval*, 2018, pp. 112–117.

[10] S. Sadiq, Y. Yan, M.-L. Shyu, S.-C. Chen, and H. Ishwaran, "Enhancing multimedia imbalanced concept detection using vimp in random forests," in *IEEE International Conference on Information Reuse and Integration*, 2016, pp. 601–608.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[12] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[13] ——, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.

[14] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Random k-labelsets for multilabel classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 7, pp. 1079–1089, 2011.

[15] H.-I. Suk, S.-W. Lee, D. Shen, A. D. N. Initiative *et al.*, "Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis," *NeuroImage*, vol. 101, pp. 569–582, 2014.

[16] Y. Huang, W. Wang, and L. Wang, "Unconstrained multimodal multi-label learning," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1923–1935, 2015.

[17] N. Thai-Nghe, Z. Gantner, and L. Schmidt-Thieme, "Cost-sensitive learning methods for imbalanced data," in *International Joint Conference on Neural Networks*, 2010, pp. 1–8.

[18] G. Douzas and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks," *Expert Systems with applications*, vol. 91, pp. 464–471, 2018.

[19] Q. Dong, S. Gong, and X. Zhu, "Imbalanced deep learning by minority class incremental rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[20] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543.

[21] S. Pouyanfar, S.-C. Chen, and M.-L. Shyu, "Deep spatio-temporal representation learning for multi-class imbalanced data classification," in *IEEE International Conference on Information Reuse and Integration*, 2018, pp. 386–393.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[23] H. Eslami Manoochehri, S. S. Kadiyala, J. Birjandtalab, and M. Nourani, "Feature selection to predict compound's effect on aging," in *ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2018, pp. 419–427.

[24] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas, "Multi-label classification of music into emotions," in *The International Society of Music Information Retrieval*, vol. 8, 2008, pp. 325–330.