MACQUARIE
University
SYDNEY·AUSTRALIA

Survey
on
Multiple Instance learning

Arian Kalantari
48549819

Macquarie University

Abstract: This survey presents an overview of Multiple Instance Learning (MIL), a weakly supervised framework suited for unstructured and partially labelled data. MIL addresses scenarios where labels are available only at the bag level, not for individual instances. We introduce core MIL concepts, compare instance-space and bag-space methods, and real-world applications in domains such as medical imaging and video anomaly detection. We then deep dive into medical imaging to review recent advances such as attention-based, graph-based, and contrastive MIL models along with key datasets like CAMELYON16. This report highlights both the strengths and challenges of current MIL approaches and outlines directions for future research to enhance scalability and interpretability.

Keywords: Multiple Instance Learning, Weakly Supervised Learning, Instance-Space Method, Bag-Space Method, Medical Imaging, Video Anomaly Detection, CAMELYON16, Interpretability, Scalability, Transformer MIL, Graph Neural Networks.

Survey Objectives

This survey aims to explore the foundation and practical relevance of Multiple Instance Learning (MIL), a weakly supervised learning framework suited for unstructured and partially labelled data. The objectives are:

- To explain core MIL principles, including its structure, assumptions, and challenges.
- To compare MIL approaches such as instance-space and bag-space methods.
- To review key real-world applications, with a focus on medical imaging.
- To evaluate recent models, highlighting strengths and challenges.
- To outline future research directions

Introduction

Conventional machine learning and feature-based models are preferred for when the dataset is well-structured and pre-defined in pretty much a table format [1]. These models also struggle to handle complexity as well as high dimensional data as the number of features increase [1]. However, it is estimated that nearly 80 percent of the real-world data is unstructured data—images, videos and audios—data types that do not conform to traditional rows and columns [2]. Integrating such diverse sources demands "distinct preprocessing techniques and model architectures" tailored to each data type, which goes beyond the scope of standard algorithms [3].

The technique Multiple Instance Learning (MIL) has gained its popularity to address these challenges. In MIL, training data are not composed of single, labelled instances, but instead are structured as labelled bags containing multiple unlabeled instances. A bag is labelled positive if it contains at least one positive instance and negative if all instances are negative. This makes MIL a form of weakly supervised learning, where the model must infer relevant patterns from coarse-grained annotations without access to individual instance labels [4].

MIL is used in many fields from medical image analysis to document analysis. For example, in medical image analysis like cancer detection, a patient is diagnosed either malignant or benign. This is similar to the MIL bag setting, an image is labelled as 1 or 0, but the cancer cell is not marked. Patches from the image are called instances, and bags are images or patients with the label. Then these labelled images are used for the model to distinguish the prominent features through weakly supervised learning process. MIL shines when labels are not on a specific patch (area), and most of the time in real-world scenario it is hard to get well-labelled and pre-processed dataset for machine learning project as annotation cost increases [4]. In short, this is why MIL has gained its popularity due to the compatibility with big unstructured data.

Concept of MIL

In MIL, training data are not individual labeled instances, but bags of instances grouped together with a bag-level label. Classic assumption when working with MIL is that bags are labelled positive if there is at least one positive instance and labelled negative if all instances are negative [4]. Model is given a bag-level labels to infer the patterns at instance level to identify the likely key instances within those bags. Instead of relying on exact annotations for every instance, the paradigm leverages the assumption. In this sense, MIL is a form of weakly supervised learning as MIL treats instance labels as latent variables and learn the relationship between the instance and bag label [1,4]. This property leads to several unique challenges such as label ambiguity, instance-level inference, scalability and assumption mismatch [4].

Firstly, as there is an absence of instance-level supervision in the training settings, MIL often suffers from determining which instances are responsible for the bag label, and this is called label ambiguity. Moreover, this ambiguity critically effects on instance-level inference interpretability as well as classification. Secondly, when the size of bags increases, meaning that the dataset becomes bigger, the computational complexity increases hence there is a scalability challenge with MIL. Lastly, the assumption mismatch could occur if the requirements for tasks we are solving with MIL are not carefully validated [4]. For example, bags could be labelled with a combination of two instances such as labelling images that have beach must need sand instance and water instance, and this violates the classical assumption.

Medical Imaging:

As briefly mentioned in the previous section, medical imaging is a key application of MIL. The main task is to diagnose diseases or detect abnormalities in medical scans, such as predicting whether a patient has cancer given multiple labelled images. Typically, a large medical image—such as a whole-slide image (WSI)—is treated as a bag, with extracted patches or regions-of-interest considered as instances [5]. Since manually annotating every cancerous region is often expensive and impractical, MIL enables models to learn from bag-level labels by assuming at least one positive instance exists in a positive bag. MIL is widely applied in pathology for cancer detection [4,5].

Datasets like CAMELYON16 and CAMELYON17 contain WSIs labelled for metastases in breast lymph nodes and are commonly used to evaluate MIL-based models. A well-known example is TransMIL (Shao et al., 2021), which uses a Transformer to model both morphological features and spatial relationships among image patches. TransMIL achieved strong performance with ~93% AUC on CAMELYON16 and over 96% on TCGA lung cancer slides [5]. Recent MIL frameworks address the patch-level inference challenge by using attention mechanisms, graph-based models, or contrastive learning to highlight cancerous patches without needing instance-level labels, while maintaining interpretability through heatmaps [5,6,7].

Video Anomaly Detection:

As Sultani et al. (2018) has introduced, another application is anomaly detection in untrimmed surveillance videos [8]. Their approach uses a deep MIL framework which

considers videos as bags that are tagged as “Anomalous” or “Normal” without mentioning the time of the event and sets of temporal clips from the videos as instances. In the paper, there is another introduction of innovational approach—MIL Ranking Loss—which encourages the model to assign higher anomaly scores to the most abnormal instance in an anomalous video. This approach enabled the model to detect a short video clip as an anomaly by just using video-level labels. In this context, the MIL formulation fits naturally, as only weak supervision is available: a video might span several minutes, but the anomalous event occurs in just a few seconds. Traditional supervised learnings require frame-level annotation, which is time-consuming and costly, whereas MIL allows learning directly from video-level tags by assuming that at least one instance in a positive video must be anomalous [8].

The UCF-Crime dataset, released by Sultani et al. (2018), has become a benchmark for video anomaly detection using MIL. The dataset contains 1,900 real-world surveillance videos with labels “Normal” or 13 types of anomalies (e.g. fighting, robbery, road accidents). All videos are untrimmed and significantly differ in length and scene content, casting a challenging for MIL-based methods.

Their work has inspired several follow-up studies and improvements, including methods that incorporate temporal attention mechanisms, self-training [9,10].

Categories of MIL

MIL algorithms are generally grouped by its learning algorithms, meaning that there are different approaches between instance-space methods and bag-space methods. These approaches differ in their underlying assumptions, learning strategies, interpretability, and scalability, making each more suitable for certain applications [11].

Instance space method is the algorithm works at the instance level, attempting to distinguish between positive and negative instances within each bag. Once predictions are made for all instances, then aggregate those instance labels to derive bag label [11]. Typical aggregation methods are max-pooling and attention mechanism. Most of time, instance space approaches are associated with the standard MIL assumption - for example, identifying a witness positive instance that made the bag to be labelled positive. This assumption is particularly suited for tasks This assumption is well suited for tasks where just one small part of the data can contribute to the bag label—for example, when a patch in a large medical image shows a tumor. However, because they must evaluate and combine potentially many instances per bag, they often require more computation and can have lower scalability or even slightly worse accuracy in practice [11].

On the other hand, bag-space methods focus on learning from the bag as a whole. Instead of looking at each instance in the bag, the methods treat each bag as a single entity and derive a global representation—for example, by computing a summary statistic or embedding across all instances [12]. A standard classifier is then applied to these bag-level vectors. Bag-space methods are typically applied under collective assumptions, where all or many instances contribute to the bag label. The key idea is that bag’s label must be the event of similar instances being in the bag, not just one single

instance. This makes them well-suited for distributed evidence tasks where the signal emerges from aggregate patterns rather than a key instance—for example, identifying the latent patterns or textures across image.)

ASPECTS	INSTANCE-SPACE	BAG-SPACE
INTERPRETABILITY	Good at instance-level explanation, meaning that the model is good to identify the latent instance. Useful when explaining the prediction under standard assumption.	Limited interpretability. Difficult to trace predictions back to specific instances as the model treats the bag holistically.
SCALABILITY	Can be computationally expensive as the method needs to go over all instances, meaning that if the number of instances increases processing time and cost increase.	More scalable. Treats each bag as one sample after embedding, though very large or complex bags may still introduce time performance overhead.
MODELLING STRATEGY	Train an instance-level classifier, then aggregate outputs to label the bag. Under an assumption that there must be an instance contributing to the label, the model finds the latent instance.	Derive a feature embedding or representation of the bag. Train a standard classifier in that space.
CORE ASSUMPTION	Standard MIL assumption (at least one positive in a positive bag assumption.)	Collective assumption: many or all instances contribute to the bag's label.

In summary, instance-space methods keep instance-level interpretability and are effective when a sparse signal exists, while bag-space methods offer more scalability and are preferred when the label depends on distributed features. Type of the task, the data structure, and the underlying label-generation process determine which method to be adopted [11,12].

Review of MIL Used in Medical Imaging

In this section, we go through recent papers on MIL in medical imaging to highlight their approaches, results and practical impacts.

In pathology, Lu et al. (2021) presented an attention-based model CLAM (clustering-constrained attention MIL) that learns to weight and cluster patch features using only WSI labels. CLAM identifies most latent regions via attention and refines patch features with an instance-level clustering loss. This means that the updated weights can be used to produce interpretable heatmaps [5].

TransMIL, mentioned in the introduction section, introduced by Shao et al. (2021), adopts a Vision Transformer to model both spatial and morphological correlations among patches: it treats patches as "tokens" and uses self-attention to capture long-range dependencies, reporting high performance on tumor detection [13].

Graph-based approaches build an explicit connectivity among patches. For example, Zhao et al. (2020) constructed a graph whose nodes are patch features and edges link neighboring patches; graph convolutional networks then aggregate context to produce a bag-level prediction [14].

Contrastive MIL approach has emerged recently. Tavolara et al. (2022) presented an unsupervised contrastive MIL which uses two augmentations of patches from the same slide form a positive pair, while patches from different slides are negatives [6]. Training with a contrastive loss yields slide-level embeddings without any labels. In general, contrastive methods first learn rich patch features from unlabeled WSIs then fine-tune MIL classifiers, reducing the need for large annotated datasets.

METHOD	MODEL TYPE	DATASETS	REPORTED AUC
TRANSMIL	Transformer-based MIL	CAMELYON16, TCGA-NSCLC [#] TCGA-RCC [#]	93.1% 96.0% 98.8%
CLAM	Attention-based MIL	CAMELYON16, TCGA-RCC	≈95.3% >99%
GRAPH MIL	Graph-based MIL	Colon and prostate histopathology (varied)	Not consistently reported

Beyond raw accuracy, MIL models demonstrate key strengths. Interpretability is a major advantage as attention heatmaps generated based on the updated weights often align with pathologist-annotated tumor regions. For instance, Courtiol et al. presented that an attention-MIL network could "attend specifically to pathologist-annotated metastases" in CAMELYON16, despite not using RoI labels [15]. CLAM models similarly highlight known histological features for each subtype [5]. Generalisability seems to be promising too as Lu et al. (2021) reported that CLAM trained on one group of patients performed on an independent test set and even on smartphone microscopy images [5]. Campanella et al. (2019) showed MIL on a very large WSI dataset (prostate and skin cancer, >25,000 slides) "outperforms strong supervision on small datasets," indicating robustness to data variations [16]. In summary, modern MIL approaches yield high AUCs and are valued for their ability to localize tumor regions and to generalize across datasets, all while using only slide-level labels [4,5,13-16].

MIL models have clear clinical appeal (e.g. by training with only slide-level diagnoses, they reduce annotation cost and leverage existing pathology reports). Moreover, automated image classifiers can assist initial diagnosis (flagging images containing abnormality) and guide pathologists to regions of interest via attention maps [5,15]. For example, a MIL-based system could triage breast lymph node scans for metastases or quantify the extent of tumor without cell-level costly annotation [4,5,16]. These methods have also been applied to tasks like predicting gene mutations or receptor status from H&E slides, potentially obviating expensive molecular assays. In principle, MIL could streamline workflows. However, challenges remain. The lack of instance-level ground truth means evaluating patch-level predictions is difficult as bag-level evaluation is only available as standard metrics. On top of that, WSIs are extremely large, so computation and memory are bottlenecks, meaning that models often need sample or crop patches to reduce the computation cost results in risking missed information [4,5,16]. Domain shift is another issue; models trained on one organisation's slides may lack generalisation due to staining or scanner differences. Moreover, because positive patches are rare, MIL networks can overfit or miss small lesions. Finally, benchmarking is hard: while datasets like CAMELYON16 exist, diverse clinical tasks have no unified evaluation framework. In sum, MIL in medical imaging shines by leveraging weak labels and offering interpretability, but it is limited by weak supervision, data heterogeneity, and heavy computational demands [4,5,13–16].

Conclusion

We have explored the foundations, categories, and real-world applications of Multiple Instance Learning (MIL). MIL addresses critical limitations of traditional supervised learning by enabling models to learn from bag-level labels without in-detail instance-level annotation. The report outlined the key differences between instance-space and bag-space methods, highlighting their assumptions, interpretability, scalability, and their suitability for different problem types.

Through a focused review of applications in medical imaging, the survey listed how MIL has been effectively used in domains where labelling is expensive or impractical. In recent papers, attention-based models; CLAM, transformer architectures such as TransMIL, and graph-based methods have shown high accuracy on benchmark datasets like CAMELYON16 and TCGA datasets. Despite their advancement, current MIL methods face challenges in scalability, interpretability, and generalisation.

Looking ahead, continued progress in MIL is likely to involve hybrid models, domain adaptation strategies, and improvements in weak supervision techniques. Techniques such as reinforced learning could be applied together with the MIL. As unstructured data continues to grow across industries, MIL will remain a critical area of research for enabling robust, scalable machine learning systems that operate effectively in real-world settings.

References

1. Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.
2. Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007
3. Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021). *Dive into deep learning*. https://d2l.ai
4. Carbonneau, M.-A., Cheplygina, V., Granger, E., & Gagnon, G. (2018). Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition, 77*, 329–353. https://doi.org/10.1016/j.patcog.2017.10.009
5. Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., & Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering, 5*(6), 555–570. https://doi.org/10.1038/s41551-020-00682-w
6. Tavolara, T. E., Gurcan, M. N., & Niazi, M. K. K. (2022). Contrastive multiple instance learning: An unsupervised framework for learning slide-level representations of whole slide histopathology images without labels. *Cancers, 14*(23), 5778. https://doi.org/10.3390/cancers14235778
7. Fang, Z., Wang, Y., Zhang, Y., Wang, Z., Zhang, J., Ji, X., & Zhang, Y. (2024). MamMIL: Multiple instance learning for whole slide images with state space models. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (accepted). https://doi.org/10.48550/arXiv.2403.05160
8. Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018)* (pp. 6479–6488). https://openaccess.thecvf.com/content_cvpr_2018/html/Sultani_Real-World_Anomaly_Detection_CVPR_2018_paper.html
9. Zhu, Y., & Newsam, S. (2019). Motion-aware feature for improved video anomaly detection. In *British Machine Vision Conference (BMVC 2019)*.
10. Feng, J.-C., Hong, F.-T., & Zheng, W.-S. (2021). MIST: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)* (pp. 1620–1629). https://openaccess.thecvf.com/content/CVPR2021/html/Feng_MIST_Multiple_Instance_Self-Training_Framework_for_Video_Anomaly_Detection_CVPR_2021_paper.html
11. Shi, X., Xing, F., Xie, Y., Zhang, Z., Cui, L., & Yang, L. (2020). Loss-based attention for deep multiple instance learning. Proceedings of the AAAI Conference on Artificial Intelligence, 34(04), 5742–5749. https://doi.org/10.1609/aaai.v34i04.6030
12. Fatima, S., Ali, S., & Kim, H.-C. (2023). A comprehensive review on Multiple Instance Learning. Electronics, 12(20), 4323. https://doi.org/10.3390/electronics12204323

13. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., & Zhang, Y. (2021). TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. arXiv preprint arXiv:2106.00908. https://doi.org/10.48550/arXiv.2106.00908
14. Zhao, Y., Yang, F., Fang, Y., Liu, H., Zhou, N., Zhang, J., Sun, J., Yang, S., Menze, B., Fan, X., & Yao, J. (2020). Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4836–4845). https://doi.org/10.1109/CVPR42600.2020.00489
15. Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., ... & Wainrib, G. (2019). Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nature Medicine, 25(10), 1519–1525. https://doi.org/10.1038/s41591-019-0583-3
16. Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., ... & Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine, 25(8), 1301–1309. https://doi.org/10.1038/s41591-019-0508-1

Datasets:

17. CAMELYON16: Grand Challenge on Cancer Metastasis Detection in Lymph Nodes. (2016). https://camelyon16.grand-challenge.org/
18. CAMELYON17: Grand Challenge on Cancer Metastasis Detection in Lymph Nodes. (2017). https://camelyon17.grand-challenge.org/
19. TCGA: The Cancer Genome Atlas Program. (n.d.). National Cancer Institute. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
20. UCF-Crime: A large real-world dataset for anomaly detection in surveillance videos. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6479–6488. https://www.crcv.ucf.edu/projects/real-world/anomaly-detection-in-surveillance-videos/