IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 16 April 2024, accepted 11 May 2024, date of publication 21 May 2024, date of current version 10 June 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3403538*

SURVEY

# Multiple Instance Learning in Medical Images: A Systematic Review

**DALILA BARBOSA**[1,2], **MARCOS FERREIRA**[3], **GERALDO BRAZ JUNIOR**[3], **MARTA SALGADO**[4], **AND ANTÓNIO CUNHA**[1,2]

[1]Institute for Systems and Computer Engineering, Technology and Science—INESC TEC, 4200-465 Porto, Portugal
[2]School of Sciences and Technology, University of Trás-os-Montes and Alto Douro, 5000-801 Vila Real, Portugal
[3]Applied Computing Center, Federal University of Maranhão (UFMA), São Luis 65080-805, Brazil
[4]University Hospital Centre of Santo António, 4099-001 Porto, Portugal

Corresponding author: Dalila Barbosa (dalila.i.barbosa@inesctec.pt)

This work was supported in part by the National Funds through Portuguese Funding Agency, FCT—Fundacão para a Ciência e a Tecnologia, under Project PTDC/EEI-EEE/5557/2020; in part by European Union under Grant 101095359; and in part by the U.K. Research and Innovation under Grant 10058099. However, the views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible.

**ABSTRACT** This article presents a systematic review of Multiple Instance Learning (MIL) applied to image classification, specifically highlighting its applications in medical imaging. Motivated by the need for a comprehensive and up-to-date analysis due to the scarcity of recent reviews, this study uses defined selection criteria to systematically assess the quality and synthesize data from relevant studies. Focusing on MIL, a subfield of machine learning that deals with learning from sets of instances or "bags", this review is crucial for medical diagnosis, where accurate lesion detection is a challenge. The review details the methodologies, advances and practical implementations of MIL, emphasizing the attention-grabbing and transformative mechanisms that improve the analysis of medical images. Challenges such as the need for extensive annotated datasets and significant computational resources are discussed. In addition, the review covers three main topics: the characterization of MIL algorithms in various imaging domains, a detailed evaluation of performance metrics, and a critical analysis of data structures and computational resources. Despite these challenges, MIL offers a promising direction for research with significant implications for medical diagnostics, highlighting the importance of continued exploration and improvement in this area.

**INDEX TERMS** Images classification, medical images, multiple instance learning (MIL).

## I. INTRODUCTION

Within the evolving domain of machine learning research, the subfield of MIL is distinguished by its approaches and methodologies. MIL has found its place in many applications, from complex medical diagnostics to the dynamic world of video surveillance [1]. Its existence is due to its remarkable ability to increase accuracy and optimise the image analysis process, making it a key technology in several domains [2].

However, the MIL field is marked by various methodologies and a rapid advance of the research frontier. This diversity and progress require a careful and insightful exploration of state-of-the-art techniques to select the most effective and context-appropriate methods.

The adoption of MIL in medical images for lesion detection is motivated by several critical factors highlighting its relevance and effectiveness. Firstly, it addresses the inherent uncertainty in the location and characterization of lesions, allowing systems to learn to identify pathological patterns even in incomplete or ambiguous information [1]. This is particularly useful in cases where accurate lesion annotation is challenging, such as in large image datasets. In addition, MIL can effectively handle inter and intra-patient variability in lesion characteristics, thus improving diagnostic accuracy [3]. MIL's ability to process and analyse sets of instances (i.e., sub-regions of the image) makes it ideal for detecting subtle abnormalities, which are often crucial in early diagnosis and evaluating responses to treatment [3]. Finally, integrating MIL with the latest machine learning and artificial intelligence techniques promises to increase

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Zuo.

the efficiency of diagnostic processes and leverage a deeper understanding of the morphological characteristics of lesions, leading to significant advances in personalized medicine and clinical decision-making [4].

Analysing existing literature reviews on MIL applied to medical images, a considerable number of publications can be observed, although most of them were carried out before 2019. A review from 2022 was found [5], which only focuses intensively on medical images of the Whole Slide Images (WSI) type. However, it is imperative to recognize the diversity and importance of other medical imaging modalities in the context of MIL. The variety of applications and clinical relevance of images from other modalities underlines the need to expand research and analysis beyond the limited scope of WSI, thus covering a more comprehensive range of images. The aim of this review is, therefore, to study more recent methods of MIL in medical imaging, not restricting itself to WSI alone but also exploring other modalities, which could provide broader insights applicable to various clinical and research contexts.

This review employs a systematic approach guided by clear selection criteria to critically assess study quality and synthesize data. It aims to enhance our understanding of Multiple Instance Learning (MIL) in imaging by providing a comprehensive analysis that facilitates future research and practical applications. The review illuminates the landscape of MIL applications in image datasets, focusing on advanced techniques that drive success in the field. It covers three main areas: the characterization of MIL algorithms across various imaging domains, the evaluation of performance metrics, and the analysis of the data structures and computational resources each method requires. This analysis not only maps current MIL trends and innovations but also evaluates the most effective solutions for both research and practical implementation.

### A. CONTRIBUTIONS

MIL research in medical imaging has shown remarkable progress, reflecting a deepening understanding and application of this advanced technique. MIL methods, characterized by their ability to handle weakly labelled datasets and learn from multiple instances, have shown promise for lesion detection in medical imaging. These approaches allow for the efficient processing of large volumes of images, identifying complex disease patterns accurately and efficiently. In the state of the art, we find a diversity of MIL implementations, each exploring different aspects of machine learning and offering valuable insights for advancing the automatic analysis of medical images. This review highlights the most recent and impactful developments in the field of MIL, underlining how these innovations are transforming the diagnosis and assessment of injuries in medical contexts.

This study undertakes a systematic review of the current state of the usage of MIL in detecting lesions in medical images. Research questions are formulated to address the identified methods found and try to provide a comprehensive understanding of the problem and pave the way for future research by answering the following questions:

**Q1:** What are the main challenges and limitations of current MIL methods in image classification?

**Q2:** How does the quality and quantity of training data influence the performance of MIL methods?

**Q3:** How accurate are MIL methods on different types of image datasets?

**Q4:** What kind of MIL algorithms are used in image classification?

TABLE 1. Inclusion (IC) and exclusion criteria (EC).

| Criteria | Description |
|---|---|
| IC0 | Published since 2018 |
| IC1 | The title, abstract, or keywords match the search query |
| EC0 | Work not published in a refereed journal or conference |
| EC1 | Literature/Systematic Review |
| EC2 | Full text is not available |
| EC3 | Mentions MIL but does not approach it |
| EC4 | The title, abstract, or keywords do not match the search query |
| EC5 | The language used is not English |

## II. REVIEW METHODOLOGY

This review used a method that is widely used in this field, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [6], as it helps to formulate the research question, choose relevant studies, and present the results, ensuring transparency and quality in the research.

### A. INCLUSION CRITERIA

We established strict inclusion criteria for selecting articles to ensure a comprehensive and up-to-date analysis of MIL methods applied to image classification.

The articles included in this systematic review must fulfil all of the following criteria: only articles published on or after January 1, 2018, were considered. Five years was chosen to ensure the inclusion of information on the latest advances and emerging technologies in the field of MIL for image classification. Studies must mention the use of MIL in image classification in their abstract. A detailed discussion of MIL in the abstract indicates direct relevance to the review's topic of interest. Articles dealing with MIL in contexts unrelated to image classification were excluded. Also, articles with full text unavailable, restricted access, or insufficient information for detailed evaluation resulted in exclusion from the study.

These criteria were applied systematically to all the literature searches conducted in the selected databases. Strict adherence to these criteria ensures that the systematic review is focused and relevant and provides a representative view of the current state of the application of MIL methods in image classification.

### B. SEARCH STRATEGY

The search strategy for this systematic review was designed to capture a broad spectrum of relevant studies applying MIL methods to image classification. We included three well-established databases in this search strategy: IEEE Xplore, Scopus, and SpringerLink.

The search strategy for this review was to select keywords to filter literature significantly relevant to our area of research. Our search terms included ''MIL'', ''Image Classification'', and ''Medical Images''. To ensure comprehensive coverage, we incorporated variants of these terms, employing Boolean operators like ''AND'' and ''OR'' to capture studies that intersect MIL with image classification, and those addressing advanced techniques and specific applications within this sphere. This thorough search was conducted across multiple databases to encompass a wide range of scholarly papers, enriching our review's robustness.

The search was initially limited to results that included the keywords in the title or abstract. The titles and abstracts were then read to identify articles related to the review's aim. Relevant articles were then selected to obtain and analyse the full text. The search was complemented by a manual investigation of the references of the selected articles to identify additional studies that might have escaped the initial search strategy. This chain search process helped ensure the inclusion of relevant papers that might have been outside the direct scope of the defined keywords.

With this search strategy, we intended to carry out a comprehensive and representative review of the current literature on the use of MIL for image classification, thus contributing to a deeper understanding of this expanding field of research.

### C. EXTRACTION OF STUDY CHARACTERISTICS

As part of this systematic review, each selected article's characteristics were extracted to capture information for the comparative analysis of MIL methods applied to image classification or detection.

Special attention was paid to how the characteristics were extracted from the images, including details on pre-processing techniques, image descriptors, and dimensionality reduction methods. The models used for classification, be they specific neural network architectures or machine learning algorithms, were carefully identified. The evaluation metrics, fundamental for comparing the effectiveness of the methods, were catalogued and included but were not limited to precision, recall, F1-score, and the area under the curve (AUC).

The systematic approach adopted for data extraction provided a solid basis for subsequent analysis, ensuring a detailed overview and comprehensive understanding of MIL strategies in the image classification literature.

## III. RESULTS

After all these methodologies that were present before, and as shown in Fig. 1, initially, there were 734 papers found in the three different databases. From that, and using the different inclusion and exclusion criteria, the number decreased significantly, and in the end, 22 papers were selected for further analysis. These underwent the process of characteristics extraction, resulting in a summary of each, which will be presented below.

In [7], the authors ascertained a MIL model based on attention and the triple kernel with contrastive learning (TGA-MIL). Several datasets were used for evaluation, including the USBC Breast Cancer Dataset, the Colon Cancer Dataset, a set based on MNIST, and the DDSM Dataset for mammogram images. The kernel, a convolution matrix, plays a key role in manipulating and highlighting specific image features. Advancing on this technique, the triple kernel concept involves using three distinct kernels, each to enhance different aspects of the medical image. The kernel functions used were Laplace, Radial Basis Function, and Inverse Multiquadric. The next step involves generating an attention map by the applied kernel functions, which is essentially a weighting mechanism that highlights the regions considered most important. The model employed ResNet and Contrastive Learning of Visual Representation (SimCLR) in the feature extraction process. This self-supervised learning approach trains the feature extractor to identify meaningful aspects of unlabelled data. This method achieved 60.9% and 81.0% accuracy on the USBC Breast Cancer and Colon Cancer datasets, respectively.

In [8], a two-phase approach has been developed for classifying Whole Slide Images (WSI) of weakly supervised learning. It uses contrastive learning to train the feature extractor in the compression stage. The learning phase combines convolutional neural networks (CNN) and transformers to capture local and global information from the images. For classification, the article proposes a two-phase model called CWC-Transformer. This model includes a compression phase, where contrastive learning is used, and a learning phase, where CNNs and transformers are combined to analyse the images. Three datasets were used for evaluation: CAMELYON16, TCGA-LUNG, and MSK. In the CAMELYON16 dataset, the CWC-Transformer achieved an accuracy of 89.14% and an AUC score of 93.85%. On the TCGA-LUNG dataset, it achieved an accuracy of 85.94% and an AUC of 94.88%. The article also mentions some limitations. One is the high consumption of computing resources and memory due to the auto-attention mechanism, especially when working with high-magnification images. In addition, the contrastive learning approach can treat very similar patches as negative pairs, affecting the model's performance.

In [9], a method for detecting lung diseases using chest CT scans employing a MIL-based approach is introduced. The main goal is to enable a more comprehensive analysis of lung diseases, including detecting semantic patterns in the lungs and predicting the mutation status of the EGFR gene. The study uses MIL as its primary methodology. Within this framework, two bag generators are employed: the Radiomic Bag Generator and the Hounsfield Units Bag Generator, which differ in the complexity of the shapes of the instances and the nature of the features. It employs bag-based

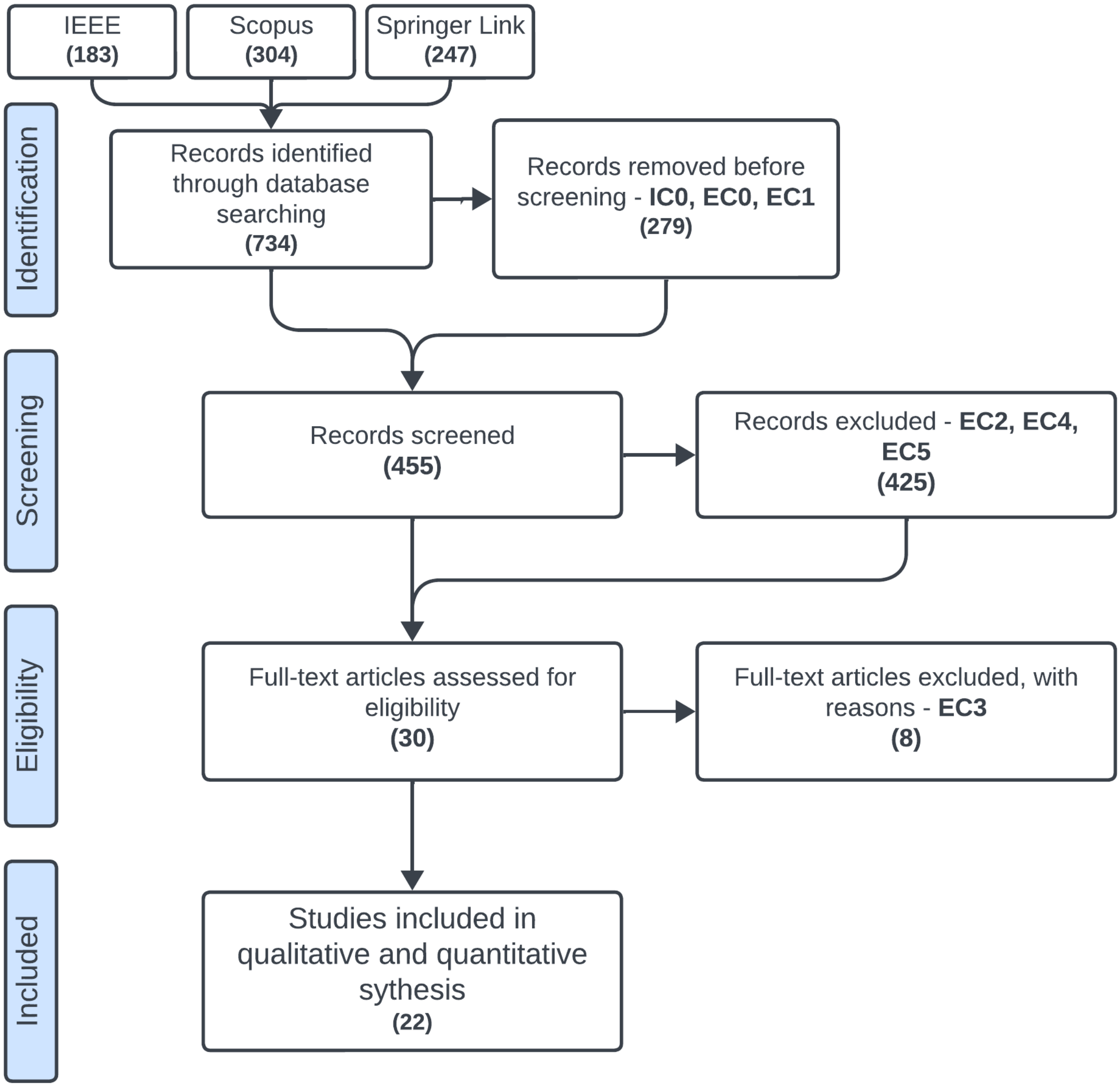
IEEE (183)
Scopus (304)
Springer Link (247)
Identification
Records identified through database searching (734)
Records removed before screening - IC0, EC0, EC1 (279)
Screening
Records screened (455)
Records excluded - EC2, EC4, EC5 (425)
Eligibility
Full-text articles assessed for eligibility (30)
Full-text articles excluded, with reasons - EC3 (8)
Included
Studies included in qualitative and quantitative sythesis (22)

**FIGURE 1. Process of the paper selection.**

classifiers for classification, focusing on the Normalised Set Kernel (NSK) algorithm from the MISVM package. The study uses three databases: the Non-Small Cell Lung Cancer Radiogenomics (NSCLC Radiogenomics) database, the Interstitial Lung Disease (ILD) database, and a private database from the Centro Hospitalar e Universitário de São João (CHUSJ). Fibrosis detection achieved an AUC of 0.89, while emphysema detection achieved an AUC of 0.72. A vital limitation of the study is the relatively small size of the databases. This limits the generalisability of the results and the application of more powerful deep learning methods.

In [4], it is presented a method for classifying medical images, called Dual Space Multiple Instance Representative Learning (DSMIRL). The main objective is to address the challenges faced in classifying medical images, particularly the imbalance between positive and negative instances in the images. DSMIRL includes two main components: Adaptive Instance Representative Selection (AIRS) and Multiple Instance Representative Learning (MIRL). AIRS uses clustering methods to select relevant subsets of instances (sub-bags), while MIRL performs aggregations in feature and label space for the final classification. The proposed model, DSMIRL, is a MIL method that integrates instance selection and aggregation in dual spaces (features and labels). It uses a ResNet50 network for feature learning and an attention module for aggregation in the feature space. DSMIR was evaluated on two sets of medical images (Camelyon16 and Pneumonia CT). In Camelyon16, it achieved 88.9% accuracy, 95.3% AUC, and 86.6% F1 score. Although the article demonstrates the effectiveness of DSMIRL, the main limitation is the complexity of the method and the need for large data sets for training.

In [10], the MIL method proposed incorporates a convolutional transform-based MIL anomaly classifier for the detection of polyps in colonoscopy videos. This method is particularly notable for its ability to operate with weakly-labelled videos, dividing each video into chunks. The main objective is to improve the learning of a MIL anomaly classifier for these chunks to optimise the accuracy of the anomaly scores at the chunk level. The classifier architecture is bifurcated into a temporal resource encoder that uses transform technology to capture the global temporal relationships between the sections and a MIL anomaly classifier responsible for generating each section's anomaly scores. To train this system effectively, a joint optimization approach is adopted that involves transform-based temporal feature learning, a contrastive snippet mining (CSM) technique, and a video classifier. The dataset used in the study is an extensive and diverse compilation of colonoscopy videos, the Hyper-Kvasir dataset, and the LDPolypVideo dataset. The methodology shows significant advances in polyp detection accuracy when compared to previous approaches, using metrics such as the area under the AUC curve and average accuracy. It achieved an AUC of 98.41% and 86.63% in the respective datasets. The classification distinguishes between normal videos and those with abnormalities, such as polyps.

In [11], a patch sampling strategy based on the sequential Monte Carlo method is proposed, specifically for the classification of histological images. MIL is implemented, where instead of labelling each patch individually, MIL treats the entire image as a single instance, assigning a general label based on the labels of the patches contained in it. This process uses neural network architectures adapted to the different data sets. For the MNIST set, an architecture like VGG with a receptive field of 40×40 pixels was used, while for the ICIAR and GTEx datasets, an architecture based on ResNet was chosen, with a receptive field of 224 × 224 pixels. In feature extraction, image points are initially sampled following a uniform distribution. Each point is then evaluated: a patch centered on the point is sampled and processed by the neural network, whose output represents the point. The points are then normalised between 0 and 1, and resampling is carried out, where points with low scores are discarded and new ones are sampled over those with higher scores, with a slight shift guided by a Gaussian distribution. As for the evaluation metrics, for the MNIST-Sparse set, the uniform sampling strategy achieved an average accuracy of 52.0%, while the Monte Carlo approach recorded a significantly higher accuracy of 82.5%.

The article [12], uses self-supervised autoencoders (MAE) for initialising the feature space. In addition, they also adopt a cluster-based resource distribution modelling method and a resource space refinement strategy based on pseudo-labels. The proposed framework, called DGMIL, employs a model to refine the latent resource space iteratively. This refinement is done through linear projection heads and a classification head, composed of a single fully connected layer. Two datasets were used: the CAMELYON16 dataset and the TCGA Lung Cancer dataset. For the CAMELYON16 dataset, the article shows an AUC of 0.9045 and FROC of 0.4887 for patch classification, an AUC of 0.8368, and an accuracy of 80.18% for slide classification. For the TCGA Lung Cancer dataset, they achieve an AUC of 0.9702 and an accuracy of 92.00% for slide classification. Common limitations in medical imaging studies include the need for large sets of annotated data, the variability in imaging techniques between different sources, and the generalisation of models to different types of cancer or tissue.

Reference [13] develops a method for classifying lung cytological images. It uses the Multiple Instance Deep Learning Algorithm (AD MIL) with an attention mechanism, Convolutional Neural Networks (CNNs) such as LeNet, AlexNet, Inception, ResNet, and DenseNet for feature extraction, and Otsu's automatic binarisation algorithm for image processing. Otsu's method is a thresholding technique used especially in situations where the object of interest and the background have different contrasts, playing an important role in the image preparation stage before being processed by the model. A dataset made up of images of lung cells was used. The best performance was obtained with the CNN AlexNet-like structure in AD MIL, achieving a classification accuracy of 91.6%.

Reference [14] focuses on analysing individual cells extracted from images by an R-CNN architecture and uses ResNet feature maps to perform classification without needing individual labelling of the cells. It uses a modified Mask R-CNN architecture with a ResNet backbone to detect individual cells in the images. It applies a CNN to classify each instance (cell) based on the weak labels of the sample. It employs an embedding level approach to MIL, where a unique representative for the set of instances is generated using a MIL clustering method, and based on the work of Ilse et al. it uses an attention mechanism to calculate a weighted average over the embeddings of instances, assigning weights learned by the neural network and associated Attention Pooling with Auxiliary Branch SIC. The study used blood samples from patients diagnosed with hereditary spherocytosis, part of the CoMMiTMenT study. The accuracy of the MIL model with attention pooling and SIC auxiliary branch was 79%, the F1 score was 78% and the AUC was 0.960. The number of cases studied was relatively small, requiring external validation to assess the practical effectiveness of the method. The study focused on specific genetic blood disorders, and its applicability to other diseases or conditions still needs to be investigated.

It presented in [15], a new Transformer model for classifying images of histopathology slides (WSI). It uses the K-means to extract a set of anchors based on the spatial clustering property of the features, EfficientNet-b0 to extract features from the WSI patches, and the cross-attention algorithm to communicate information between the kernels and the patch tokens. The study evaluates two databases: Gastric2K and Endometrial-2K. KAT achieved 98.3% accuracy in the subtyping task and 96.7% accuracy in the binary classification task on the Endometrial-2K dataset. In the Gastric-2K dataset, KAT achieved 98.3% accuracy in the subtyping task and 96.9% accuracy in the binary classification task. Limitations in studies such as this may include dependence on the quality and diversity of images in the database and the need for fine-tuning the model for different tissue types or diseases.

Reference [16] focused on developing a computer-aided diagnosis (CAD) system for diagnosing breast cancer through the classification of histopathological images. It uses a transfer learning approach and the selection of discriminative patches. The methodology employs algorithms such as CNN for extracting features from the patches, dendrogramming and clustering for selecting discriminative patches, and EfficientNet models for classifying the images. It also uses the Support Vector Machine (SVM) technique for classification based on extracted features. The study tested on the BreakHis dataset, achieving a maximum accuracy of 99.81%, 99.26%, 99.49% and 99.14% for binary classification at magnification levels of 40x, 100x, 200x and 400x, respectively, using EfficientNet-B7, and accuracy of 96.99%, 95.17%, 94.71% and 91.66% for multiclass classification at the same magnification levels. The limitations of the study include the dependence on the quality and variety of the images in the BreakHis dataset, the possibility of omitting relevant information due to the selection of discriminative patches, and the variability of performance depending on the specifics of the EfficientNet architecture and the SVM parameters.

In [17], it is demonstrated a method called SA-AbMILP (Self-Attention Attention-based MIL Pooling) for image classification using the MIL paradigm in a weak learning context. This method stands out for its ability to capture global dependencies between instances within a set, combining self-attention and attention-based pooling techniques to transform these instances into a fixed-size vector for classification. In this process, there are two distinct neural networks: the first is responsible for generating representations of the instances, while the second, comprising self-attention mechanisms, attention-based pooling, and a classifier, uses these representations to identify fungal species. To generate these representations, they recommend using deep architectures such as ResNet-18 and AlexNet, pre-trained on ImageNet. In addition to SA-AbMILP, other derived models, such as GSA-AbMILP, IQSA-AbMILP, LSA-AbMILP, and MSA-AbMILP, are explored in the article. These models were tested on several datasets, including MNIST, breast, and colon cancer histological datasets, DIFaS microbiological, and a retinal image screening set. Although the article does not explicitly discuss the limitations of the method, it is possible to infer some weaknesses, such as the dependence on the type of kernel used, issues of interpretability of the results, and the scalability of the method for larger data sets. These considerations highlight future research's importance in improving and adapting SA-AbMILP for different applications and data scales.

In [18], a graph-based multiple instance learning (GMIL) model is proposed for binary and multiclass classification tasks on unbalanced breast cancer datasets. The article uses a GMIL model, which employs a graph neural network (GNN) and a MIL framework. The model also incorporates a Gated Attention Module for efficient information fusion from nodes in a graph. In addition to the proposed GMIL model, the paper compares its performance with other high-performance deep learning models in recent years, including CSDCNN, Inception-V3, Inception-ResNet-V2, IRRCNN, BreastNet, and C-Net. The BreakHis dataset is used to evaluate the model. For binary classification, GMIL achieved an accuracy of 99.75% and an AUC of 99.69%. In multiclass classification, it achieved an accuracy of 96.40% and an AUC of 98.57%. These results demonstrate the superiority of GMIL compared to the other models evaluated. Although the article does not explicitly specify limitations, it is common for models such as GMIL to face challenges such as the need for large sets of labelled data, the computational complexity for large-scale graph processing, and the interpretability of deep learning models.

The main objective of [19] is the classification of WSIs at high magnification (40x), using a machine learning approach called TransMIL, which is a MIL model based on transformers, complemented by a new bag embedding loss (BEL). Within the methodology, the authors use the TransMIL model, which employs a recently proposed method called the Neystrom Method to approximate self-attention, allowing many instances to be processed. In addition, pre-processing is used, which includes patch extraction from WSIs and feature extraction using DenseNet-121. The data used includes two datasets: BRACS and CAMELYON17. TransMIL with BEL achieved an accuracy of 60.0% and an F1 score of 57.0%, while in CAMELYON17, it achieved 73.0% accuracy and 48.0% in F1 score. The article mentions that despite the improvements with BEL, there are still challenges due to poor annotation and the large size of the "bags" (sets of instances). In addition, the performance improvement seems to depend heavily on maximising the distance in the BEL loss term. This indicates that, although effective, the approach may still be sensitive to the quality of the class representations and the configuration of the BEL hyper-parameters.

[20] developed and validated an advanced machine learning model called Attention MIL with Transformer (AMIL-Trans) to classify breast cancer WSIs. The model aims to improve the selection of discriminating instances and the aggregation of bag-level features by integrating channel attention and self-attention. AMIL-Trans is combined with ResNet-50 with the efficient channel attention module (ECA) for selecting discriminating instances and a Transformer encoder for aggregation at the bag level. In addition, attention-based and Transformer methods are used to capture discriminant information and correlations

between instances. Two datasets were used: Camelyon-16 and MSK. AMIL-Trans achieved optimal AUC values of 94.3% on the Camelyon-16 dataset and 84.2% on the MSK dataset. Typical challenges in such approaches include the need for large amounts of labelled data for effective training, the computational complexity associated with processing large WSIs and generalising the model to different tissue types or pathological conditions beyond breast cancer.

It is shown in [21] a model called Shuffle Attention Multiple Instance Learning (SAMIL) for classifying breast cancer WSI. The methodology employs MIL and integrates the following algorithms and techniques: Shuffle Attention (SA) to capture pixel-level relationships and channel dependencies; Multi-Head Attention (MHA) and Long Short-Term Memory (LSTM), which are used to build an aggregator for instance features. ResNet with Shuffle Attention (SA) is used for instance selection and MHA and LSTM for bag-level prediction. The Camelyon-16 dataset was used. The SAMIL model achieved an accuracy of 96.0%, AUC of 95.3%, precision of 92.1%, recall of 97.2%, and an F1-score of 94.6%.

In [22], the study is based on WSI classification using a MIL method with weak supervision. The feature extraction process uses self-supervised contrastive learning to generate high-quality representations for MIL. An innovative aspect is incorporating a pyramid fusion mechanism, which integrates features from different scales of the WSIs, providing a more holistic and detailed approach. The classification model is an MIL network with a dual-flow architecture. This model incorporates a MIL aggregator, which uses a trainable distance measure to model the relationships between instances. Self-supervised contrastive learning plays a vital role in effectively extracting representations for the MIL. The model was validated using two databases: Camelyon16 and the TCGA dataset. In Camelyon16, the DSMIL (Dual-Stream Multiple Instance Learning Network) model had an accuracy of 86.8% and an AUC of 89.4%. In the TCGA lung cancer dataset, the accuracy was 91.9%.

A Multi-View Attention-guided Multiple Instance Detection Network (MA-MIDN) model is proposed in [23]. The methodology integrates MIL, a new Multi-View Attention (MVA) algorithm, and a convolutional neural network (CNN) in an end-to-end structure. A Deep Mutual Learning (DML) scheme is also used for training. The main model is MA-MIDN, which combines MIL, MVA, and CNN. The BreaKHis, BACH, and PUIH datasets were used. In the BreaKHis dataset, the MA-MIDN model achieved an AUC of over 99.0%. In the other sets, it outperformed reference models by a significant margin. Some intrinsic limitations include the dependence on the quality and resolution of histopathological images and as the need for large volumes of data for effective model training.

A method for classifying the stages of retinopathy of prematurity (ROP) using deep learning techniques is developed in [24]. In the methodology, a Fully Convolutional Neural Network (FCN) extracts high-level features from fundus images. It generates a spatial score map (SSM) with this, the MIL trains and classifies the stages of ROP using SSM patches. The data was collected from various hospital institutions, resulting in 6209 retinal images. The method proved effective and achieved promising performance in classifying ROP phases. The approach proposed by the study is innovative in that it combines FCN and MIL with an attention module, which can significantly increase accuracy in classifying the stages of ROP, a challenging task due to the similarity between the early stages of the disease and the small region of the lesions concerning the full fundus images.

Reference [25] demonstrated a Transformer-based network architecture for Vision called Local-Global Vision Transformer (LGViT). LGViT aims to combine the advantages of Transformers in learning global representations and CNNs in capturing local features. The methodology involves using a mechanism called Local-Global Multi-head Self-attention (LGMSA) and a Ghost Feed-forward Network (GFFN). LGMSA is a self-attention mechanism that effectively captures local and global features of images with low computational cost. The GFFN brings locality to the network using a simple depth convolution. The database used was the PatchCamelyon (PCam) dataset. The LGViT model achieved results of 91.8% accuracy, 92.2% precision, 91.8% recall and 91.8% F1-Score. Limitations include the dependence on image quality and variability, the need for large sets of annotated data for training and validation, and the generalisability of the model to different types of data and image conditions.

A machine learning model called Multi-scale Efficient Graph Transformer (MEGT) for classifying WSIs in cancer pathology was developed in [26]. MEGT is a dual-branch Transformer model that aggregates image patches of different resolutions to improve the accuracy of cancer diagnosis in WSIs. This model is notable for its ability to integrate information from multiple scales and to capture spatial information relevant to diagnosis. MEGT's main features include the Efficient Graph-Transformer (EGT), a component that improves the ability of the branches in MEGT to learn spatial information in WSIs. It integrates a WSI graph representation with a Transformer to learn both the WSI's spatial information and the long-range dependencies between image patches. Multi-scale Feature Fusion Module (MFFM), is designed to learn multi-scale features and reduce the semantic gap between patches of different resolutions. The model was evaluated using The Cancer Genome Atlas Renal Cell Carcinoma (TCGA-RCC) and CAMELYON16 datasets. MEGT achieved an accuracy of 96.9% on TCGA-RCC and 96.89% on CAMELYON16.

In [27], a new method called MIST (Multiple Instance Learning for Whole Slide Image Classification of Colorectal Adenomas) was proposed to classify colorectal adenomas in whole slide images. MIST is based on the Swin Transformer for feature extraction, employing a three-stage process. Initially, patches are extracted from the images at magnifications of 2.5x and 5x. Subsequently, two Swin Transformer feature extractors are trained using self-supervised contrastive learning, and the resulting embeddings are combined to train the MIL aggregator, completing the classification of the entire blades. In the MIL implementation, MIST uses a MIL network with two branches. One identifies the critical instance with the highest score through max pooling, while the other evaluates the similarity of the instances to the critical instance. The characteristics are then aggregated to calculate the final bag score, which is derived from the average of the scores of the two branches. MIST uses the Swin Transformer complemented by models such as CLAM and DSMIL for classification. The research used a database with 666 images of whole colorectal cancer slides and 273 additional images MIST achieved an AUC of 78.5% in the internal validation set and 92.1% in the external validation set, as well as an accuracy of 78.4% and an F1-score of 73.6% in the latter set.

## IV. DISCUSSION

While exploring recent advances in MIL algorithms for classifying medical images, it is essential to highlight this methodology's distinct characteristics and specific challenges. MIL, unlike traditional supervised learning approaches, operates under the premise of bags of instances, where labels are only available for sets of instances and not for individual instances. This approach is particularly pertinent in medical scenarios, such as the analysis of histopathological images for the diagnosis of breast cancer, where the precise identification of pathological areas in large volumes of tissue is fundamental. MIL techniques have recently incorporated advances such as CNNs and transformers, adapting them to deal with medical data's ambiguous and often sparse nature. While offering significant promise, each technique faces inherent challenges, such as the need for extensively annotated datasets and issues of generalisation and interpretability critical in clinical application.

Based on the latest research, the following analysis explores the use of MIL techniques for classifying medical images. Attention mechanisms and transformers will be given special attention, as they are the most commonly used techniques. However, the analysis will also cover other innovative approaches. After this, we will go through the research questions and answer them. Table 2 summarises the main points of each article analyzed in this systematic review.

Given the predominant use of breast cancer images in these studies, it was decided to compile a comparative table to discern which models perform best with these images. This is motivated by the role that advanced image analysis plays in the early detection and accurate diagnosis of breast cancer. The comparative table, Table 3, aims to shed light on the diversity of MIL methodologies applied in different research, offering information on their effectiveness and the challenges they face when dealing with breast cancer imaging datasets.

Furthermore, to address the application of MIL beyond breast cancer images, Table 4 presents the metric results from models applied to a wide range of datasets, including datasets containing images of the colon, lung, skin cancers, blood cells, gastric and endometrial images, retina, among others. This expansion emphasises the versatility and adaptability of MIL techniques to the unique challenges presented by different types of data and pathologies.

By juxtaposing various studies, it provides a clear overview of the algorithms used, the specific breast cancer datasets utilized, and the resulting performance metrics such as accuracy and AUC. Including a wider variety of datasets in the analysis enhances understanding of the applicability and effectiveness of MIL techniques in a broader spectrum of medical imaging contexts, highlighting significant advances and persistent challenges in this area of research.

### A. ATTENTION MECHANISMS

The attention mechanisms emerged recently as a transformative component in the field of Deep Learning, adjusting the importance attributed to different parts of the input data, enabling deep learning models to identify and prioritise relevant information for carrying out specific tasks, improving DL models' interpretability and providing significant advances in a variety of applications including in MIL approaches [48].

After a thorough analysis of every article selected, [7], [13], [14], and [17], [18], [21], [23], [24], are the ones that mention the use of attention mechanisms. The combination of MIL with the triple kernel, [7], allows for a more refined manipulation of image features, and the generation of an attention map by kernel functions highlights important regions of the images, which is essential for identifying critical areas in complex images. Also, in [17], depending on the type of kernel used, the combination of self-attention and attention pooling can capture global dependencies between instances, which is crucial. The integration of attention mechanisms with multiple CNNs was also found, allowing the model to focus on specific features, [13], [23], or to calculate weighted averages on the instance's embeddings, which can be very efficient in terms of time and resources [14]. It is also observed in the integration of MIL with shuffle attention, in which the central idea is to "shuffle" the elements within the feature maps [21], helping capture pixel-level relationships and channel dependencies. In [23], a new form of attention, MVA, is used to identify the relevant features in different perspectives of the images. A different approach to implementation attention mechanisms is a graph-based model with GNNs, which is important for efficiently merging node information into a graph [18].

In all these studies, attention mechanisms play a crucial role in improving the accuracy and effectiveness of MIL models, allowing them to focus on the most important features of medical images. This method is particularly useful in clinical settings, where the identifi-

**TABLE 2.** Resume of the papers selected.

| Reference | Dataset | Models Used | Metrics |
|---|---|---|---|
| Hu, Huafeng, et al. 2023 [7] | USBC breast cancer, colon cancer, Musk1, Musk2, Fox, Tiger, Elephant, MNIST-based dataset | TGA-MIL (Triple-kernel Gated Attention-based Multiple Instance Learning with Contrastive Learning) | Accuracy, Precision, Recall, F-score, AUC |
| Y Wang, et al. 2023 [8] | Camelyon16, Tcga-Lung and MSK | CWC-Transformer | AUC, Accuracy |
| Frade, Julieta, et al. 2022 [9] | Non-Small Cell Lung Cancer Radiogenomics(NSCLC Radiogenomics), Interstitial Lung Disease (ILD), a private database from the Centro Hospitalar e Universitário de São João (CHUSJ) | Normalised Set Kernel (NSK) | AUC |
| Zhang, Xiaoxian, et al. 2022 [4] | Camelyon16, Pneumonia CT | Dual Space Multiple Instance Representative Learning (DSMIRL) | Accuracy, AUC, F1-Score |
| Tian, Yu, et al. 2022 [10] | HyperKvasir, LDPolypVideo | Convolutional transformer MIL anomaly classifier, Contrastive snippet mining (CSM) approach | AUC |
| Combalia, Marc, et al. 2018 [11] | MNIST-Sparse, MNIST-Clustered, ICIAR Grand Challenge 2018 Part A, GTEx Skin | VGG-like architecture for MNIST, ResNet-based architecture for ICIAR, and GTEx | Accuracy |
| Qu, Linhao, et al. 2022 [12] | Camelyon16, Tcga-Lung | Distributed Guided MIL for WSI, Self-Supervised masked Autoencoders (MAE) | AUC, Accuracy |
| Teramoto, Atsushi, et al. 2021 [13] | Dataset made of lung cells (Self-created) | MIL Deep Learning Algorithm with attention (ADMIL), CNNs, Otsu algorithm | Accuracy |
| Sadafi, Ario, et al. 2020 [14] | Blood samples, part of CoMMitMenT study | Mask R-CNN, MIL Attention Pooling | AUC, F1-Score |
| Zheng, Yushan, et al. 2022 [15] | Gastric2K, Endometrial-2K | Kernel Attention Transformer (KAT) | Accuracy |
| Ahmad, Nouman, et al. 2022 [16] | BreakHis | Suport Vector Machine (SVM), EfficientNet | Accuracy |
| Rymarczyk, Dawid, et al. 2021 [17] | MNIST, Breast Cancer Histologic dataset, DIFaS, Retinal Image Set | Self-Attention Attention-Based MIL Pooling (Sa-AbMILP), Resnet-18 | Accuracy |
| Sens, Daniel, et al. 2023 [19] | BRACS, Camelyon17 | TransMIL, Bag Embedding Loss (BEL) | Accuracy, F1-Score |
| Zhang, Jianxin, et al. 2022 [20] | Camelyon16, MSK | Attention MIL Transformer (AMIL-Trans) | AUC |
| Hou, Cunqiao, et al.2022 [21] | Camelyon16 | Shuffle Attention (SA), Multi-Head Attention (MHA), Long Short Term Memory (LSTM) | AUC, Precision, Recall, F1-Score |
| Li, Bin, et al. 2021 [22] | Camelyon16, Tcga-Lung | Dual-Stream (DSMIL), Attention Pooling | Accuracy, AUC |
| Li, Guangli, et al. 2021 [23] | BreakHis, BACH, PUIH | Multi-View Attention-guided Multiple Instance Detection (MA-MIDN) | AUC |
| Chen, Shaobin, et al. 2021 [24] | Data collected from various hospital institutions | Fully Convolutional Neural Networks (FCN), Space Score Map (SSM) | AUC |
| Wang, Lang, et al. 2023 [25] | PatchCamelyon | Local Global Vision Transformer (LGViT) | Accuracy, Precision, Recall, F-score |
| Ding, Saisai, et al. 2023 [26] | Cancer Genome Atlas Renal Cell Carcinoma (TCGA-RCC), Cammelyon16 | Multi-scale Efficient Graph-Transformer (MEGT) | Accuracy |
| Cai, Hongbin, et al. [27] | Colorectal Cancer images | Swin Transformer, CLAM, DSMIL | AUC, Accuracy, F1-Score |
| Li, Xiaoyu, et al. 2023 [18] | BreakHis | Graph-based multiple instance learning (GMIL) | AUC |

cation of subtle patterns can be crucial for diagnosis and treatment.

### B. TRANSFORMERS

Introduced in 2017 by Vaswani et al. [49], they have revolutionised DL with its self-attention mechanism that processes sequences in parallel, in contrast to previous architectures based on recursion or convolutions. Essential in natural language processing tasks, they have also gained ground in computer vision and medical image analysis, helping to detect and classify pathological features. Their ability to generate contextualised representations of data is especially valuable in medicine, where they integrate information from various sources for accurate diagnoses, continuing to evolve to meet the complexities of medical imaging.

In analysing the articles that mention the use of transformers [8], [10], [19], [25], [26], [27], we observed various implementations and significant impacts in the field of medical image analysis. In the article [8], the CWC-Transformer model was developed for WSI classification, combining contrastive learning with CNNs and transformers. This approach stands out for its ability to capture both local and global information from the images, providing a richer

**TABLE 3.** Results from the studies that used breast cancer datasets.

| Dataset | Model Used | Accuracy | AUC |
|---|---|---|---|
| USBC breast cancer [28] | TGA-MIL [7] | 77.0% | 75.6% |
| Camelyon16 [29]/ MSK [30] | CWC Transformer [8] | 89.1%/ 92.6% | 93.9%/ 94.7% |
| Camelyon16 [29] | DSMIRL [4] | 89.9% | 95.3% |
| ICIAR Grand Challenge 2018 Part A [31] | ResNet based architecture [11] | 77.6% - 84.7% | - |
| Camelyon16 [29] | MAE [12] | 80.2% | 83.7% |
| BreakHis [32] | SVM, EfficientNet [16] | 97.0% - 99.6% | - |
| Breast Cancer Dataset [28] | Sa-AbMILP [17] | 65.5% - 76.7% | 85.8% - 86.7% |
| BRACS [33]/ Camelyon17 [34] | BEL [19] | 60.0%/ 68.0% | 76.0%/ 68.0% |
| Camelyon16 [29]/ MSK [30] | AMIL-Trans [20] | - | 94.3%/ 84.2% |
| Camelyon16 [29] | MHA, LSTM [21] | 96.0% | 95.3% |
| Camelyon16 [29] | DSMIL [22] | 86.8% | 89.4% |
| BreakHis [32], BACH [35], PUIH [36] | MA-MIDN [23] | 96.0%- 98.8% | 99.0% - 99.8% |
| PatchCamelyon [37] | LGViT [25] | 91.8% | - |
| BreakHis [32] | GMIL [18] | 96.4% - 99.8% | 99.7% - 98.6% |

**TABLE 4.** Results from the other datasets * - Datasets not referenced.

| Dataset | Model Used | Accuracy | AUC |
|---|---|---|---|
| Colon Cancer dataset [38] | TGA-MIL [7] | 92.7% | 98.3% |
| Tcga-Lung [39] | CWC Transformer [8] | 92.6% | 94.7%/ |
| NSCLC [40]/ ILD [41] | NSK [9] | - | 59.0% - 89.0%/ 59.0% |
| Pneumonia CT | DSMIRL [4] | 93.0% | 96.7% |
| HyperKvasir [42] & LDPolypVideo [43] | CSM [10] | - | 98.4% |
| GTEx Skin [44] | ResNet based architecture [11] | 82.6% - 94.2% | - |
| Tcga-Lung [39] | MAE [12] | 92.0% | 97.0% |
| Self made Lung Cells Dataset* | ADMIL [13] | 91.6% | - |
| Blood samples [45] | Mask R-CNN [14] | 79.0% | 96.0% |
| Gastric-2K*/ Endometrial-2K* | KAT [15] | 91.5%/ 94.9% | 96.7%/ 98.3% |
| Retinal Image Dataset [46] | Sa-AbMILP [17] | 76.3% | - |
| Tcga-Lung [39] | DSMIL [22] | 91.9% | 96.3% |
| Retina Data collected from various hospitals* | FCN/SSM [24] | 94.4% | 97.2% |
| TCGA-RCC [47] | MEGT [26] | 96.9% | 97.9% |
| Colorrectal Cancer Images* | Swin Transformer/ CLAM/ DSMIL [27] | - | 78.4% |

and more detailed analysis. In [10], a MIL method based on transformers is presented for the detection of polyps in colonoscopy videos with the ability to operate with weakly labelled videos and capture global temporal relationships, surpassing previous techniques and demonstrating the effectiveness of transformers in video contexts. In [19], the TransMIL model uses Neystrom's Method [50] to process many instances in high-resolution images, an important innovation to show the usefulness of transformers in dealing with complex, high-dimensional data. The article [25] introduces LGViT. This architecture combines the advantages of Transformers in learning global representations with the ability of CNNs to capture local features, highlighting the synergy between global and local learning provided by Transformers. MEGT, discussed in [26], is a dual-branch Transformer model that integrates image patches of different resolutions to improve the accuracy of cancer diagnosis in WSIs. This ability to combine information from multiple scales stands out as a crucial approach for the detailed analysis of pathological images. To conclude, the article [27] describes the use of MIST, a method based on the Swin Transformer, to classify colorectal adenomas in WSIs. This approach improves classification accuracy, benefiting from the efficiency of the Swin Transformer in combination with a three-stage process and a MIL aggregator.

These studies collectively illustrate how transformers offer a range of benefits in medical image analysis, from improved accuracy to the ability to process and analyse complex data, playing a crucial role in advancing accurate and detailed medical diagnosis.

### C. TRANSFORMERS AND ATTENTION MECHANISMS

The combination of transformers and attention mechanisms in machine learning, especially in MIL contexts, represents a powerful fusion of advanced artificial intelligence techniques. These techniques allow MIL models to identify key features within a bag of instances and understand the complex relationships between these instances, significantly improving the accuracy and effectiveness of image-based diagnoses or analyses. [51]

And since, in the articles selected, studies that approach both techniques [15], [20] were found, there will be some discussion about that. In [15], KAT uses the K-means algorithm to extract a set of anchors based on the spatial clustering property of the features and EfficientNet-b0 to extract features from WSI patches. In addition, a cross-attention algorithm is used to facilitate information communication between kernels and patch tokens. The use of transformers and attention mechanisms in the KAT aims to highlight the most relevant regions of the images for more accurate classification, resulting in high accuracy in the subtyping and classification tasks in the datasets. In [20], the AMIL-Trans model integrates ResNet-50 with an efficient channel attention (ECA) module to select discriminating instances and a Transformer encoder for bag-level aggregation. This model's implementation of attention mechanisms and transformers is designed to capture discriminant information and correlations between instances, improving feature selection and aggregation at the bag level.

These studies exemplify how integrating transformers and attention mechanisms into MIL models offers a more refined and detailed analysis of medical images. This approach improves diagnostic accuracy by highlighting crucial discriminating features.

### D. OTHER METHODS

In the context of MIL for medical image analysis, techniques such as CNNs and supervised and unsupervised learning methods offer valuable alternatives to attention mechanisms and transformers, contributing significantly to accurate and efficient diagnosis [52].

In the articles [4], [9], [11], [12], [16], and [22], we see the application of various techniques in the context of MIL for analysing medical images, each chosen for its specific advantages.

In [9], the method for detecting lung diseases uses CT scans and combines different bag generators with the NSK algorithm, providing a more comprehensive and accurate analysis of lung diseases. DSMIRL, presented in [4], addresses the challenges of the imbalance between positive and negative instances in medical images, offering a balanced solution through a combination of adaptive instance selection and multi-instance representative learning that, although complex, this method improves classification accuracy. In [11], the patch sampling strategy based on the Monte Carlo method is chosen for its effectiveness in improving classification accuracy in histological images, overcoming the limitations of uniform sampling. The paper [12] uses masked autoencoders and a pseudo-label-based refinement strategy to improve the initialisation of the feature space, resulting in high accuracy and AUC. In [16], the CAD system for breast cancer employs transfer learning and discriminative patch selection, aiming for high classification accuracy. Finally, [22] uses MIL with weak supervision and self-supervised contrastive learning, choosing this approach to integrate a pyramid fusion mechanism and provide a more holistic and detailed view in WSI analyses despite the complexity of the model.

Each of these studies highlights the importance of selecting appropriate techniques in MIL to deal with the specificities of medical images, seeking to balance accuracy, processing capacity, and suitability for the available data.

### E. LIMITATIONS

By analysing the limitations presented in the studies from [7] to [27], we observe common and specific challenges faced in MIL approaches to medical image analysis.

In [7], the implementation of the TGA-MIL model reveals challenges such as the need to fine-tune for different tissues and the dependence on image quality. Similarly, [8] with its CWC-Transformer model highlights the high consumption of computational resources, especially with high magnification images, and potential problems in treating similar patches such as negative pairs. The study [9] on the detection of lung diseases faces limitations due to the restricted size of the databases, compromising the generalizability of the results. Similarly, [4] and [10] show challenges related to the complexity of their methods and the need for large data sets. The article [11], which proposes a Monte Carlo-based sampling strategy, may face challenges with the computational effort required for effective implementation. In [12], limitations include the need for large, annotated datasets and challenges in generalising the model. The study [13] on classifying lung cytology images may face similar, though unspecified, challenges. The analysis of individual cells in [14] is limited by the small number of cases studied, requiring external validation and investigation of applicability to other conditions. The article [15] on the KAT model highlights the dependence on image quality and the need to fine-tune the model. The study [16] on the CAD system for breast cancer diagnosis highlights limitations such as the possibility of omitting relevant information due to the selection of discriminative patches. The [17] faces challenges related to dependence on the type of kernel used and scalability issues. The challenges of poor annotations and the large size of the "bags" are evident in [19], as is the sensitivity to the quality of the class representations. The paper [20], with its AMIL-Trans model, highlights the need for large amounts of labelled data and challenges in generalisation. The study [21] on the SAMIL model for breast cancer WSI classification can address model complexity and the need for adjustments for different types of data. In [22], WSI classification with MIL and self-supervised learning faces implementation challenges due to its complexity. The article [23] with the MA-MIDN model highlights the dependence on image quality and the need for large volumes of data. [24] addresses the classification of stages of retinopathy of prematurity, facing challenges due to the similarity between early stages of the disease. The [25] with LGViT, also has limitations, such as the dependence on image quality and the need for large data sets. The MEGT model in [26] faces challenges in integrating information from various scales. Finally, [27] the MIST method highlights common challenges in medical imaging studies, including the need for large, labelled datasets and variability in imaging techniques.

These limitations underline the complexity of developing effective MIL models for medical image analysis, highlighting the importance of data quality, model generalisation, and the balance between accuracy and computational complexity.

### F. RESEARCH QUESTIONS

In this section, we explore research questions related to challenges, influences, effectiveness, and diversity of algorithms in Multiple Instance Learning (MIL) applied to image classification. Based on the summaries of recent studies, this analysis seeks to elucidate critical aspects that shape the MIL field, ranging from technical and methodological limitations to the nuances in algorithm performance on different types of image datasets.

**Q1:** The main challenges and limitations of current MIL methods in image classification include the need for large sets of annotated data, computational complexity, and dependence on image quality and diversity. As seen in the studies [7], [8], [9], and others, these challenges directly impact the generalizability of the results and the applicability of the models to different pathological conditions or tissue types.

**Q2:** The quality and quantity of training data significantly influence the performance of MIL methods. Insufficient or low-quality data can limit the model's ability to learn accurate and generalisable representations, as evidenced in [9] and [12]. Large sets of annotated data are crucial for effective training, especially in complex models such as those described in [4] and [20].

**Q3:** The accuracy of MIL methods varies on different types of image datasets. For example, in [15] and [16], high accuracy was observed in classifying histopathology and breast cancer images. However, this accuracy can be influenced by several factors, including the complexity of the image, the nature of the classification problem, and the specificity of the MIL algorithm used.

**Q4:** Several MIL algorithms are used in image classification, ranging from techniques based on kernels and self-attention, as in [7] and [17], to approaches that integrate contrastive learning and pyramid fusion, as seen in [22]. In addition, algorithms that combine CNNs, transformers, and attention strategies, as in [8] and [20], demonstrate the versatility and adaptability of MIL methods to the specific needs of image classification.

Therefore, the field of MIL in image classification is marked by a diversity of approaches and techniques, each with its challenges and advantages. Choosing the appropriate algorithm and the availability of high-quality data are crucial to successfully applying these methods in practical medical image analysis contexts.

## V. CONCLUSION

This systematic review article rigorously explores the dynamic field of multiple instance learning (MIL) applied to image classification, offering valuable insights into current trends, challenges, and technological advances in this area. Through carefully analysing the selected studies, we addressed the defined research questions, uncovering crucial aspects of MIL and its application in various imaging contexts, with a special emphasis on medical imaging.

From this review, although MIL offers considerable promise for image analysis, especially in medical applications, significant challenges remain. These include the need for greater interpretability of models, more efficient methods for dealing with large volumes of data, and the integration of specific domain knowledge into learning processes.

This study serves as a basis for future research, suggesting the exploration of advanced machine-learning techniques and the integration of specialised knowledge to improve the accuracy and usefulness of MIL methods in practical applications. Furthermore, it emphasizes the importance of high-quality, representative datasets for the field's continued evolution.

In short, MIL continues to be a vital and expanding area of research, with significant potential to positively impact various applications, especially in medical imaging. As technology advances and new approaches are developed, MIL solutions are expected to become even more effective and widely adopted, contributing to significant advances in various fields.

## REFERENCES

[1] Ł. Struski, D. Rymarczyk, A. Lewicki, R. Sabiniewicz, J. Tabor, and B. Zieliński, "ProMIL: Probabilistic multiple instance learning for medical imaging," 2023, *arXiv:2306.10535*.

[2] S. Fatima, S. Ali, and H.-C. Kim, "A comprehensive review on multiple instance learning," *Electronics*, vol. 12, no. 20, p. 4323, Oct. 2023.

[3] J. Jian, W. Xia, R. Zhang, X. Zhao, J. Zhang, X. Wu, Y. Li, J. Qiang, and X. Gao, "Multiple instance convolutional neural network with modality-based attention and contextual multi-instance learning pooling layer for effective differentiation between borderline and malignant epithelial ovarian tumors," *Artif. Intell. Med.*, vol. 121, Nov. 2021, Art. no. 102194.

[4] X. Zhang, S. Huang, Y. Zhang, X. Zhang, M. Gao, and L. Chen, "Dual space multiple instance representative learning for medical image classification," in *Proc. 33rd Brit. Mach. Vis. Conf. (BMVC)*. London, UK, Nov. 2022, pp. 768–779.

[5] M. Gadermayr and M. Tschuchnig, "Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential," *Computerized Med. Imag. Graph.*, vol. 112, Mar. 2024, Art. no. 102337.

[6] D. Moher, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *Ann. Internal Med.*, vol. 151, no. 4, p. 264, Aug. 2009.

[7] H. Hu, R. Ye, J. Thiyagalingam, F. Coenen, and J. Su, "Triple-kernel gated attention-based multiple instance learning with contrastive learning for medical image analysis," *Appl. Intell.*, vol. 53, no. 17, pp. 20311–20326, Sep. 2023.

[8] Y. Wang, J. Guo, Y. Yang, Y. Kang, Y. Xia, Z. Li, Y. Duan, and K. Wang, "CWC-transformer: A visual transformer approach for compressed whole slide image classification," *Neural Comput. Appl.*, vol. 2023, pp. 1–13, Jan. 2023.

[9] J. Frade, T. Pereira, J. Morgado, F. Silva, C. Freitas, J. Mendes, E. Negrão, B. F. de Lima, M. C. D. Silva, A. J. Madureira, I. Ramos, J. L. Costa, V. Hespanhol, A. Cunha, and H. P. Oliveira, "Multiple instance learning for lung pathophysiological findings detection using CT scans," *Med. Biol. Eng. Comput.*, vol. 60, no. 6, pp. 1569–1584, Jun. 2022.

[10] Y. Tian, G. Pang, F. Liu, Y. Liu, C. Wang, Y. Chen, J. Verjans, and G. Carneiro, "Contrastive transformer-based multiple instance learning for weakly supervised polyp frame detection," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 88–98.

[11] M. Combalia and V. Vilaplana, "Monte-Carlo sampling applied to multiple instance learning for histological image classification," in *Proc. Int. Workshop Deep Learn. Med. Image Anal.* Cham, Switzerland: Springer, 2018, pp. 274–281.

[12] L. Qu, X. Luo, S. Liu, M. Wang, and Z. Song, "DGMIL: Distribution guided multiple instance learning for whole slide image classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 24–34.

[13] A. Teramoto, Y. Kiriyama, T. Tsukamoto, E. Sakurai, A. Michiba, K. Imaizumi, K. Saito, and H. Fujita, "Weakly supervised learning for classification of lung cytological images using attention-based multiple instance learning," *Sci. Rep.*, vol. 11, no. 1, p. 20317, Oct. 2021.

[14] A. Sadafi, A. Makhro, A. Bogdanova, N. Navab, T. Peng, S. Albarqouni, and C. Marr, "Attention based multiple instance learning for classification of blood cell disorders," in *Proc. 23rd Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Lima, Peru. Cham, Switzerland: Springer, 2020, pp. 246–256.

[15] Y. Zheng, J. Li, J. Shi, F. Xie, and Z. Jiang, "Kernel attention transformer (KAT) for histopathology whole slide image classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 283–292.

[16] N. Ahmad, S. Asghar, and S. A. Gillani, "Transfer learning-assisted multi-resolution breast cancer histopathological images classification," *Vis. Comput.*, vol. 38, no. 8, pp. 2751–2770, Aug. 2022.

[17] D. Rymarczyk, A. Borowa, J. Tabor, and B. Zielinski, "Kernel self-attention for weakly-supervised image classification using deep multiple instance learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1720–1729.

[18] X. Li, B. Yang, T. Chen, S. Lv, Z. Gao, and H. Li, "A weakly supervised multi-instance learning based on graph neural network for breast cancer pathology image classification," in *Proc. Int. Conf. Commun., Comput. Artif. Intell. (CCCAI)*, Jun. 2023, pp. 47–51.

[19] D. Sens, A. Sadafi, F. Paolo Casale, N. Navab, and C. Marr, "BEL: A bag embedding loss for transformer enhances multiple instance whole slide image classification," 2023, *arXiv:2303.01377*.

[20] J. Zhang, C. Hou, W. Zhu, M. Zhang, Y. Zou, L. Zhang, and Q. Zhang, "Attention multiple instance learning with transformer aggregation for breast cancer whole slide image classification," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2022, pp. 1804–1809.

[21] C. Hou, Q. Sun, W. Wang, and J. Zhang, "Shuffle attention multiple instances learning for breast cancer whole slide image classification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 466–470.

[22] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14313–14323.

[23] G. Li, C. Li, G. Wu, D. Ji, and H. Zhang, "Multi-view attention-guided multiple instance detection network for interpretable breast cancer histopathological image diagnosis," *IEEE Access*, vol. 9, pp. 79671–79684, 2021.

[24] S. Chen, R. Zhang, G. Chen, J. Zhao, T. Wang, G. Zhang, and B. Lei, "Attention-guided deep multi-instance learning for staging retinopathy of prematurity," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2021, pp. 1025–1028.

[25] L. Wang, J. Liu, P. Jiang, D. Cao, and B. Pang, "LGVIT: Local-global vision transformer for breast cancer histopathological image classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[26] S. Ding, J. Li, J. Wang, S. Ying, and J. Shi, "Multi-scale efficient graph-transformer for whole slide image classification," 2023, *arXiv:2305.15773*.

[27] H. Cai, X. Feng, R. Yin, Y. Zhao, L. Guo, X. Fan, and J. Liao, "MIST: Multiple instance learning network based on Swin transformer for whole slide image classification of colorectal adenomas," *J. Pathol.*, vol. 259, no. 2, pp. 125–135, Feb. 2023.

[28] E. Drelie Gelasca, J. Byun, B. Obara, and B. S. Manjunath, "Evaluation and benchmark for biological image segmentation," in *Proc. 15th IEEE Int. Conf. Image Process.*, 2008, pp. 1816–1819.

[29] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," 2016, *arXiv:1606.05718*.

[30] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16514–16524.

[31] A. Rakhlin, A. Shvets, V. Iglovikov, and A. A. Kalinin, "Deep convolutional neural networks for breast cancer histology image analysis," 2018, *arXiv:1802.00752*.

[32] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1455–1462, Jul. 2016.

[33] N. Brancati, A. M. Anniciello, P. Pati, D. Riccio, G. Scognamiglio, G. Jaume, G. De Pietro, M. Di Bonito, A. Foncubierta, G. Botti, M. Gabrani, F. Feroce, and M. Frucci, "BRACS: A dataset for BReAst carcinoma subtyping in H&E histology images," *Database*, vol. 2022, Oct. 2022, Art. no. baac093.

[34] P. Bándi et al., "From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 550–560, Feb. 2019.

[35] G. Aresta et al., "BACH: Grand challenge on breast cancer histology images," *Med. Image Anal.*, vol. 56, pp. 122–139, Aug. 2019.

[36] R. Yan, F. Ren, Z. Wang, L. Wang, T. Zhang, Y. Liu, X. Rao, C. Zheng, and F. Zhang, "Breast cancer histopathological image classification using a hybrid deep neural network," *Methods*, vol. 173, pp. 52–60, Feb. 2020.

[37] B. E. Bejnordi et al., "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, vol. 318, no. 22, pp. 2199–2210, 2017.

[38] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. J. Snead, I. A. Cree, and N. M. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1196–1206, May 2016.

[39] Y. Zhao, Y. Gao, X. Xu, J. Zhou, and H. Wang, "Multi-omics analysis of genomics, epigenomics and transcriptomics for molecular subtypes and core genes for lung adenocarcinoma," *BMC Cancer*, vol. 21, no. 1, p. 257, Dec. 2021.

[40] S. Bakr, O. Gevaert, S. Echegaray, K. Ayers, M. Zhou, M. Shafiq, H. Zheng, J. A. Benson, W. Zhang, and A. N. C. Leung, "A radiogenomic dataset of non-small cell lung cancer," *Sci. Data*, vol. 5, no. 1, pp. 1–9, 2018.

[41] A. Depeursinge, A. Vargas, A. Platon, A. Geissbuhler, P.-A. Poletti, and H. Müller, "Building a reference multimedia database for interstitial lung diseases," *Computerized Med. Imag. Graph.*, vol. 36, no. 3, pp. 227–238, Apr. 2012.

[42] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, D. Johansen, C. Griwodz, H. K. Stensland, E. Garcia-Ceja, P. T. Schmidt, H. L. Hammer, M. A. Riegler, P. Halvorsen, and T. de Lange, "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *Sci. Data*, vol. 7, no. 1, p. 283, Aug. 2020.

[43] Y. Ma, X. Chen, K. Cheng, Y. Li, and B. Sun, "LDPolypVideo benchmark: A large-scale colonoscopy video dataset of diverse polyps," in *Proc. 24th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 387–396.

[44] J. Lonsdale et al., "The genotype-tissue expression (GTEx) project," *Nature Genet.*, vol. 45, no. 6, pp. 580–585, 2013.

[45] A. Sadafi, N. Koehler, A. Makhro, A. Bogdanova, N. Navab, C. Marr, and T. Peng, "Multiclass deep active learning for detecting red blood cell subtypes in brightfield microscopy," in *Proc. 22nd Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Shenzhen, China. Cham, Switzerland: Springer, 2019, pp. 685–693.

[46] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J.-C. Klein, "Feedback on a publicly distributed image database: The messidor database," *Image Anal. Stereol.*, vol. 33, no. 3, p. 231, Aug. 2014.

[47] P. Cao, J.-Y. Wu, J.-D. Zhang, Z.-J. Sun, X. Zheng, B.-Z. Yu, H.-Y. Cao, F.-L. Zhang, Z.-H. Gao, and W. Wang, "A promising prognostic risk model for advanced renal cell carcinoma (RCC) with immune-related genes," *BMC Cancer*, vol. 22, no. 1, p. 691, Dec. 2022.

[48] G. Brauwers and F. Frasincar, "A general survey on attention mechanisms in deep learning," 2022, *arXiv:2203.14263*.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, ''Attention is all you need,'' in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[50] S. Wang, L. Luo, and Z. Zhang, ''SPSD matrix approximation VIS column selection: Theories, algorithms, and extensions,'' 2014, *arXiv:1406.5675*.

[51] A. M. Hafiz, S. A. Parah, and R. U. A. Bhat, ''Attention mechanisms and deep learning for machine vision: A survey of the state of the art,'' *Res. Square*, Jun. 2021, pp. 12–18, doi: 10.21203/rs.3.rs-510910/v1.

[52] K. Raza and N. K. Singh, ''A tour of unsupervised deep learning for medical image analysis,'' *Current Med. Imag. Rev.*, vol. 17, no. 9, pp. 1059–1077, 2021.

**DALILA BARBOSA** received the bachelor's degree in biomedical engineering from the University of Trás-os-Montes and Alto Douro. Currently, she is a Researcher in computer assisted gastric cancer diagnosis, a project with InescTec and the University of Trás-os-Montes e Alto Douro, her work focuses mainly on deep learning models and MIL algorithms research for medical imaging classification.

**MARCOS FERREIRA** received the master's degree in computer science from the Federal University of Maranhão (UFMA), where he is currently pursuing the Ph.D. degree. His research interests include computer vision, deep learning, and medical image processing.

**GERALDO BRAZ JUNIOR** received the Ph.D. degree in electrical engineering from the Federal University of Maranhão. He is currently a Professor with the Federal University of Maranhão. He has experience in computer vision, machine learning, deep learning, and medical image processing.

**MARTA SALGADO** received the degree in medicine from the University of Porto, in 1997. She completed her specialty in gastroenterology, in 2005. She is currently a Graduate Hospital Assistant with the Gastroenterology Department, University Hospital Centre of Porto. She is also a Guest Lecturer on the master's degree in medicine with the Abel Salazar Biomedical Sciences Institute and the author of dozens of articles presented at scientific meetings and published in scientific journals.

**ANTÓNIO CUNHA** is currently a Ph.D. Senior Researcher and an Auxiliary Professor with the Engineering Department, University of Trás-os-Montes and Alto Douro (UTAD). He has participated as a member in seven funded research projects. His research interests include medical image analysis, bio-image analysis, computer vision, machine learning, and artificial intelligence, particularly in computer-aided diagnosis applied in several imaging modalities, e.g., computed tomography of the lung and endoscopic videos. He is part of the organization Committee HCIST—International Conference on Health and Social Care Information Systems and Technologies (2013–2015) and (2020–2023) and the Organization Chair (2012) and an Advisory Board (2016–2023). He has been a member of the Centre for Biomedical Engineering Research (C-BER), INESC TEC, since 2015, where one of the final goals is to create a CAD system based on DL approaches to assist clinical doctors in different biomedical image-related diagnoses/screening tasks. Thus, he has plenty of experience in medical imaging analysis and student supervision. In the last five years, he supervised two Ph.D. students and 22 M.Sc. students and published 17 journal articles and 34 Scopus conference papers in this area. He is a referee of several international journals and conferences and participated yearly in organizing several international scientific events. He is the General Chair of the MobiHealth2022 International Conference.