# Hierarchical and Multimodal Classification of Images from Soil Remediation Reports

Korlan Rysbayeva, Romain Giot, Nicholas Journet

# Hierarchical and Multimodal Classification of Images from Soil Remediation Reports*

Korlan Rysbayeva, Romain Giot, and Nicholas Journet

Univ. Bordeaux, Bordeaux INP, CNRS, LaBRI, UMR5800, F-33400 Talence, France
{korlan.rysbayeva,romain.giot,nicholas.journet}@u-bordeaux.fr

**Abstract.** When soil remediation specialists clean up a new site, they have a long time manually revising digital reports previously written by other experts, where they look for necessary information in accordance with similar characteristics of polluted fields. Important information lies in tables, graphs, maps, drawings and their associated captions. Therefore, experts have to be able to quickly access these content-rich elements, instead of manually scrolling through each page of entire reports. Since this information is multimodal (image and text) and follows a semantically hierarchical structure, we propose a classification algorithm that takes these two constraints into account. In contrast to existing works using either multimodal system or hierarchical classification model, we explore the combination of state-of-the-art methods from multimodal systems (image and text modalities) and hierarchical classification systems. By this combination, we tackle the constraints of our classification process: small dataset, missing modalities, noisy data, and non-English corpus. Our evaluation shows that the multimodal hierarchical system outperforms the unimodal and that the performance of multimodal system with a joint combination of hierarchical classification and flat classification on different modalities provides promising results.

**Keywords:** Multimodal Classification · Hierarchical Classification · Operational Constraints

## 1 Introduction

Hazardous chemicals have been regularly used, utilized and spilled for decades without considering long-term issues in commercial and industrial facilities [27]. Thus, regulation services of various countries monitor the state of soil pollution for health purposes of citizens. Environmental protection agencies, companies or parties are responsible for cleaning up soil contamination. This is a laborious process which includes steps such as risk evaluation, laboratory experiments, pilot tests, field tests and further observations [10]. After performing each of these stages, specialists describe their observations and analysis of the obtained results with texts, diagrams, tables, aerial maps and photographs in a dedicated report [17] that can span from thirty pages to hundreds. Nevertheless, they are organized

* This work is supported by Abai-Verne scholarship and Innovasol Consortium.

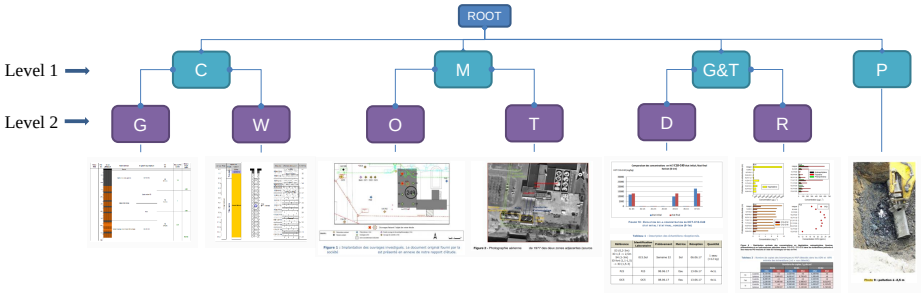


**Fig. 1.** *Hierarchical Tree* defined by soil remediation specialists that consist of 2-levels hierarchy. The samples illustrated for the $2^{nd}$ level of classes, according to the hierarchical tree also relates to the corresponding $1^{st}$ level of classes.The distribution of classes is following. **Level 1 classes:** C - cross section (505), M - maps (171), G&T - graphs and tables (416), P - photos (47). **Level 2 classes:** G - geology (195), W - well (310), O - one time (81), T - temporary (90), D - description (209), R - results (207).

in similar formats and some parts may contain similar semantic hierarchical plan. Experts use experience and knowledge from previous reports to clean up a new area or re-clean a current one. Due to the problem of finding similarities between two remediation zones in terms of different characteristics from huge data source and also from specific data, it can only be personally carried out by experts. Our long-term objective is to provide a tool to assist them during structuring the huge data to search for necessary information and similarity of technical documents.

The valuable information is kept in figures and their corresponding texts. It is convenient firstly to classify the elements (figures, texts) in the document by relevant categories, which will save the time in accessing and navigating the information, instead of searching them manually. This article illustrates the first step of this objective, and presents a method for classifying the content of the different documents parts in accordance of pre-established hierarchical plan given by soil remediation specialists (Fig. 1). Our purpose is to correctly classify the data for the last level of hierarchy ($2^{nd}$ level). Our proposal combines state-of-the-art method for hierarchical classification [3,2,26], and multimodal classification [16,29]. We classify the data independently of each other by adding a constraint that respects the hierarchy. Moreover, to classify the data for the classes that belong to $2^{nd}$ hierarchy level, which are more semantically defined, it is more convenient to additionally use text information. We are dealing with a small data source: about 1.2K images and 0.5K corresponding captions from 35 reports written by various companies with an average volume size of 50 pages. Considering that reports were provided by different experts with their inherit way of delivering information, there are constraints that includes missing modalities (some images provided without caption) and noisy data. The reports are written in French language so it is impossible to use mainstream NLP networks.

To our knowledge, several works exist on multimodal classification (*e.g.*, [25,14]) and hierarchical classification (*e.g.*, [31,28]), but there are very few

works [6,23] dedicated to joint combination of multimodality and hierarchy of the classes, and even fewer works related to classification task [6]. The purpose of this article is to show the possibility of applying the combination of multimodality system and hierarchical classification on standardized data. Secondly, to show that it outperforms the individual application of multimodality system and hierarchical classification. The novelty of this work lies in comparing the multimodal hierarchical classification with multimodal classification that considers the relationship of classes as independent. Moreover, the current paper studies how the model deals with missing modality aspect and illustrates how the different fusion techniques affect the performance of hierarchical classification. Specifically, the way to fuse the modality representations affects the performance of local, global or combined approach in hierarchical classification (Section 2.3).

Section 2 presents the previous work. Section 3 describes the proposed methodology. Section 4 specifies the protocol of the conducted experiment and Section 5 contains the results of the experiment, Section 6 concludes.

## 2 Previous work

There are few literature on MultiModal Hierarchical Classification (MMHC). We emphasize methods related to multimodal systems or hierarchical classification. Due to the nature of our application, we focus on those dedicated to classification of text and images, classification of small and imbalanced databases.

### 2.1 MultiModal Hierarchical Classification

Specificity of MMHC lies on both the *fusion strategy*, which belongs to modality processing and the *classification approach*, which deals with the hierarchy. The SemEval-2020 Task 8 [19] proposed three tasks: Sentiment Classification, Humor Classification and Scales of Semantic Classes for a dataset with memes of images and their corresponding short text. Das *et al.* [6] have considered these tasks as a hierarchy of classes and proposed the multi-task learning system, which combines the image feature block (ResNet [8]), text feature block (bi-LSTM [9] and GRU [4] with contextual attention) to learn all three tasks at once. They used early fusion strategy by concatenation of feature vectors to aggregate the feature block from two modalities to create task-specific features. Aggregated feature vectors passed on to smaller networks with dense layers, each one assigned to predict a sole type of fine grain sentiment label. The hierarchy dependency is managed by transfer learning of knowledge between the levels of hierarchy.

### 2.2 Multimodal Classification

Existing research on multimodal systems is focused on how to influence the information from multimodal feature spaces to get better performance than from their single modality counterparts. The fusion strategy can be roughly classified into *early fusion* (at feature level) and *late fusion* (at score level) [16].

*Early fusion* creates a joint representation of input features from multiple modalities, which is subsequently processed by a single model. The straightforward way of integrating a text with image features is a simple concatenation of their feature vectors. Wang *et al.* have used trainable CNN-RNN architecture for highlighting the meaningful text words and image regions in the Text-Image Embedding network (TieNet) [25], which is applied for classification of chest X-rays by using image and text features extracted from reports. Joint learning concatenated the two forms of representations and it uses final-fully connected layer to produce output for multi-label classification. There are also more complex fusion techniques, such as *gated summation* (using the gate value to weigh the summation of modality representations to create the fusion vector) or *bilinear transformation* (the filter that integrates the information of two vectors is concatenated with modality representations to create the fusion vector) [12,30].

*Late fusion* methods use a specific model on each modality and then combine the outputs to the final outcome. They integrate scores delivered by the classifiers on various features through a fixed combination rule commonly based on principle of indifference, treating all classifiers equally. The weighted averaging combines classifiers by evaluating optimal weights for each of them. Other approaches are Borda count rule, where the classifiers rank classes by giving each class the points corresponding to the number of classes ranked lower. Majority vote rule defines output by the largest number received among classifiers [11]. Separately extracting the embedding of the text and image views using pre-trained networks for classification task, Narayana *et al.* used simple weighted averaging to fuse softmax scores from image and text representations [16]. Yang *et al.* acquired the final prediction by mean-max pooling the bag-concept layer of different modalities [29]. This bag concept layer for modalities was obtained by dividing the raw articles into two modal bags - images and text paragraphs - and by calculating each modality with different networks.

### 2.3 Hierarchical Classification

In flat classification (FCl), the classifier is learned from labeled data instances without information about the semantic relationships between the classes. In contrast in hierarchical classification (HCl) models deal with hierarchy dependency of the labels in classification task. All existing state-of-the-art models on HCl can be divided for several approaches [26,1].

*Local approach* does the classification by traversing the hierarchy in top-down or bottom-up manner, learning from parent or sub-classes, accordingly. Each classifier is responsible for prediction of particular nodes or particular level of hierarchy, later combining this local prediction to generate the final classification. Fine-tuning, as a local approach, transfers the parameters from upper level to lower-level hierarchy classes while training and fine-tunes the parameters of lower levels to avoid the training from scratch. Hierarchical structure of categories for multi-label short text categorization is effectively utilized by transferring

parameters of CNN trained in the upper levels to the lower levels [21]. Zhang *et al.* transferred the Ordered Neuron LSTM (ONLSTM) [20] parameters from upper level to lower level for training, and fine tuned parameters of ONLSTM between adjacent layers. They used two gates mechanism and new activation function that constructed the order and hierarchical differences among neurons. To deal with multi-label models, which suffer on categories with few training examples, Banerjee *et al.* from parent category classifiers initialized the parameters of child category and later fine-tuned the model in HTrans recursive strategy for hierarchical text classification [3].

*Global approach* treats the hierarchical classification as a flat multi-label classification problem, where the authors use a single classifier for all classes. *Hidden-layer initialization* approach works by initializing the final hidden layer with a layer where the co-occurrence of labels in different hierarchy is shown [2,1]. For each occurrence, the value $w$ assigned to associated classes and value *0* assigned to the rest. The motivation for this assignment is to incline units in hidden layer by triggering only the corresponding label nodes in the output layer when they are active. *Hierarchy-aware attention mechanism* was used by Zhou *et al* [33] to deal with the hierarchy of the classes, where the initial label embedding is directly fed into a treeLSTM [24] as input vector of specific nodes, then the output hidden state represents as the hierarchy aware label features. Furthermore, sum of product of attention value and text representation is calculated to obtain the label aligned text features. *Deep hashing* was considered using the semantic hierarchy for large-scale image retrieval problem [32]. Authors handled the hierarchy of image classes by firstly mapping the image to binary code using deep hashing model, where the class center (*e.g.*, mean value) is calculated for the lowest hierarchy level. Secondly, they updated the class center for all levels in hierarchy, which depend on the ground-level class center.

*Combined approach* [26] follows the hybrid method by simultaneously optimizing both local and global loss functions. The global loss tracks the dependency of labels in the hierarchy as a whole. Each local loss function enhances the propagation of gradients leading to the correct encoding of local information between classes that corresponds to the hierarchical level. Moreover, to ensure the predictions, which obey the hierarchical structure, the hierarchy violation penalty was presented. The model was trained on 21 datasets related to protein function prediction, medical images or text classification.

### 2.4 Discussion

Multimodal systems and hierarchical classification has been mostly studied separately using large volume of data. Even if the number of classes is large, the number of samples for the classes in deeper levels of the hierarchy is usually sufficient. Most works on text focused on the English-based corpus. Some literature related to handling the multimodal systems considers the missing modality aspect, but focused on cross-modal retrieval task [23].

## 3 Combined Multimodal and Hierarchical Approaches for Technical Document Content Classification

We focus on multimodal classification with hierarchy dependency of classes and study how different fusion techniques of modality representations affect the performance of hierarchical classification and show a way of improving state-of-the-art methods by using the combination of joint multimodal classification system and hierarchical classification system. Hierarchical part considers both global and local approaches [26]. In multimodal system, the concatenation (early fusion) [6] of feature representations, weighted averaging rule [16] as simple fusion rules and the mean-max pooling (late fusion) [29] are chosen as fusion techniques for text and image modalities. As this work concerns data with few training samples, where text modality can be missing, and labeling is hierarchical we revised various works of the literature tackling these constraints [26,11,29,23]. They are not directly usable in our context as they consider either hierarchy of classes [26], or the multimodality aspect [11,29] of data, but not the combination of them. Some works consider the same constraints but focuses on cross-modal retrieval task [23]. Consequently, the adaptations of state-of-the-art methods for our problem are necessary.

### 3.1 Proposed Architecture

The architecture of the model consists of 3 parts. (i) The raw data is processes through feature extraction stage to obtain a compact representation that contains enough information for the classification process. This representation is classified with the (ii) hierarchical and (iii) multimodal parts. Depending on the fusion technique, precisely early fusion or late fusion techniques which explained in Section 2.2 the hierarchical model applied after or before multimodal system respectively. This section presents feature extraction process, the hierarchical classification model and finally discusses the multimodal fusion techniques.

*Feature extraction* process for *Image modality* resizes the input images to $224 \times 224$ pixels (Fig. 2), then passes it to VGG16 [22] using pre-trained weights from ImageNet [7] dataset; this part of the network is kept frozen. The obtained vector goes through FC layer with 1024 dimensions and is used as input for classifier layers. *Text modality* extraction process uses bag-of-words (BoW) technique, precisely term frequency–inverse document frequency (TF-IDF) [18] to extract features from captions, as our preliminary experiments illustrated their superiority to camemBERT [13]. The vector obtained by TF-IDF has around 2500 dimensions and passed through FC layer with 512 dimensions.

*The hierarchical part* considers the application of the feed-forward (HMCN-F) [26] for both image and text modalities. Different combinations of layers' dimensions were tested, and finally it was decided to use 512, 256 nodes for image modality and 256, 128 nodes for text modality for each Layers 1, Layers 2 and Layers 3 (Fig. 2). The following explanation focuses on image modality, but they

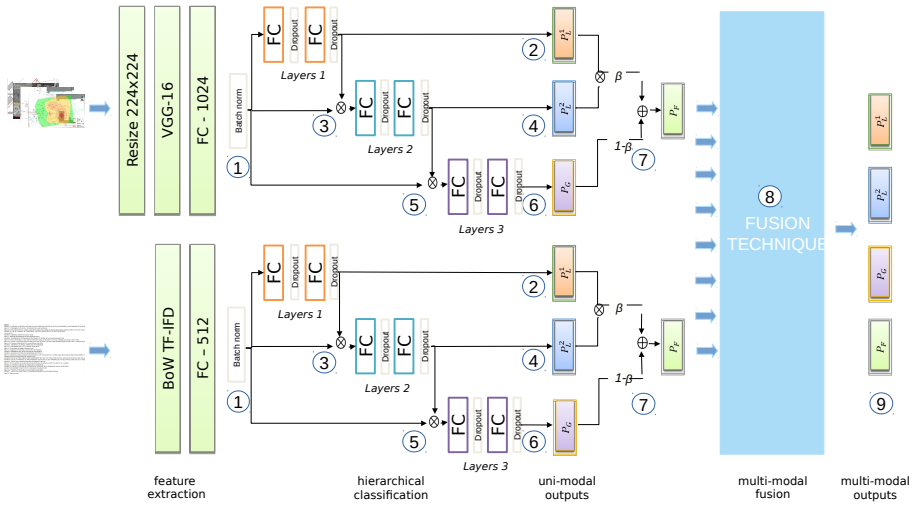


**Fig. 2.** The pipeline of architecture of multimodal (MM) hierarchical classification (HCl) model used during this work. The image representation passes through VGG16 pre-trained model, lately fed to the layers of hierarchical classification model ①. The text modality representations follow the same architecture except it was trained from scratch. HCl model based on the combination of local ②,④ and global approaches [26] ⑥. The hierarchy of classes handled by fine-tuning approach ③,⑤. The final score is the linear combination of two vectors from hierarchical approaches ⑦. The multimodal outputs is received after the fusion of representatives of modalities ⑨. Figure illustrates the late fusion technique of image and text modalities [16,29] ⑧; early fusion slightly simplifies the pipeline using a single branched network from ① (Section 3.1).

are similar for text modality. The feature vector of image modality is passed through Layers 1 ①. Using softmax function, the local prediction $P^1_{L,image}$ for the 4 classes of the $1^{st}$ level of hierarchy is calculated ②. At the same time, the output of the last dense layer of Layers 1 is concatenated with the feature vector, which is present to Layer 2 in the architecture of the model ③. Using softmax function, the local prediction $P^2_{L,image}$ for the 6 classes of $2^{nd}$ level of hierarchy is calculated ④. The last dense layer of Layer 2 is concatenated with feature vector, which was followed by passing through Layer 3 ⑤. Since global approach considers hierarchical classification as flat multi-label classification problem, 2 softmax function used for each hierarchy level to get the probability for the global prediction $P_G$(image) for overall 10 classes ⑥. The final prediction for 10 classes is calculated by Equation 1, where $\beta$ regulates the importance of local and global information from the class hierarchy ⑦ and $\otimes$ represents the concatenation.

$$P_{F,image} = \beta \left[ P^1_{L,image} \otimes P^2_{L,image} \right] + (1 - \beta)\, P_{G,image} \tag{1}$$

*The multimodal system* chosen fusion methods correspond to early fusion technique with 9.6M trainable parameters (*i.e.*, concatenation [6]) and late fusion technique with 6.5M trainable parameters (*i.e.*, weighted averaging [16], mean-max pooling [29]). During the experiments, each fusion technique is separately applied for the representatives of modalities. For late fusion, the fusion technique is applied for image and text representatives after being individually classified as explained in *Hierarchical part*. On the other hand, for early fusion, the fusion technique applied for feature vectors.

The weighted averaging is applied after the hierarchical model; it is the mean of two softmax tensors coming from the image and text modalities. Apart from the softmax tensors obtained for each modality $P^1_{L,image}$, $P^2_{L,image}$, $P_{G,image}$, $P_{F,image}$ and $P^1_{L,text}$, $P^2_{L,text}$, $P_{G,text}$, $P_{F,text}$, the model also calculates the corresponding 4 *multimodal* outputs, $P^1_L$, $P^2_L$, $P_G$ and $P_F$ ⑨. As one of the late fusion techniques, the mean-max pooling also utilizes the row-wise max pooling for the softmax tensors of image and text modalities, by firstly stacking the tensors beforehand, and consequently receives the corresponding 4 multimodal outputs. The concatenation pipeline differs from the illustration shown in Fig. 2 by transforming the fusion technique ⑧ after the feature extraction for image and text modalities respectively. Consequently, after hierarchical classification the model obtains only 4 multimodal outputs, $P^1_L$, $P^2_L$, $P_G$ and $P_F$ ⑨.

### 3.2 Loss function

The optimizer minimizes the sum of the local ($L^1_L$, $L^2_L$) and global ($L_G$) loss functions with the categorical cross-entropy loss. To guarantee the consistency of hierarchical path minimizing the proposed loss function is insufficient: to penalize predictions with hierarchical violation is required [26]. The hierarchical loss appears when parent and child classes are connected. The violation of hierarchy happens when $2^{nd}$ level class score is higher than the score of its parent class. Since there are more classes in $2^{nd}$ level, the score is supposed to be lower for each class. The characteristic equation $x(p, c)$ is determined to define the connection between the hierarchy levels, where $p$ indicates the class number in $1^{st}$ hierarchy level, $c$ the class number in $2^{nd}$ hierarchy level. $x(p, c) = 1$, if $p$ and $c$ connected, $x(p, c) = 0$, if $p$ and $c$ is disconnected. The hierarchical violation loss is calculated for each sample $i$ when holds the statement $Y^i_{c,2} > Y^i_{p,1}$ and exist the connection between parent and child classes:

$$L_H = \lambda \sum_i \sum_{p,c} max\{0, Y^i_{c,2} - Y^i_{p,1}\}^2 * x(p, c) \tag{2}$$

where for sample $i$ $Y^i_{c,2}$ represents the *child score* of class $c$ and $Y^i_{p,1}$ represents the *parent score* of class $p$. $\lambda \in R$ is employed for regulating the importance of the penalty for hierarchical violations in the overall hierarchical loss function. The final loss function for whole system is optimized by

$$L_F = L^1_L + L^2_L + L_G + L_H \tag{3}$$

where $L_F$ is final loss, $L^1_L$, $L^2_L$ and $L_G$ is the loss for local and global networks.

## 4 Experimental Protocol

This section gives details of the experimental procedure and describes the evaluation of interest targeted during the experiment.

### 4.1 Operational Dataset

The system presented in Section 3 was tested on a real world but private dataset. Confidentiality issues does not allow to reveal this dataset. To our knowledge, public dataset that contain multi-modal data and hierarchically defined classes does not exist. We have created our dataset by extracting images and captions from 35 reports related to soil remediation procedure. Around 500 images and their corresponding captions were automatically extracted [5] and supplemented by a manual extracting of 700 additional images with no caption. The extracted images have dimensions from $100 \times 100$ to $2000 \times 2000$ pixels, whereas the average caption length is around 44 words. Moreover, since the dataset of interest was created on the basis of reports from French companies, the considered text dataset (captions) is based on French. Soil remediation experts have labeled the images. According to Fig. 1, the dataset is spread across 2 hierarchical levels defined by the soil remediation specialists. The $1^{st}$ level consists of 4 classes, the $2^{nd}$ level consist of 6 classes. The samples distribution is shown in Fig. 1. To get the thorough results, the parent class *Photos*, which does not have the subclasses was transmitted for the $2^{nd}$ level hierarchy, thereby the $2^{nd}$ level hierarchy artificially consists of 7 classes, and global level consists of 11 classes.

*Train/Validation* and *Test* datasets are generated with a 10-folds stratified cross-validation that keeps the class distribution of the full dataset in each split of the data; on its turn *Train* and *Validation* dataset divided as 80% and 20%.

### 4.2 Hyper-parameters configuration

Adam optimizer uses a learning rate of $10^{-3}$; batches contains 16 samples; batch normalization layers are applied after first FC layer (Fig. 2); Dropout layers (0.6) are used after remaining FC layers to mitigate overfitting. Wehrmann *et al.* [26] considered two types of fusion parameters. $\beta$, which correlates the influence of local or global approach on final result (Fig. 2), and $\lambda$, which correlates the importance of hierarchical violation during calculation of loss function (Equation 2). The value of these parameters is chosen in accordance of the best performance on the dataset of interest. For Image modality system the parameters are $\beta = 0.5$ and $\lambda = 0.1$, for Text modality system the parameters are $\beta = 0.1$ and $\lambda = 0.1$.

### 4.3 Evaluation metric

Similar to the most recent works on hierarchical classification [1,21] we report weightedF1 scores for our experiments. F1 scores consider the precision ($P$) and recall ($R$) of the categories and changes between $[0, 1]$, while weightedF1 computes

the average of the F1 scores considering the number of instances $GT_{class}$ of each class $c$.

$$weightedF_1 = \frac{\sum_c^C (F_1(c) * GT_{class}(c))}{\sum_c^C GT_{class}(c)}, F_1(c) = 2 * \frac{P(c) * R(c)}{P(c) + R(c)} \quad (4)$$

where $C$ is the total number of classes per hierarchy level, $GT_{class}$ is the total number of samples per class. The weightedF1 score is calculated independently for $1^{st}$ and $2^{nd}$ hierarchy levels in local, global and combined approaches.

### 4.4 Evaluations of interest

Several questions arise and are treated in independent evaluations.

*Evaluation 1: How state-of-the-art multimodal fusion techniques perform combined with hierarchical classification for the given dataset?* Three multimodal fusion techniques are combined with hierarchical classification model: weighted averaging (late fusion) weighting the modality representations equally, mean-max pooling (late fusion) and concatenation (early fusion). After receiving the results on Evaluation 1, the weighted averaging fusion of modality representation was chosen for the remaining evaluations (Section 5, Table 1).

*Evaluation 2: Which hierarchical approach (local or global) needs to be investigated in future works?* We have compared the performances of hierarchical local approach with hierarchical global approach (fine-tuning) while using the weighted averaging fusion technique for modality representations.

*Evaluation 3: What is the best strategy to handle the missing modalities?* Multimodal system requires the presentation of a sample of each modality, but captions maybe absent for some images. We have compared two strategies to overcome this issue: replace the missing samples with empty vectors or fallback to the monomodal system of the available modality.

*Evaluation 4: How the multimodal system performs with the combination of hierarchical classification (HCl) and flat classification (FCl) models?* We firstly compare the multimodal system combined with HCl and FCl systems separately, to study which modality performs better in given context, then to validate results, the multimodal system tested along with the combination of hierarchical classification and flat classification systems, taken the best of each modality performances on previous tests.

## 5 Results and discussion

The results illustrated are computed globally by fusing 10 folds results (Section 4.1). The performance is shown for *local* and *global* approaches separately. The *final* performance represents the performance of combined approach (Fig. 2, ⑦). The statistical significance of the results performed by Mann-Whitney rank test [15] on all folds, where the results of $p$-values outline superiority of one specification to another. Any sample in the dataset is assigned with one class

**Table 1. weighted F1** : Comparison of different fusion techniques of modality representations for hierarchical classification model. Early fusion shows better performance than the late fusion techniques. MM indicates to multimodal system.

| | **Late fusion** | | | | | | **Early fusion** |
|---|---|---|---|---|---|---|---|
| | Weighted averaging | | | Mean-max pooling | | | Concatenation |
| | Image | Text | MM | Image | Text | MM | MM |
| Local level1 | 95.88 | 78.68 | **96.85** | 95.51 | 80.79 | 41.70 | **97.63*** |
| Global level1 | 95.61 | 78.07 | **96.40** | 96.48 | 76.89 | 38.39 | **97.54*** |
| Final level1 | 96.06 | 78.66 | **96.66** | 94.53 | 78.45 | 80.27 | **97.80*** |
| Local level2 | 79.27 | 50.56 | **83.26** | 76.34 | 49.26 | 13.27 | **83.59*** |
| Global level2 | 79.05 | 51.25 | **83.28*** | 72.03 | 26.57 | 19.38 | **83.09** |
| Final level2 | 79.17 | 50.79 | **83.10** | 71.53 | 29.75 | 56.71 | **83.95*** |

from each level of the hierarchy. The local and global approach consist of two classifiers each. Thereby the performance is illustrated in accordance with the results obtained in Level 1 and Level 2. Using Equation 1, the final results is supposed to be shown for 11 overall classes, but to be consistent with the illustration of results in local and global approaches, the final result is split for Level1 and Level2 categories. The early fusion (concatenation) technique creates a joint representation of image and text representations. Then the concatenated vector is used for the hierarchical classification model. Thereby the scores for Image and Text modality results is absent, and only MM classification is present.

*The combination of MM and HCl systems are studied (Evaluation 1, Evaluation 2).* Table 1 describes the performance of the tested multimodal fusion techniques considering hierarchical structure of labels. The concatenation fusion technique outperforms the rest probably because of the difference in number of learning weights, for Level 1 and Level 2 final scores with 97.80% and 83.95% respectively. Mann-Whitley test has confirmed that with $p$-value = 0.041 for Level 2. The mean-max pooling MM results shows that this fusion technique performs poorly for the tested dataset, whereas weighted averaging late fusion technique performs almost the same as early fusion technique ($p$-value = 0.079 for Level 1). The combination of local and global approaches works only for early fusion technique, considering the outperforming result (97.80% and 83.95%) of final scores in hierarchy levels. For the late fusion technique, the local approach performance is better for the $1^{st}$ level of hierarchy. In contrast, for the $2^{nd}$ level of hierarchy local and global approaches perform similarly. Furthermore, the MM outperforms the unimodality performances for both hierarchy levels, which shows that fusion of information from two modalities (Image, Text) can increase the results. Since the concatenation cannot show the performance of each modality separately and the mean-max pooling does not work for the dataset of interest, the weighted averaging fusion of modality representation was chosen for the remaining experiments.

**Table 2. weighted F1** : Comparison of MM HCl system performances. *Empty vectors* corresponds to the performances of the system where the missing modality data is replaced with empty vectors and *Image fallback* corresponds to the performance of the system that fallback to the monomodal image system when text modality is unavailable.

| | **MM HCl** (Weighted averaging) | | |
|---|---|---|---|
| | **Unimodal** | **Multimodal** | |
| | Image only | Empty vectors | Image fallback |
| Local level1 | 95.88 | **96.85** | **97.68*** |
| Global level1 | 95.61 | **96.40** | **96.50*** |
| Final level1 | 96.06 | **96.66** | **97.80*** |
| Local level2 | 79.27 | **83.26** | **83.40*** |
| Global level2 | 79.05 | **83.28** | **83.69*** |
| Final level2 | 79.17 | **83.10** | **83.22*** |

*The missing modality constraint is studied (Evaluation 3).* Table 2 compares the performance of the image monomodal system to two variants of the multimodal system: use of an empty vector when text modality is missing or fallback to the image monomodal system otherwise. Unimodal system represents the results where the missing modality information does not influence the performance of the model. However Table 2 illustrates that the performance of the model is close for multimodal system, and unimodal performance shows worse results for both hierarchy levels (97.80% over 96.06% and 83.22% over 79.17%). This indicates that the text modality enhances the overall performance, nevertheless artificially complementing the missing text representatives with empty vectors. This ensures that empty vectors force the system to take decision with no information.

*The multimodal system performance with combination of Flat Classification (FCl) and Hierarchical Classification (HCl) systems is studied (Evaluation 4).* Firstly, we tested the multimodal system coupled with the HCl and FCl systems separately, to define best modality performance in each case. In the FCl system the pipeline of architecture changes by eliminating the concatenation of input vectors with the output from each network related to hierarchy approaches (Fig. 2, ③, ⑤), thereby the networks was trained independently. Secondly, the multimodal system is tested on the combination of HCl model on text modality and FCl model on image modality. Table 3 shows that the HCl model performs better on text modality (78.66% and 50.79%), FCl model performs better on image modality (96.91% and 81.33%). Moreover, the multimodal system tested on FCl model slightly outperforms multimodal system tested on HCl model because of simplicity of hierarchy classes with $p$-value = 0.018, except for Level 2 in local and global approaches ($p$-value = 0.104). However, the novel results were delivered by the application of multimodal system with combination of FCl and HCl models, which for modality representation overpass the performances on both modalities (97.71% and 86.00% for respective hierarchy level with ($p$-value = 0.014). This implies that HCl model can contribute to multimodal

**Table 3. weighted F1** : Comparison of Flat classification (FCl) with Hierarchical classification (HCl) performance. The best results of each modality is represented in bold: Image modality achieves highest result under FCl system, Text under HCl system, while multimodality (MM) outperforms under combination of FCl and HCl systems.

| | **FCl** | | | **HCl** | | | **FCl + HCl** | | |
|---|---|---|---|---|---|---|---|---|---|
| | Image | Text | MM | Image | Text | MM | Image (FCl) | Text (HCl) | MM |
| Local Level1 | **96.45** | 67.25 | 97.36* | 95.88 | **78.68** | 96.85 | 96.93 | 67.70 | **97.11** |
| Global Level1 | **96.20** | 67.02 | 97.10 | 95.61 | **78.07** | 96.40 | 97.11 | 68.62 | **97.19*** |
| Final level1 | **96.91** | 67.57 | 97.18 | 96.06 | **78.66** | 96.66 | 97.19 | 68.09 | **97.71*** |
| Local level2 | **79.43** | 44.93 | 82.34 | 79.27 | **50.56** | 83.26 | 82.24 | 45.65 | **84.61*** |
| Global level2 | **79.59** | 43.81 | 82.16 | 79.05 | **51.25** | 83.28* | 80.20 | 46.71 | **82.97** |
| Final level2 | **81.33** | 45.09 | 84.90 | 79.17 | **50.79** | 83.10 | 82.90 | 46.06 | **86.00*** |

system performances, therefore this results acts as a foundation for perspective future research on this area. Additionally, the proposed method is competitive against humans as it completes classification in 30 minutes whereas the manual classification of complete dataset takes 30 hours.

## 6 Conclusion and Perspectives

This paper proposes a multimodal and hierarchical classification framework of information contained in soil remediation reports. The main contribution of this work is the proposal of an efficient combination of state-of-the-art methods to overcome specific constraints: small dataset, missing modalities, noisy data and non-English corpus. The most relevant results are that the multimodal system enhances the performance of the unimodal systems, and regarding the hierarchical approaches, the combination of local and global approaches only works with concatenation of modality representations. For the late fusion of representations, the local approach needs to be investigated for future works. Moreover, early fusion technique for multimodal system performs better than the late fusion technique for our dataset. The multimodal system combined with hierarchical classification model performs worse than with flat classification model. However, multimodal system with hierarchical classification model for text modality and flat classification model for image modality performs better than the rest.

During this study, the trained hierarchical classification system applies both local and global hierarchical approaches, instead of having two separate networks that consider either local or global approaches, which supposedly had an impact on the obtained result. Thereby, for the future work, it is interesting to analyze the performance using these distinct networks. Moreover, in order to study whether the classifier assigns labels to classes that do not conform with the hierarchy of the categories, it is convenient to compare the post-processing label correlation [1,2].

## References


1. Aly, R., Remus, S., Biemann, C.: Hierarchical multi-label classification of text with capsule networks. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. pp. 323–330 (2019)
2. Baker, S., Korhonen, A.L.: Initializing neural networks for hierarchical multi-label text classification. In: Proceedings of the BioNLP workshop. pp. 307–315 (2017)
3. Banerjee, S., Akkaya, C., Perez-Sorrosal, F., Tsioutsiouliklis, K.: Hierarchical transfer learning for multi-label text classification. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 6295–6300 (2019)
4. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1724–1734 (Oct 2014)
5. Clark, C., Divvala, S.: PDFFigures 2.0: Mining figures from research papers. In: 2016 IEEE/ACM Joint Conference on Digital Libraries. pp. 143–152. IEEE (2016)
6. Das, S.D., Mandal, S.: Team neuro at semeval-2020 task 8: Multi-modal fine grain emotion classification of memes using multitask learning. arXiv preprint arXiv:2005.10915 (2020)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. IEEE (2009)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. arxiv 2015. arXiv preprint arXiv:1508.01991 (2015)
10. Hyman, M., Dupont, R.R.: Groundwater and soil remediation: Process design and cost estimating of proven technologies. p. 367–422. American Society of Civil Engineers (2001)
11. Kittler, J., Hatef, M., Duin, R.P., Matas, J.: On combining classifiers. IEEE transactions on pattern analysis and machine intelligence pp. 226–239 (1998)
12. Lu, D., Neves, L., Carvalho, V., Zhang, N., Ji, H.: Visual attention model for name tagging in multimodal social media. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1990–1999 (2018)
13. Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, , Seddah, D., Sagot, B.: Camembert: a tasty french language model. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020)
14. Masakuna, J.F., Utete, S.W., Kroon, S.: Performance-agnostic fusion of probabilistic classifier outputs. In: 2020 IEEE 23rd International Conference on Information Fusion (FUSION). pp. 1–8. IEEE (2020)
15. McKnight, P.E., Najab, J.: Mann-Whitney U Test, pp. 1–1. American Cancer Society (2010)
16. Narayana, P., Pednekar, A., Krishnamoorthy, A., Sone, K., Basu, S.: Huse: Hierarchical universal semantic embeddings. arXiv preprint arXiv:1911.05978 (2019)
17. Pastor, J., Gutiérrez-Ginés, M.J., Bartolomé, C., Hernández, A.J.: The complex nature of pollution in the capping soils of closed landfills: Case study in a mediterranean setting. In: Environmental Risk Assessment of Soil Contamination, pp. 199–223. IntechOpen, Rijeka (2014)

18. Sammut, C., Webb, G.I. (eds.): Encyclopedia of Machine Learning, chap. TF–IDF, pp. 986–987. Springer US, Boston, MA (2010)
19. Sharma, C., Bhageria, D., Paka, Scott, W., P Y K L, S., Das, A., Chakraborty, T., Pulabaigari, V., Gambäck, B.: SemEval-2020 Task 8: Memotion Analysis-The Visuo-Lingual Metaphor! In: Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020) (2020)
20. Shen, Y., Tan, S., Sordoni, A., Courville, A.: Ordered neurons: Integrating tree structures into recurrent neural networks. arXiv preprint arXiv:1810.09536 (2018)
21. Shimura, K., Li, J., Fukumoto, F.: HFT-CNN: Learning hierarchical category structure for multi-label short text categorization. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 811–816 (2018)
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Computational and Biological Learning Society pp. 1–14 (2015)
23. Sun, C., Song, X., Feng, F., Zhao, W.X., Zhang, H., Nie, L.: Supervised hierarchical cross-modal hashing. In: Proceedings of the 42nd International ACM SIGIR on Research and Development in Information Retrieval. pp. 725–734 (2019)
24. Tai, K.S., Socher, R., Manning, C.D.: Improved semantic representations from tree-structured long short-term memory networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1556–1566 (2015)
25. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9049–9058. IEEE (2018)
26. Wehrmann, J., Cerri, R., Barros, R.: Hierarchical multi-label classification networks. In: International Conference on Machine Learning. pp. 5075–5084 (2018)
27. Wuana, R.A., Okieimen, F.E.: Heavy metals in contaminated soils: A review of sources, chemistry, risks, and best available strategies for remediation. Heavy Metal Contamination of Water and Soil: Analysis, Assessment, and Remediation Strategies p. 1 (2014)
28. Xue, H., Liu, C., Wan, F., Jiao, J., Ji, X., Ye, Q.: Danet: Divergent activation for weakly supervised object localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6589–6598 (2019)
29. Yang, Y., Wu, Y.F., Zhan, D.C., Liu, Z.B., Jiang, Y.: Complex object classification: A multi-modal multi-instance multi-label deep network with optimal transport. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2594–2603 (2018)
30. Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D.: Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE transactions on neural networks and learning systems (Vol.29) (12), 5947–5959 (2018)
31. Zhang, Q., Chai, B., Song, B., Zhao, J.: A hierarchical fine-tuning based approach for multi-label text classification. In: 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA). pp. 51–54. IEEE (2020)
32. Zhe, X., Ou-Yang, L., Chen, S., Yan, H.: Semantic hierarchy preserving deep hashing for large-scale image retrieval. arXiv preprint arXiv:1901.11259 (2019)
33. Zhou, J., Ma, C., Long, D., Xu, G., Ding, N., Zhang, H., Xie, P., Liu, G.: Hierarchy-aware global model for hierarchical text classification. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1106–1117 (2020)