Vrije University Amsterdam
MSc AI, Spring 2021

Semi-supervised Multiple Instance Learning using Variational Auto-Encoders

A. Nihat Uzunalioglu

Supervisors: Jacub Tomczak & Tameem Adel

Master thesis, Faculty of Science

Major: Artificial Intelligence

Vrije University

Abstract

In this thesis, a hybrid model comprised of an Attention-based MIL classifier (AD-MIL) and a Variational Autoencoder (VAE) is defined in the direction of developing a deep generative framework for multiple-instance learning (MIL). For such purpose, this dissertation follows the subsequent research questions; (1) Integrating a VAE and an Attention-based Deep MIL classifier, (2) Investigating whether the hybrid learning approach can provide state-of-the-art performance in the semi-supervised setting, (3) Evaluating the proposed approach on the semi-supervised scenario and comparing it with baselines. The experiments are evaluated on the MNIST-BAGS and one real-life histopathology (the COLON CANCER) datasets. One appealing aspect of the results is that the ssMILVAE can integrate the density of the features $p(x)$ and the predictive distribution $p(y \mid x)$. The availability of the joint density $p(x, y)$ brings an opportunity to gain additional capabilities such as detecting uncertain decision-making of the model and enabling semi-supervised learning. In addition to this, the ssMILVAE achieves better accuracy than the purely predictive baseline model and samples in a similar quality to a VAE when the model encounters an insufficient number of labeled training data. Nevertheless, the attention mechanism provides more accurate attention weights over the instances of a bag which can be used to create heatmaps for representing the region of interests (ROIs). The code is available at https://github.com/anu43/ssMILVAE.

Keywords – Semi-supervised Learning, Hybrid Learning, Multiple Instance Learning, Attention-based MIL classification, Variational Autoencoder

Contents

List of Figures

List of Tables

1 Introduction

1.1 Image Classification

Image classification is one of the quintessential challenges in machine learning problems. It may vary from binary to multiclass categorization. Several domains like medical (Tufail et al., 2020), quality control (Diaz and Morales-Menendez, 2018; Escobar and Morales-Menendez, 2018), and information retrieval (Cui et al., 2002; Nallapati, 2004) can be addressed to the binary classification where this strategy includes only two possible outputs. Whereas the multiclass problem includes three or more class labels to predict. Over the past decade, most research in machine learning has emphasized the use of image classification (Goetz et al., 2015; Litjens et al., 2017; Liu et al., 2012; He et al., 2018) that plays an important role in the maintenance of analyzing many images by human labor in daily life. Thus, a great deal of interest and investments into this domain has been extremely increased to build such a mechanism with a high success rate in predicting the category. This widespread commitment helped dramatic improvements (Thompson et al., 2020) in computational power such as deep learning being ported to GPUs that speeds up from 5 to 15 times more (Oh and Jung, 2004; Raina et al., 2009) and image analysis algorithms as in the advent of the Convolutional Neural Networks (CNNs) (Lecun et al., 1998). In addition, it gave a rise to powerful developments in computer-assisted analytical approaches to various domains such as medical.

1.2 Image Analysis in Medical Domain

Image analysis is fast becoming a key instrument in medical domain (Cireşan et al., 2013; Xu et al., 2016; Sirinukunwattana et al., 2016; Jimenez-del Toro et al., 2017; Rakhlin et al., 2018). It is at the heart of the understanding of the case as an expert tool in the medical area (Nahid et al., 2018). Yet, it hosts various challenges (Lee and Yoon, 2017). In this study, three substantial challenges are scrutinized and relevant solutions are presented.

1.3 Challenges

There exists significant challenges concerning feature representation (Yetisgen-Yildiz and Pratt, 2005; Liu et al., 2019), finding sufficient number of labeled training data (Shin et al., 2016; Chen et al., 2021) and building computationally and algorithmically efficient environments (Oh and Jung, 2004) to construct such image analysis systems in the medical domain.

Feature representation is a notable area of interest within the field of image classification (Xu et al., 2014). One notable issue is that the data may have varied patterns. Sirinukunwattana et al. (2016) state that histopathologic images may have a very diverse morphology and distribution that makes detection of a sequence very hard. In addition, there is a high chance of encountering complex tissue architectures. Particularly, colorectal adenocarcinoma, dysplastic and epithelial cells often have unrelated chromatin textures and are dispersed with no clear boundary. The quality of the inferior image may fail due to the poor fixation and staining during the tissue preparation. Thus, an expert should involve in this procedure which makes the dataset creation progress very expensive. Another problem emerges as human interactions are drawn in the progress. Experts in the medical domain can evaluate a tissue image differently (Lafarge et al., 2017) for several reasons such as different staining and scanning techniques (Ciompi et al., 2017).

Another challenge is that the insufficient number of labeled training images in the medical domain. The insufficiency splits into two directions. One of the main factors causing this problem is a small number of studied disease incidence. Less frequency in studying various topics and diseases means fewer images and analyses. Therefore, collecting medical images becomes costly. Another reason is that the labor work needed to manually label images. This procedure especially brings a significant amount of intense labor work to process over the collection of images with extra care by manual annotations. This extension of labor work demands a great deal of effort. Consequently, the lack of annotated training data may cause several problems such as overfitting (Tajbakhsh et al., 2016) that occurs in exploiting few data in deep learning approaches. Even overcoming this step does not solve the problem completely. Because the quantification of medical interpretations is hard enough and the manual annotations that are done by clinical experts also become essentially ambiguous (Lafarge et al., 2017).

A primary concern of image classification is computational complexity. Alongside the computational improvement over the recent years (Oh and Jung, 2004; Raina et al., 2009), technological developments appear in the mechanics as well. Computers can deal with higher-level dimensionality compared to the past, the tissue scanners have also been under development that the quality of images became more qualified than ever. The algorithms need to be both time and memory-efficient, in addition to that, the model should be able to extract as much information as possible from the large images. High-resolution images that can yield from 20,000x20,000 to 100,000x100,000 pixels can be captured to describe a whole slide of a tissue sample (Vu et al., 2019). The algorithms are confronted with memory issues and computation time to iterate over all large-sized images in a dataset.

1.4 Solutions

In the history of machine learning developments, data is the leading cause of accurate results of a system. One of the most important events of the 1960s was the advent of the semi-supervised learning approach (Scudder, 1965). A heuristic approach of self-training takes the first steps in history as the oldest method to semi-supervised learning setup (Chapelle et al., 2006) and the examples started following later. Nevertheless, Vapnik and Chervonenkis (1974) makes the first formal introduction of the transductive learning framework. Semi-supervised setup deviates from supervised learning by taking advantage of unlabeled data. The idea is to benefit from the few amount of labeled data while the unlabeled data is used in conjunction with the small amount of labeled data which leads to an improvement in the learning accuracy. Surveys such as that conducted by Zhou (2006) and Van Engelen and Hoos (2020) show that the semi-supervised approach contributes to a partial decrease in the intensive workload of data labeling.

Semi-supervised learning refers to either transductive (Gammerman et al., 2013) or inductive learning (Zhu, 2005). Transductive learning tries to infer the correct label of the given unlabeled data while the goal of inductive learning is to infer the correct mapping from the training data to the label. The methods are used in semi-supervised learning are as follows. The generative modeling approach seeks to estimate $p(x|y)$, which represents the probability distribution of the data points that belong to each class. The distribution that the model tries to approach is obtained by the Bayes' rule. The following concept of

semi-supervised learning is low-density separation (Zhu, 2005). A considerable number of methods seek to build boundaries of regions with the use of few data points without necessarily picking labeled or unlabeled data. Another approach is graph-based methods that construct a graph representation with nodes containing both the labeled and the unlabeled data (Zhu, 2005). Apart from mentioning the details of constructing the graph representations, the method attempts to connect each data point by the interaction of its k-nearest neighbors (Fix and Hodges, 1951) or some custom set distance metric ϵ.

However, semi-supervised architecture saves a major problem (finding a vast number of labeled images) with this kind of application, it does not bring any solution to the high dimensionality problem of the medical images. Bellman (1960) refers to this phenomenon as the curse of dimensionality when considering problems in dynamic programming. The common issue is, especially in classification, to end up with a sparse dataset with high dimensionality while organizing the feature representation in machine learning applications. In a medical dataset, as mentioned in Section 1.3, an image may vary from $20,000$ by $20,000$ to $100,000$ by $100,000$, meaning that a single image may contain more than one million features. This high dimensionality issue would indeed reflect on the computational power and energy, as well as the time consumption in a training session. On the other hand, there may be an opportunity in data usage. Instead of a whole property, the most important part of data can be utilized. Feature extraction, *i.e.* dimensionality reduction (Van Der Maaten et al., 2009), plays a significant role. An important aspect of reduction in the dimensionality of the data is to employ only the substantial part of the input into training.

According to the present challenges (see Section 1.3), a primary concern of analyzing a large-sized image to classify whether it has a tumor or a certain cancer subtype is to include a large image without knowing the crucial aspects of the given input to the model. This may result in model confusion or unnecessary computational time. One of the solutions can be named dimension reduction. Dimension reduction is a type of transformation of the data from high-dimensional space to low dimensional. The lower-dimensional representation not only contributes to the training time and computational power need but also retains the meaningful properties of the data, ideally to the intrinsic dimensions or the plainest (Ding et al., 2002). The extracted features may pretend to the representatives

of the image. Dimension reduction can be achieved with the use of most two common algorithms, namely principal component analysis (PCA) (Pearson, 1901) and autoencoders (Kramer, 1991). PCA performs a linear mapping of the data. Thus, it contributes to machine learning problems that involve linear feature representation within the data. On the contrary, Kramer (1991) states that an auto-associative network can discern the significant patterns in data, can be greatly facilitated by reducing dimensionality in non-linear data which puts the autoencoders in a more favorable position when dealing with a non-linear dataset as in the medical domain. The main idea behind the encoding is to learn a latent representation for a set of data which can be also referred to as dimensionality reduction. Yet, the autoencoders are mainly used for dimensionality reduction (Hinton and Salakhutdinov, 2006), data denoising (Van Der Maaten et al., 2009; Gondara, 2016), information retrieval (Salakhutdinov and Hinton, 2009) or image compression (Theis et al., 2017; Ballé et al., 2016). It was also highlighted that the autoencoders can help denoising or reducing the number of dimensions of the data but it would not fully solve the challenges in the image classification. A richer latent representation of a generative model creates room for data generation that is a more preferable environment where the data is lack.

As described earlier in the semi-supervised setup, generative modeling offers a great deal to synthesize data in various ways (Xu et al., 2015a; Creswell et al., 2018). The reason generative modeling is successful on sampling is that unlike discriminative modeling focuses on learning a predictor given the observation, generative modeling emphasizes learning how the data is generated and reflects the underlying causal relations. One of the breakthrough contributions comes from (Kingma and Welling, 2013; Rezende et al., 2014) that they offer a reparametrization trick for making the training suitable of an autoencoder. Variational autoencoder (VAE) can both be present at the dimensionality reduction (encoding) and the data generation (sampling). A fundamental property of a VAE is to have the ability to generate data concerning the distribution of the latent variables. The model can produce samples once the model learns the latent space. One fundamental feature of this kind of approach is conditional sampling (Kingma et al., 2014) that leads to labeled data generation in an environment that lacks data. Moreover, VAE belongs to the variational Bayesian method that is carried out for latent representation learning (Kingma and Welling, 2019). The objective function is constituted as the difference between the causal

relation (Kullback and Leibler, 1951) and the reconstruction error in VAE. If the focus will be only on the reconstruction error as in the standard autoencoder, an uneven distribution that does not belong to any of the observed data emerges. On the other hand, if the main attention will be only on the latent distribution being similar to the prior ($X \sim \mathcal{N}(0, 1)$), the latent representative distribution will be narrowed down. So, the VAE architecture differs from the other generative model approaches as it learns the prior distribution z that allows us to generate images from sampling instead of replicating. The causal relation works as a regularization to restrict the model to learn a close latent space representation to the encoded data while the reconstruction error forces to have similar results compared to the original data after decoding. VAE demands a lower-dimensional feature representation by reducing the size of the image but the decoder must turn back to the same size as the original input data. Regular medical image data will still hold its high dimensional property which would again cause a problem during the training. In response to this, Keeler et al. (1991) and Dietterich et al. (1997) explore the area of multiple instance learning (MIL).

Multiple instance learning is a type of supervised learning in machine learning. The actual term of multiple instance learning was first introduced by Dietterich et al. (1997), while the authors were investigating the problem of drug activity prediction. Instead of dealing with the instances individually, the method takes care of the learning session by having multiple instances in a bag. It flawlessly fits into the medical image classification domain since the bag setup relies on considering whether a specific type of disease is included in a bag (Herrera et al., 2016). Considering a binary MIL setup, the bag is labeled as *positive*, otherwise *negative*. MIL setup makes analyzing the low-level structure possible, meaning that instead of attempting to learn the whole cell at once, it will give a chance to look at the various parts of a cell in more detail. The model will be more aware of the actual causal relationship between the factor and the non-factor parts of the disease in the given cell image.

1.5 Overview

Considering the challenges of image classification procedure (Section 1.3) especially in the medical domain, semi-supervised learning has a significant potential to go beyond the

data tagging process, which holds most of the workload of experts. As the quality of the images gets better with the improvements in the mechanical devices, labeled data can only contribute to image classification to a degree. As the depth of the information of a single image can contain between 400,000 and more than a million, it makes the training procedure excessively harder and longer. To overcome such an issue, the MIL setup can be useful in terms of gaining the ability to look at the inner scale of the real images while handling the dimensions problem.

However, a major problem emerges with this kind of application that is to achieve a higher success rate with a few numbers of labeled data in a discriminative model (Wiens, 2003). Recent developments in generative models (Goodfellow et al., 2014; Kingma et al., 2014; Rezende et al., 2014) allow for rich and tractable generative architectures. Several researchers have reported that a conditional training for the probability of the target variable given the data alone, $p(y|x)$, does not provide a full understanding of the input (Tulyakov et al., 2017; Nalisnick et al., 2019). Because the results can easily change with a slight difference in the input (*e.g.*, noise). Most studies in the field of hybrid modeling have only focused on single image classification. Moreover, several attempts have been made to a hybrid approach in a MIL problem (Zafra et al., 2013; Jiao and Zare, 2017). A very recent study contributes to this area by considering a non-i.i.d environment (Zhang, 2021). Although some research has been carried out on the MIL problem in a hybrid setup, no single study exists which pays attention to uncertainty rating by utilizing $p(y|x)p(x)$ on the decision-making.

This dissertation attempts to show the cooperation of generative and discriminative models in a MIL setup. By design, the methodological approach taken in this study is a mixed methodology based on a variational autoencoder (Kingma and Welling, 2013; Rezende et al., 2014) and an Attention-based Deep MIL classifier (Ilse et al., 2018) in a hybrid setup. For this purpose, the data will be prepared concerning the binary MIL problem. An Attention-based MIL classifier (Ilse et al., 2018) is charged with returning a loss value if a labeled bag is served into the system during the training. The unlabeled bags will be utilized by the generative model to get more qualified reconstructed and sampled images. While the current proposals (Kingma et al., 2014; Rezende et al., 2014; Maaløe et al., 2016; Zhang, 2021) put more emphasis on the label ($p(x|y)p(y)$) that cares about the

generation. Another perspective, namely more utilization of both labeled and unlabeled data by the joint probability $p(y|x)p(x)$ that works for assigning better probabilities and contributes more on understanding the reason behind the selection will proceed in this approach (Nalisnick et al., 2019).

The main goal of this study is to shine new light on developing a deep generative framework for multiple-instance learning. To clarify the challenges (Section 1.3) among the machine learning approaches, it is now possible to automate the gathering of data at unprecedented scales. In many applications, the annotation of ground-truth labels is still done manually. Following that, learning from partially labeled data has emerged as a significant challenge in machine learning. Consequently, the gap between our data collecting and labeling capacity has increased further. As a solution to overcome this gap, weakly supervised and generative learning has been developed as an active topic of machine learning (Section 1.4).

One essential approach that draws interest over the recent years is MIL that establishes learning from labels available only for instance groups. As opposed to the standard supervised learning framework, the MIL approach allows the model to advance in many compelling real-world applications that require learning more structured data (Doran and Ray, 2016). For instance, in the text domain, the task may be to classify a single document that contains multiple passages or paragraphs. A bag of words treatment to the data may cause to lose the internal structure of the document. As a result, determining which texts relate to the category of interest might be difficult. Similarly, image classification becomes very challenging since, in some domains such as medical, it is reasonably common to coincide labels where the annotators might disagree on the final decision. For such issues, a richer representation of structure objects as sets is provided by the MIL setting.

The latter increasing trend is to include the abundant unlabeled data by employing a generative setup in a semi-supervised environment. Some generative techniques use semi-supervised learning to model class distributions and infer the label of each positive bag instance based on which of these two distributions best describes that instance (Adel et al., 2013; Foulds and Smyth, 2011). For data that naturally arises in MIL shape, generative modeling techniques are beneficial and productive. For numerous reasons, predicting with a generative model is particularly well suited to medical domains. Originally, expert domain

knowledge may be included in generative models in an understandable fashion, resulting in a desirable inductive bias in the modeling assumptions. Additionally, even with a limited training set, a model with excellent inductive bias that is elicited from specialists may provide remarkably accurate predictions. Most significantly, in contrast to discriminative learning, which requires either modeling the posterior class probabilities $p(y \mid x)$ or explicitly learning a decision boundary to divide the classes, the generative method to classification involves modeling the joint distribution of the input data $p(x, y)$. Generative models get their name from the fact that they may be sampled to generate synthetic data. Although generative classifiers are less accurate than discriminative classifiers when huge amounts of labeled data are available (Ng and Jordan, 2002), they can perform better when little amounts of data are available and can employ unlabeled data. Due to these reasons, this study underlies the motivation for developing a deep generative structure for the MIL environment.

In addition to the main objective, there are three supplemental directions of this study. (1) Integrating a VAE and an Attention-based Deep MIL classifier. (2) Investigating whether the hybrid learning approach can provide state-of-the-art performance in the semi-supervised setting. (3) Evaluating the proposed approach on the semi-supervised scenario and comparing it with baselines.

Drawing upon three stands of research into hybrid modeling in a MIL problem, this study help to address that the semi-supervised MIL VAE can utilize the unlabeled data and provide an uncertainty rate by learning the true distribution of the training data. Therefore, the method yields a higher loss rate along with the classifier prediction for the unseen distribution of the data. Although this uncertainty rate can be used for unseen distributions, it can be utilized while observing true/false classification rates. An additional offer of this work is to employ the attention mechanism in the MIL setup. One of the fundamental aspects of the attention mechanism is its capacity to recognize the most relevant information to accomplish a task, including an increment in the performance (Ilse et al., 2018; Shi et al., 2020). Attention mechanism serves as an importance distribution over the instances which helps to focus on the effective factors in a bag.

This study provides an exciting opportunity to advance the understanding of how the generative models can contribute to the discriminative model performance considering

accuracy and certainty on the decisions. In order to consolidate the results, a combination of quantitative and qualitative approaches was used in the data analysis. The results show that there are several important areas where this study makes an original contribution to the decision-making where labeled data is insufficient for a discriminative model in a MIL setup. Furthermore, the attention mechanism that is integrated into the classifier produces importance weights for the instances in a bag which shows the relevance of the case and the effect on finalizing the bag label. Additionally, the generative part of the proposed method can provide samples regarding the original data.

The thesis does not engage with competing against a discriminative model using all the data. This thesis intends to determine the extent to which the proposed approach achieves better accuracy rates than a single classifier under the insufficient number of labeled data and whether the hybrid structure provides a certainty rate on the decision-making. Due to practical constraints, this dissertation cannot provide a comprehensive comparison with similar methods.

This dissertation has been organized in the following way. The overall structure of the study takes the form of six chapters, including the introductory section. Chapter two begins by laying out the theoretical dimensions of the research and looks at how the components of semi-supervised MIL VAE evolve by sharing recent studies that have a common approach. The third chapter is concerned with the methodology used for this study. The fourth section presents the findings of the research, focusing on the three key themes that are; (1) integrating a VAE and an Attention-based Deep MIL classifier, (2) whether the hybrid approach can provide state-of-the-art performance on the decision-making in a semi-supervised setting, (3) whether the proposed approach performs better in classification than a discriminative model alone in a MIL setup. Chapter 5 analyses the results of qualitative and quantitative experiments and focus the positive and the negative side of the proposed methods. The final chapter draws upon the entire thesis, tying up the various theoretical and empirical strands in order to consolidate the findings and possible future works.

Throughout this dissertation, the acronyms AD-MIL and ssMILVAE will be used to refer to Attention-based Deep MIL and semi-supervised MIL VAE, respectively. Moreover, the proposed methods (Section 3.3.4) will be abbreviated by their first letter as an addition

at the end of ssMILVAE (*e.g.*, ssMILVAE$_p$ for the plain approach in Section 3.3.4.1). Besides the abbreviations, in favor of clarity in mentioning and comparing the models including the loss results, ELBO (or $\log p(x)$) and MIL loss will be used to refer to the VAE and AD-MIL losses, respectively. Furthermore, the ssMILVAE loss results will be brought down into ELBO and MIL losses in some sections where the model components are compared with the baselines. For simplicity, whenever the results talk about *the proposed model loss* should be understood as ELBO + MIL (or $\log p(x, y)$).

2 Background

In classic supervised learning tasks, we are given an ordered set of l labelled data points, denoted by $\mathcal{D} = \{x_i, y_i\}$. Each data point (x_i, y_i) is made up of an item $x_i \in \mathcal{X}$ from a specific input space $\mathcal{X}$ and a label y_i, where y_i can belong to regression or classification problems. Supervised learning approaches aim to infer a function that can correctly predict the label y of any previously unknown input x based on a collection of these data points, commonly referred as the *training data*.

However, there occurs another fruitful collection of data points $\mathcal{D} = (x_i)$, the labels of which are unknown in many real-world classification problems. For example, the data points for which are conjectured, commonly referred to as the *test data*, are not defined. Half-controlled classification approaches try to use unlabeled data points to build a model whose performance goes above the accomplishment of the conventional architecture generated by just utilizing the labeled data.

Unlabeled data can contribute to the performance of a classifier in many cases. Considering an example of a deep neural network architecture that classifies between certain images ($x_i \in \mathcal{D}$) of objects that can successfully predict the proper class of image with a high probability ($p(y|\mathbf{x})$, Szegedy et al. (2014) show that the model performance can easily be manipulated by adding a certain imperceptible perturbation. This reveals a crucial property that is absent in the systems where they utilize only labeled data $p(y|\mathbf{x}$ (Caruana and Niculescu-Mizil, 2006). One of the notable deficiencies of supervised learning is that the conditional distribution lacks the semantic comprehension of the data. Lakshminarayanan et al. (2016) and Nalisnick et al. (2019) demonstrate that discriminative modeling alone is not enough to create a system that can be relied on how to make a decision containing uncertainty about the environment. They, further, point out that distribution over objects, $p(\mathbf{x})$, can be utilized. In some areas of expertise, specifically in the medical domain, a large amount of unlabeled data exists next to the labeled data. However, the weak annotations in labeled images (Lu et al., 2019) and the high dimensional images (*i.e.*, medical domain emerge a solution called Multiple Instance Learning (MIL). Keeler et al. (1991) first explore the area of MIL in his work. Next, Dietterich et al. (1997) introduce the actual term while they investigated the drug activity prediction problem.

MIL algorithms can generally be split into three groups based on whether the algorithm focuses on an instance, bag, or embedding level (Amores, 2013; Ilse et al., 2020). The first group specifically deals with building an instance-based classifier under the assumption that a positive bag contains at least one positive target in a bag while a negative includes none (Maron and Lozano-Pérez, 1998; Zhang and Goldman, 2001; Andrews et al., 2002; Bunescu and Mooney, 2007). The second approach looks for similarities among bags by using different methods such as distances (Wang and Zucker, 2000; Zhang et al., 2007; Belongie et al., 2002), and kernels (Gärtner et al., 2002). The last approach operates in the same manner as the instance level except the ordering of the modeling architecture (Bunescu and Mooney, 2007).

With the advent of deep neural networks, the classifiers could be trained from end to end. Particularly, deep neural networks can learn which characteristics best reflect within the provided training data. Many deep learning techniques are based on this concept: networks comprised of many layers that discover a mapping from the input space (e.g., data) to the output space (e.g., class label) while learning successively higher-level characteristics (Zupan, 1994).

Numerous studies have attempted to employ several neural network architectures in a MIL problem. Li et al. (2012) utilize parallelism and feature extraction by integrating fully connected layers. With the use of Convolutional Neural Networks (CNNs) (Lecun et al., 1998; O'Shea and Nash, 2015), Pinheiro and Collobert (2015) can differentiate between various classes at a pixel level and segments the objects with weakly supervision including transfer learning (Athiwaratkun and Kang, 2015). Pathak et al. (2014) analyze the input dimension flexibility of CNNs while they target the weakly supervised image segmentation problem by computing the multi-class logistic loss. Feng and Zhou (2017) describe a deep neural network architecture that generates instance representation. Wang et al. (2018) revisit the MIL problem with a neural network approach by utilizing a permutation-invariant pooling operation at bag level. Oquab et al. (2014) operate max-pooling to hypothesize the object location by the maximum score. Other alternative pooling operations vary such as ISR (Keeler et al., 1991), generalized mean (Ramon and De Raedt, 2000), and Noisy-or (Zhang et al., 2005). However, these operations require pre-defined functions. Therefore, they are not trainable. More flexible pooling functions

are proposed later as learnable pooling in conjunction with bag+instance loss (Zhou et al., 2017), and a fully connected Conditional Random Field (CRF) (Chen et al., 2014).

Despite the fact that a large and growing body of literature shows incentive upon the attention mechanism in various tasks (Xu et al., 2015b; Zhang et al., 2019; Wang et al., 2017; Xu et al., 2018; Xu et al., 2019), less effort is seen on attention mechanisms for MIL. Pappas and Popescu-Belis (2014) conduct a novel multiple instance regression (MIR) model that makes less simplifying assumptions by assigning importance weights to each instance in a bag. Subsequently, Pappas and Popescu-Belis (2017) extend the previous idea by employing a single neural net to learn the instance weights. Ilse et al. (2018) add one more layer to the attention network to learn attention weights by attaching a sigmoid function to predict the bag class. The attention weights act as an importance weight for each instance that shows how much contribution a single instance produces. They further use them to create a heatmap in the medical domain. In this study, the classifier component consists of the same structure as Ilse et al. (2018) with regards to its performance comparing the other approaches (LeCun et al., 1998; Andrews et al., 2002). Shi et al. (2020) extend the idea by connecting the loss function to the attention mechanism which helps the model produce instance and bag labels at once with the attention weights.

Among unsupervised generative algorithms, Hinton et al. (1995) utilize a recognition model to approximate the true posterior. The approach results in two flaws. First, the recognition weights training becomes wasteful since the process includes the parts where there is no data. Second, the recognition weights only approximately follow the gradient of the variational bound on the log probability of the data which leads to incorrect mode-averaging. To conquer such issues, variational Bayesian methods are consulted for intractable integrals arising in machine learning.

Variational Bayesian methods are a class of techniques used in Bayesian inference and machine learning to approximate intractable integrals. They are frequently employed in complicated statistical models that include observable variables (commonly referred to as "data"), unknown parameters, and latent variables, as well as many forms of connections between the three categories of random variables, as defined by a graphical model. There is noticeably an interest in the variational inference lately (Hoffman et al., 2013). Ranganath

et al. (2014) develop a new algorithm for the variational inference that is a stochastic optimization of the ELBO by Monte Carlo sampling from the variational posterior to compute the noisy gradient. They also bring several methods to reduce the variance of the gradient. Paisley et al. (2012) present an alternative stochastic optimization scheme that allows controlling the high variance of the gradient. They show how control schemes can reduce the high variance of the posterior approximation. Salimans and Knowles (2013) propose a flexible algorithm that minimizes the Kullback-Leibler divergence (Kullback and Leibler, 1951) of a given distribution to an intractable posterior. The algorithm supplies the flexibility upon approximation to any posterior distribution in the exponential family.

The semi-supervised MIL VAE architecture connects directed probabilistic models (since a part of the objective includes a variational objective), auto-encoders, and hybrid modeling because the model architecture utilizes $p(x)$ in its objective. Auto-encoder contains the reconstruction error to the loss criterion (Vincent et al., 2010). It is shown that the training criterion is the maximization of a lower bound. However, Bengio et al. (2013) point out that the reconstruction error is insufficient for learning meaningful representations on its own. Lately, Kingma and Welling (2013) and Rezende et al. (2014) question the intractability of the marginal likelihood and offer the reparameterization trick to connect the auto-encoders and directed probabilistic models. During the study, the proposed approach incorporates a VAE architecture.

One of the first initiatives of a combination of generative and predictive models is seen from Jaakkola et al. (1999) by defining a kernel function to bridge between the generative model and the classifier such as Support Vector Machines (SVMs) (Cortes and Vapnik, 1995). Some studies (Efron, 1975; Ng and Jordan, 2002) share insightful outputs by presenting the trade-offs between the generative and the predictive models. Raina et al. (2003) provide an in-depth analysis of the work of a hybrid model that performs even better than each component alone by sharing a small subset of parameters with the discriminative model. Soon after, Lasserre et al. (2006) share all parameters with both the generative and the predictive model. Recent developments in deep generative modeling and stochastic variational inference have enabled the neural networks to involve in the above-mentioned scheme. These advancements are seen in various studies (Kingma et al., 2014, Maaløe et al., 2016, Kuleshov and Ermon, 2017, Tulyakov et al., 2017, Gordon and

Hernández-Lobato, 2017, Rezende and Mohamed, 2015 Nalisnick et al., 2019).

As a result of this research done so far, to the best knowledge of the author, there has not been a publication same as the proposed methods. The closest structure comes from a very recent study that is conducted by Zhang (2021) that they utilize a VAE to build a hybrid model in a MIL problem. The proposed method differs as Zhang (2021) build an architecture that would both exploit from the labeled bags and unlabeled instances to hold a non-i.i.d environment in the training. Additionally, that study does not focus on providing an uncertainty rate on the decision-making. Hou et al. (2016) conduct another study employing the Expectation-Maximization-based method on a classification task in the medical domain. Similar work is seen for a segmentation task by Papandreou et al. (2015). Ghaffarzadegan (2018) have studied an interesting architecture of two generative models to learn the latent representations of negative and all instances and one classifier to distinguish between the positive and negative bags. Javadi et al. (2020) combine Independent Conditional Variational Auto Encoder (ICVAE) with a MIL network. They generate synthetic data while training against the insufficient number of the dataset. Even more interesting work is conducted by Zhao et al. (2020) as they combine a VAE with a generative adversarial network (VAE-GAN).

Together, these studies outline that although the hybrid structure is exploited in many pieces of research, a gap still exists in the MIL area. Overall, these studies provide important insights into the potential to integrate an Attention-Based MIL classifier into a generative framework for the MIL problem. According to the literature review done so far, there are no exhaustive studies that focus on utilizing $p(y|x)p(x)$ in a MIL problem. Moreover, given all that has been mentioned so far, one may suppose that a hybrid approach by combining a VAE and Attention-based MIL classifier will perform better in terms of accuracy and attention weights to the instances of a bag.

3 Methodology

In this section, each component of the semi-supervised MIL VAE approach is explained in detail. First, the variational auto-encoders (Kingma and Welling, 2013) is overviewed along with the loss criterion, then the Attention-based Deep MIL classifier (Ilse et al., 2018) is mentioned. Next, attention mechanism and pooling operations are examined. At last, these components will be merged as a hybrid model in a MIL setup. The semi-supervised MIL VAE architecture is produced with four different proposals. The testing of these four approaches will be done on the MNIST-BAGS and proceed with the most successful structure for the COLON CANCER dataset.

3.1 Variational Auto-Encoder (VAE)

3.1.1 Problem formulation

Given data X consists of M i.i.d sample of continuous or discrete variables x. A model generates the data by the involvement of a continuous random variable z, in some random process. The process involves two important steps. Initially, random variable z_i generation from some prior distribution $p_\theta(z)$. Then, the generation of x_i from some conditional distribution $p_\theta(x_i|z_i)$. As much as it looks like a classic optimization problem in a machine learning problem, the generation procedure for a single x is as follows.

$$\begin{aligned} P(X = x) &= \int p(X = x, Z = z)dz \\ &= \int p(X = x \mid z; \theta)p(z)dz \\ &= \int p_\theta\left(x; g(z; \theta), \sigma^2 * I\right) p_\theta(z; 0, I)dz \\ &\approx \frac{1}{M}\sum_{m=1}^{M} p_\theta\left(x; g\left(z_m; \theta\right), \sigma^2 * I\right) \qquad \text{where } z_m \sim \mathcal{N}(0, I) \end{aligned} \tag{3.1}$$

Equation 3.1 shows the probability of a single sample is the joint probability of x and z by marginalizing out Z. The integral approximation can be achieved by taking the average over M samples from $z \sim \mathcal{N}(0,\, I)$.

$$\log P(X) \approx \sum_{i=1}^{N} \log\left(\frac{1}{M}\sum_{m=1}^{M} p_\theta\left(x_i; g\left(z_m; \theta\right), \sigma^2 * I\right)\right) \tag{3.2}$$

However, an intractability problem appears when trying to have the log-likelihood of the model across all M observations since z_m is a high dimensional vector. The model needs to process a large amount of z values for each x_i example. Dealing with many values brings the curse of dimensionality (Bellman, 2015) in which each new dimension of z exponentially increases the number of samples required to estimate the volume of the space. It would be feasible for a small region, but it is not likely to coincide with such circumstances in a real-world example. If the size of the samples is few, then the approximation would be poor. Thus the integral is intractable which also makes the true posterior density $p_\theta(z|x) = p_\theta(x|z)p_\theta(z)/p_\theta(x)$ is intractable. Consequently, it concludes that an EM algorithm cannot be used. These kinds of intractability cases frequently appear in simple likelihood functions such as nonlinear hidden layers in a neural network.

Kingma and Welling (2013) offer three significant solutions to the above scenarios. The first solution is an efficient ML approximation or MAP estimation to the parameters θ. These parameters, which is commonly referred to as *the decoder*, also allow to simulate of the hidden random procedure and produce artificial data that maps onto the real data. Secondly, a component, namely *the encoder* or *the recognition* model that is an efficient posterior inference approximator of the latent variable z, given the data x. Thirdly, an efficient marginal inference approximation of the input data x where it can be useful in some computer applications needs sampling or data production.

3.1.2 The Evidence Lower Bound (ELBO)

The idea behind the variational inference is to adjust parameters ϕ of Q_ϕ to infer upon P as close as possible. Using the reverse KL divergence (Kullback and Leibler, 1951) can derive the closeness term between distributions by putting more importance on Q as it will try to avoid spreading the approximation.

$$KL\left(Q_\phi(Z \mid X) \| P(Z \mid X)\right) = \sum_{z \in Z} q_\phi(z \mid x) \log \frac{q_\phi(z \mid x)}{p(z \mid x)} \tag{3.3}$$

Substituting $p(z \mid x) = \frac{p(x,z)}{p(x)}$ into the Eq. 3.3;

$$\begin{aligned} KL(Q\|P) &= \sum_{z \in Z} q_\phi(z \mid x) \log \frac{q_\phi(z \mid x) p(x)}{p(z,x)} \\ &= \sum_{z \in Z} q_\phi(z \mid x) \left(\log \frac{q_\phi(z \mid x)}{p(z,x)} + \log p(x) \right) \\ &= \left(\sum_z q_\phi(z \mid x) \log \frac{q_\phi(z \mid x)}{p(z,x)} \right) + \left(\sum_z \log p(x) q_\phi(z \mid x) \right) \\ &= \left(\sum_z q_\phi(z \mid x) \log \frac{q_\phi(z \mid x)}{p(z,x)} \right) + \left(\log p(x) \sum_z q_\phi(z \mid x) \right) \\ &= \log p(x) + \left(\sum_z q_\phi(z \mid x) \log \frac{q_\phi(z \mid x)}{p(z,x)} \right) \end{aligned} \tag{3.4}$$

The term, $\sum_z q_\phi(z \mid x) \log \frac{q_\phi(z|x)}{p(z,x)}$, should be minimized since the goal is to minimize $KL(Q\|P)$ with regard to the parameters ϕ. So, Eq. 3.4, can be re-written considering $\log p(x)$ does not depend on ϕ.

$$\begin{aligned} \sum_z q_\phi(z \mid x) \log \frac{q_\phi(z \mid x)}{p(z,x)} &= \mathbb{E}_{z \sim Q_\phi(Z|X)} \left[\log \frac{q_\phi(z \mid x)}{p(z,x)} \right] \\ &= \mathbb{E}_Q \left[\log q_\phi(z \mid x) - \log p(x,z) \right] \\ &= \mathbb{E}_Q \left[\log q_\phi(z \mid x) - (\log p(x \mid z) + \log(p(z))) \right] \\ &= \mathbb{E}_Q \left[\log q_\phi(z \mid x) - \log p(x \mid z) - \log(p(z))) \right] \end{aligned} \tag{3.5}$$

The original goal is to find an approximation that makes $q_\phi(z|x)$ close to the true posterior. For this purpose, q_ϕ is varied in order to minimize the KL term between $q_\phi(z|x)$ and the posterior, $p(z|x)$. To accomplish such intention, the Equation 3.5 should be minimized. In other words, minimizing this quantity is equal to maximizing the $\mathcal{L}$. And, it can be re-written as,

$$\begin{aligned} \max \mathcal{L} &= -\sum_z q_\phi(z \mid x) \log \frac{q_\phi(z \mid x)}{p(z,x)} \\ &= \mathbb{E}_Q \left[-\log q_\phi(z \mid x) + \log p(x \mid z) + \log p(z) \right] \\ &= \mathbb{E}_Q \left[\log p(x \mid z) + \log \frac{p(z)}{q_\phi(z \mid x)} \right] \end{aligned} \tag{3.6}$$

Now, $\mathcal{L}$ can be split into two procedures. The LHS (Left Hand Side) term, can be expressed as the conditional likelihood, in other words, a stochastic decoder. The RHS (Right Hand Side) term represents the KL-divergence between the $Q_\phi(Z|X)$ and the posterior $P(Z)$. These quantities

can be observed by rearranging $\mathcal{L}$ that is also known as the *variational lower bound*,

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_Q \left[\log p(x \mid z) + \log \frac{p(z)}{q_\phi(z \mid x)}\right] \\ &= \mathbb{E}_Q[\log p(x \mid z)] + \sum^{K} q_\phi(z \mid x) \log \left[\frac{p(z)}{q_\phi(z \mid x)}\right] \\ &= \mathbb{E}_Q[\log p(x \mid z)] - KL(Q(Z \mid X) \| P(Z))\end{aligned} \tag{3.7}$$

Sampling $z \sim Q(Z|X)$ stands for the encoding process that extracts latent variables, z, from an observation x, while $x \sim P(X|Z)$ represents the decoding procedure that reconstructs the observation as close as possible to the data from the latent variables. In other words, $\mathcal{L}$ seeks for the quality of reconstruction. In addition to that, the KL divergence regularizes the distance between the approximation on the latent variables and the prior on Z.

By substituting $\mathcal{L}$ into the Eq. 3.3,

$$\begin{aligned}KL(Q \| P) &= \log p(x) - \mathcal{L} \\ \log p(x) &= \mathcal{L} + KL(Q \| P)\end{aligned} \tag{3.8}$$

The log-likelihood of a data point x under the true distribution, $\log p(x)$, is received that is the sum of $\mathcal{L}$ and a regularization term, $KL(Q\|P)$, that determines the distance between $Q(Z \mid X = x)$ and $P(Z \mid X = x)$ at a particular value of X. The reason $\mathcal{L}$ is a *lower bound* is that $\log p(x)$ must be greater than $\mathcal{L}$ since $KL(Q\|P) \geq 0$. It is also referred as *evidence lower bound (ELBO)* by a variant formulation:

$$\mathcal{L}(x) = \log p(x) - KL(Q(Z \mid X) \| P(Z \mid X)) = \mathbb{E}_Q[\log p(x \mid z)] - KL(Q(Z \mid X) \| P(Z)) \tag{3.9}$$

3.2 Multiple Instance Learning (MIL)

3.2.1 Problem Formulation

The goal of a supervised learning problem is to predict the value of a target variable, $y \in \{0, 1\}$, for a given instance, $x \in \mathbb{R}^D$. However, instead of a single instance representation of the input, a bag of instances, $X = \{x_1, ..., x_K\}$, appears in a form that is neither dependent nor ordered. Ilse et al. (2018) assume that the number of k instances can vary in different bags. Further, individual labels, $y_1, ..., y_k$, exist in a bag where y_k is a binary variable. During the training, nonetheless, those labels are not accessible as they stay unknown. Therefore, considering the

assumptions, reshaping the MIL problem into a more compact form by the maximum operator resembles as;

$$Y = \max_k \{y_k\} \tag{3.10}$$

For at least two factors, using a maximum operator over instance labels would cause problems in model optimization. To start with, the model would face with vanishing gradient issue for all gradient-based learning mechanisms. Subsequently, the formulation in Eq. 3.10 is convenient only if an instance-level classifier is practiced.

The bag probability $\theta(X)$ must be *permutation-invariant*, meaning that the model does not assume any spatial relationship between the features, since it is assumed that neither ordering nor reliance of instances within a bag. Accordingly, the MIL problem can be considered as a three-step approach for classifying a bag of instances. First, using a function f to transform the instances; secondly, a symmetric (permutation-invariant) function σ to combine the transformed instances; thirdly, a g function to transform the combined instances transformed by the function f. Lastly, the expressiveness of the score function is determined by the function classes chosen for f and g. The choice of the functions determines a distinct way to modeling the label probability that is *instance-level* and *embedding-level* approaches.

The instance-level approach carries out a transformation with function f that returns scores for each instance within a bag. Afterward, MIL pooling aggregates the individual scores to obtain $\theta(X)$. In the embedding level approach, examples are mapped by a function f to a low-dimensional representation. Then, MIL pooling obtains a bag representation that is not related to the number of instances in a bag. Lastly, a bag-level classifier processes the embeddings to provide $\theta(X)$.

In terms of bag level classification performance, Wang et al. (2018) states that the embedding level approach is more favored. Since the instance-level classifier evaluates the instances in a bag while the individual labels are unknown, it is likely to have the model trained insufficiently. Hence an additional error is observed in the final prediction. On the other hand, the embedding-level approach deals with increased bias through a joint representation of a bag.

3.2.2 MIL with Neural Networks

Some tasks such as an image or text analysis require further processing more than f being identity function. For this reason, a class of transformations that are run by neural networks $f_\psi(\cdot)$ with parameters ψ that convert the k-th instance into a low-dimensional representation,

$h_k = f_\psi(\mathbf{x}_k)$ is offered for the embedding-based approach, where $h_k \in \mathcal{H}$ such that $\mathcal{H} = \mathbb{R}^M$. The following transformation $g_\phi : \mathbf{H^K} \rightarrow [0, 1]$ decides the parameter $\phi(X)$. The transformation g_ϕ could be parameterized by a neural network in the embedding-based approach. The notion of utilizing neural networks to parameterize all transformations is interesting since the entire technique is arbitrarily flexible. Additionally, it can be trained end-to-end via backpropagation on the condition that the MIL pooling is differentiable.

3.2.3 Attention-based MIL Pooling

There exist two MIL pooling operators that assure the score function (*i.e.*, the bag probability) is symmetric (Ilse et al., 2018), namely, the maximum operator:

$$\forall_{m=1,\dots,M} : z_m = \max_{k=1,\dots,K} \{\mathbf{h}_{km}\}, \tag{3.11}$$

and the mean operator:

$$z = \frac{1}{K} \sum_{k=1}^{K} h_k. \tag{3.12}$$

Other operators can be found as, noisy-or (Maron and Lozano-Pérez, 1998), noisy-and (Kraus et al., 2016), the convex-maximum operator (i.e., log-sum-exp) (Ramon and De Raedt, 2000), Integrated Segmentation and Recognition (Keeler et al., 1991). Despite the differentiability of the proposed operators, those operators are pre-defined and non-trainable. For instance, the max-operator may be suitable in the instance-based method but not in the embedding-based method. Similarly, the mean operator is certainly a poor MIL pooling to aggregate instance scores, while it could compute the bag representation. Consequently, a flexible and adaptable MIL pooling might potentially produce higher outcomes. Such MIL pooling should ideally be interpretable as well.

Ilse et al. (2018) propose that a weighted average of instances (low-dimensional embeddings) is used with the weights generated by a neural network. Furthermore, the weights must add up to 1 to be independent of bag size. Then, the attention MIL pooling can be formulated as,

$$\mathbf{z} = \sum_{k=1}^{K} a_k \mathbf{h}_k, \tag{3.13}$$

where $H = \{\mathbf{h}_1, \ldots, \mathbf{h}_K\}$ and:

$$a_k = \frac{\exp\left\{\mathbf{w}^\top \tanh\left(\mathbf{V}\mathbf{h}_k^\top\right)\right\}}{\sum_{j=1}^{K} \exp\left\{\mathbf{w}^\top \tanh\left(\mathbf{V}\mathbf{h}_j^\top\right)\right\}}, \tag{3.14}$$

where $\mathbf{w} \in \mathbb{R}^{L\times 1}$ and $\mathbf{V} \in \mathbb{R}^{L\times M}$ are parameters. Moreover, this approach uses the hyperbolic tangent $\tanh(\cdot)$ element-wise non-linearity to include both negative and positive gradients to the flow. The suggested structure enables the discovery of (dis)similarities between instances.

3.3 Semi-supervised MIL VAE

3.3.1 Problem Formulation

Since only a subset of the observations has corresponding class labels in semi-supervised learning, the data is available in two forms as labeled and unlabeled. While unlabeled data is only for the generative descriptions, the latent variables of the labeled data should also be put into training by a classifier, namely a structure that also takes advantage of the label. Labeled data appears as pairs $(X, Y) = \{(x_1, y_1), \ldots, (x_k, y_k)\}$, with the i-th observation $x_i \in \mathbb{R}^D$, and the corresponding label, $y_i \in \{0, 1\}$. Each x in the data represents a bag with several instances. The goal is to build a model that not only utilizes the labeled data as in supervised learning but also exploits the unlabeled data. So that it gives better results than a single discriminative approach in cases where the number of labeled data is not sufficient.

3.3.2 Objective

There remain two cases to consider for this model. Mainly, the ELBO loss (see Section 3.1.2) is calculated for both labeled and unlabeled bags in VAE flow. Further, an additional loss comes from the log-likelihood that is added up to the ELBO if the hybrid architecture receives a labeled bag in the training session. The case where the label is missing is treated as a Vanilla VAE. Meaning the data does not meet with the classifier.

Equation 3.15 describes the joint probability of the data X and the labels Y. The hybrid architecture assures that the abundant unlabeled data is involved in the process. In this approach, the latent variables, $q_\phi(z_k \mid x_k)$, are used by the AD-MIL instead of the input data, X. It is worth noting that there is a certain relationship between X and Z that is shared by the parameters ϕ

in the network.

$$\begin{aligned} p(X,Y) &= p(Y \mid X)p(X) \\ &= \int p(X,Y,Z)p(Z)dZ \\ &= \int p(Y \mid X,Z)p(X \mid Z)p(Z)dZ \\ &= \int p(Y \mid Z)p(X \mid Z)p(Z)dZ \\ &\overset{i.i.d}{=} \int p(Y \mid Z) \left(\prod_{k=1}^{K} p\left(x_k \mid z_k\right) p\left(z_k\right) \right) dZ, \end{aligned} \tag{3.15}$$

where the z_k is the output of the encoder and $p(z_k)$ is the latent distribution, also known as the prior. $p(Y \mid Z)$ illustrates the training of the AD-MIL with the latent variables. The derivation of the objective comprises of two steps. Originally, the VAE loss (ELBO) is received by the model for the unlabeled data (Section 3.1.2). Following, the log-likelihood value is added to the ELBO once the model receives a labeled data:

$$\begin{aligned} p(Y \mid X) &= \int p(Y \mid Z)p(Z \mid X)dZ \\ &\overset{i.i.d}{=} \int p(Y \mid Z) \cdot \prod_{k=1}^{K} p\left(z_i \mid x_i\right) dZ, \end{aligned} \tag{3.16}$$

where $Z = z_1, \ldots, z_k$ and $z_i = f\left(x_i\right)$. In order to model Equation 3.17, the Direc's Delta function is utilized by including attention.

$$p(Y \mid X) = p(Y \mid f(x_1), \ldots, f(x_k)) \tag{3.17}$$

The loss function of the ssMILVAE model is constructed as the integral of the VAE and the AD-MIL losses over the data points. The VAE architecture utilizes a variational Bayes method to derive a tractable lower bound on the marginal log-likelihood (see Section 3.1.2 for a detailed derivation). An addition to the ELBO calculation, a log-likelihood (LL) loss is integrated for the latent variables of the input bags. By inserting the Equation 3.16 into the Equation 3.15 and

having the logarithm of the both sides of the equation;

$$\begin{aligned}\ln p(X,Y) &= \ln \int \frac{Q(Z \mid X)}{Q(Z \mid X)} p(Y \mid Z) \left(\prod_k p\left(x_k \mid z_k\right) p\left(z_k\right) \right) dZ \\ &\geqslant \int Q(Z \mid X) \left[\ln p\left(Y \mid Z\right) + \sum_k \ln p\left(x_k \mid z_k\right) p\left(z_k\right) - \sum_k \ln q\left(z_k \mid x_k\right) \right] dZ \\ &= \int Q(Z \mid X) \left[\ln p(Y \mid Z) + \sum_k \ln p\left(x_k \mid z_k\right) + \sum_k \left[\ln p\left(z_k\right) - \ln q\left(z_k \mid x_k\right) \right] \right] dZ,\end{aligned} \tag{3.18}$$

where $Q(Z \mid X) = \prod_k q_\phi\left(z_k \mid x_k\right)$. In the objective function (Equation 3.18), the discriminative model $p(y \mid z)$ contributes only when the model receives the labeled data which is an undesirable case since the amount of the unlabeled data surpasses the labeled input. All model and variational parameters should ideally learn in all circumstances. For this reason, an additional α parameter in front of the classifier loss in the objective function to remedy such cases with the best intentions of hoping it work as a natural regularizer. Now, the marginal likelihood for the whole dataset can be expressed as,

$$\mathcal{L}(x,y) = \mathbb{E}_{q_\phi(z|x)}\big[\mathcal{L}(x) + \alpha\left[-\log p(y \mid z)\right]\big], \tag{3.19}$$

where $\mathcal{L}(x)$ stands for the objective of a VAE (see Equation 3.9) and the hyper-parameter α is a regulatory weight to keep the balance between generative and purely discriminative learning in the hybrid approach (Kingma et al., 2014). The final objective incorporates the variational lower bounds of labeled and unlabeled cases. Assuming N labeled and M unlabeled examples, the following objective is obtained:

$$\mathcal{L}(x,y) = \alpha \sum_{n=1}^{N} -\log p(y_n \mid x_n) + \sum_{m=1}^{M} \mathcal{L}(x_m) \tag{3.20}$$

3.3.2.1 Prior Choices among Datasets

The distribution for the latent variables is chosen based on how to express the latent components in data. Typically, $\mathbf{z}$ is treated as a vector of continuous random variables, $\mathbf{z} \in \mathbb{R}^M$. In this case, the true posteriors are assumed that they are approximately Gaussian with an approximately diagonal covariance. Therefore, the choice of the latent priors are let as multivariate Gaussians with a diagonal covariance structure (Kingma and Welling, 2013):

$$\begin{aligned} q_\phi(\mathbf{z} \mid \mathbf{x}) &= \mathcal{N}\left(\mathbf{z} \mid \mu_\phi(\mathbf{x}), \operatorname{diag}\left[\sigma_\phi^2(\mathbf{x})\right]\right) \\ p(\mathbf{z}) &= \mathcal{N}(\mathbf{z} \mid 0, \mathbf{I}) \end{aligned} \tag{3.21}$$

where $\mu_\phi(\mathbf{x})$ and $\sigma^2_\phi(\mathbf{x})$ are outputs from the encoder network.

According to the MNIST dataset properties, the pixels were additionally binarized in the early stages of loading the dataset. The goal is to pick the closest prior to the original data to have better results in reconstruction. For this purpose, Bernoulli distribution (Uspensky, 1937) is applied $p_\theta(\mathbf{x} \mid \mathbf{z}) = \text{Ber}(\mathbf{x} \mid \theta(\mathbf{z}))$, for MNIST bags.

Salimans et al. (2017) brings a solution for computing the conditional probability of the observed pixel values. As in a VAE (Kingma et al., 2016), they assume a continuous distribution for the latent color intensity ν, which is then rounded to the nearest 8-bit representation to provide the observed sub-pixel value x. As demonstrated in Eq. 3.22, this continuous univariate distribution to be a blend of logistic distributions, which allows to quickly compute the probability on the observed discretized value x. Hence, discretized mixture of logistic distribution (Salimans et al., 2017) serves as the prior for the COLON CANCER bags.

$$\begin{aligned} \nu &\sim \sum_{i=1}^{K} \pi_i \,\text{logistic}\left(\mu_i, s_i\right) \\ P(x \mid \pi, \mu, s) &= \sum_{i=1}^{K} \pi_i \left[\sigma\left(\left(x + 0.5 - \mu_i\right)/s_i\right) - \sigma\left(\left(x - 0.5 - \mu_i\right)/s_i\right)\right] \end{aligned} \tag{3.22}$$

where $\sigma(.)$ is the logistic sigmoid function. In case of a edge 0, $x - 0.5$ is replaced by $-\infty$ but when the pixel is 255, $x + 0.5$ is replaced by ∞. Further, the number of mixtures (components) was set to 5 in the experiments.

The ELBO becomes with the given prior distributions as,

$$ELBO(\mathcal{D}; \theta, \phi) = \sum_{n=1}^{N} \Big[\ln \text{Op}\left(\mathbf{x}_n \mid \theta\left(\mathbf{z}_{\phi,n}\right)\right) + \left[\ln \mathcal{N}\left(\mathbf{z}_{\phi,n} \mid \mu_\phi\left(\mathbf{x}_n\right), \sigma^2_\phi\left(\mathbf{x}_n\right)\right) + \ln \mathcal{N}\left(\mathbf{z}_{\phi,n} \mid 0, \mathbf{I}\right)\right]\Big] \tag{3.23}$$

where *Op* shows the chosen distribution in the ELBO calculation and vary depending on the dataset.

3.3.3 Model Structure

Building a model that gives an embedding or feature representation of the data is commonly used method. The embeddings enable the grouping of similar observations in a latent feature space, allowing for accurate classification even with a small number of labels. Instead of linear embedding, VAE (see Section 3.1) by Kingma and Welling (2013) will be employed to supply more robust set of latent features from the model. The generative model will benefit from

$p(\mathbf{z}) = \mathcal{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I})$; $p_\theta(\mathbf{x} \mid \mathbf{z}) = f(\mathbf{x}; \mathbf{z}, \boldsymbol{\theta})$ where $f(\mathbf{x}; \mathbf{z}, \boldsymbol{\theta})$ is an applicable likelihood function such as a Gaussian or Bernoulli distribution whose probabilities are created by a non-linear transformation of a collection of latent variables z with parameters θ. This non-linear modification must allow the density model to capture unique representative features of data. The latent variables of a labeled bag that are obtained from the recognition (*encoder*) model are used as features to train a classifier that predicts the bag labels y. The classifier choice is AD-MIL (see Section 3.2) proposed by Ilse et al. (2018). The classifier performs a lower-dimensional space classification since it uses the latent variables of the labeled bag that the dimensionality is less than the observations.

3.3.4 Proposed Methods

The following sections introduce the four different architectures in the ssMILVAE approach. For clarity, the proposed methods will be abbreviated by their first letter as an addition at the end of ssMILVAE (*e.g.*, ssMILVAE$_p$ for the plain approach). For a simple non-repetitive explanation of the methods, the model components, in general, are as follows. Every model comprises a VAE and AD-MIL. In addition to this, the ssMILVAE$_a$ employs one more encoder to process the input data twice for different purposes.

3.3.4.1 Plain Approach

The ssMILVAE$_p$ is defined as a plain integration of a VAE and AD-MIL. The aim is to utilize abundant unlabeled data. The discriminative model of the hybrid approach deals with the latent variables if the labels are introduced to the model otherwise the method acts as a VAE that derives only $\log p(x)$ (see Equation 3.8). When the model receives a labeled batch, the objective becomes as in Equation 3.18.

3.3.4.2 Disentangling Approach

The ssMILVAE$_d$ seeks for separated latent representations to be dealt with. Accordingly, the disentangling method refers to the latent variables that are chunked into three tensors after the encoding either labeled or unlabeled bags, namely z_1, z_2, z_3. These three detached latent spaces are evaluated in the following way. The connection of z_1 and z_2 are decoded in the VAE flow as in Equation 3.7. The only difference is that the amount of the latent representatives decreases in one-third of the ssMILVAE$_p$ approach. If a labeled data is received by the model, z_2 and z_3 are shared with AD-MIL to receive the classifier loss. The latent variables z_2 become a commonly shared variable tensor both in the generative and the discriminative models.

3.3.4.3 Separate Optimizer Approach

The ssMILVAE$_s$ setup is constructed the same as the ssMILVAE$_p$ (see Sec. 3.3.4.1). But instead of sharing the same optimizer, the proposed approach uses two optimizers, one for the VAE and the other for the AD-MIL. Both optimizers are defined as an Adam optimizer (Kingma and Ba, 2014). The generative optimizer parameters are found by a grid search experiment, and the classifier parameters are applied according to Ilse et al. (2018) findings. Parameters may be found in the Table A1.5 in Appendix.

3.3.4.4 Auxiliary Encoder Approach

The ssMILVAE$_a$ approach integrates an additional encoder for the input data, $q_{\phi_v}(z_v \mid x)$. The training is as follows. The two encoders, q_{ϕ_x} and q_{ϕ_v}, process the same input X. Originally, the latent representatives that are encoded by q_{ϕ_x} is used in the regular VAE flow;

$$\mathcal{L}(x) = \mathbb{E}_{q_{\phi_x(z_x|x)}, q_{\phi_h(z_v|x)}} \left[\log p_\theta(x \mid z_x) - KL(q_{\phi_x}(z_x \mid x \| p(z))\right], \tag{3.24}$$

where the parameters ϕ_x and ϕ_v represent the auxiliary and the regular encoders for the input data, respectively. Then, the latent variables that are received from the auxiliary encoder are introduced with the AD-MIL:

$$\begin{aligned} p(Y \mid X) &= \int p(Y \mid Z) p(Z \mid X) dZ \\ &= \int p(Y \mid Z_v) \cdot \prod_{k=1}^{K} p\left(z_{v_i} \mid x_i\right) dZ, \end{aligned} \tag{3.25}$$

where $z_{v_i} = f(x_i)$. Now, the bound on the marginal likelihood for the complete dataset can be defined in a hybrid setting by integrating the Equation 3.24 and 3.25,

$$\mathcal{L}(x, y) = \sum \mathcal{L}(x) - \alpha \sum \log p(y \mid z_v) \tag{3.26}$$

where the hyper-parameter α serves as a natural regularizer to balance the model between generative and entirely discriminative learning as in Equation 3.19. Besides, the auxiliary encoder architecture can be examined in Table A1.4 in Appendix.

4 Analysis

In the analysis, the goal is to evaluate the proposed approach, namely the ssMILVAE. The evaluation was conducted on two MIL datasets; an MNIST-based image dataset (MNIST-BAGS) and one real-life histopathology dataset(the COLON CANCER). During the experiments, three primary research questions are going to be verified: (i) Integrating a VAE and an Attention-based Deep MIL classifier. (ii) Investigating whether the hybrid learning approach can provide state-of-the-art performance in the semi-supervised setting. (iii) Evaluating the proposed approach on the semi-supervised scenario and comparing it with baselines.

To establish a fair comparison, a typical assessment approach is used, namely 10-fold cross-validation and five repetitions of each experiment. MNIST-BAGS are constructed with a predefined divide into training and test sets. To create test bags, images are sampled only from the MNIST set. During the training, the sampled images are solely used to construct the training bags from the MNIST training set. All classification models that are used in experiments are mimicked from previous works (Ilse et al., 2018; Sirinukunwattana et al., 2016). However, the first convolutional layers are needed to be adapted to smaller dimensions (see Table A1.3 in Appendix) since this approach seeks to classify the latent variables that come from the encoder. An upsampling operator (Oppenheim, 1999) is employed to scale up the dimensions of the latent variables to fit into the classifier structure. In addition, the MIL pooling layer is located before the last layer of the model, namely the embedded-based approach as in Ilse et al. (2018). Dimensions of the classifier are tried to be kept as same as possible with the original structure to make a fair comparison. As a result, there was not a parameter search upon the classifier. In this study, it is observed that different VAE architectures can significantly change the hybrid model performance. Hence, much focus is directed toward seeking various parameters (see Appendix) to find the optimum VAE architecture compatible with the MIL classifier.

Initially, a search was conducted to choose between different model types namely, fully linear, semi-linear, and convolutional, and lastly fully convolutional upon the MNIST set. The search results showed that fully linear and semi-linear/convolutional architectures of VAE do not perform well on the MNIST dataset at all. Accordingly, the fully convolutional

structure is carried out for both MNIST and Colon Cancer datasets. After exploring the VAE structure, the hidden layers were investigated by $3, 4$, and 5. Consequently, the number of channels is tested along with the learning and weight decay weights of the Adam optimizer. The β_1 and β_2 parameters of the optimizer are kept default as 0.9 and 0.999, respectively. Furthermore, all layers are initialized according to He et al. (2015), namely He initialization. Lastly, all experiments are run for 100 epochs and the best model was picked based on the validation accuracy and loss.

In order to assess the ssMILVAE, both qualitative and quantitative methods are used. To test the classifier performance, the proposed method is compared with the AD-MIL (Ilse et al., 2018). Subsequently, the VAE performance is conducted by running a plain VAE (Kingma and Welling, 2013) to compare the ELBO results. Moreover, four different model architectures provide their results depending on the best performances from parameter search. The experimental evaluation in COLON CANCER is conducted with the method that gives the highest accuracy and lowest loss rates in MNIST-BAGS. A detailed explanation about the model structures can be followed in Section 3.3.4. The experimental results consist of parameter search, 10-fold cross-validation experiments with 5 times running including accuracy, precision, recall, f-score, and AUC rates. Moreover, attention weights, reconstruction, and sampling visualization are demonstrated. In the final step of the comparison, ELBO results in histograms represent the distribution over truly classified bags, including the false classifications.

4.1 Data

This section examines the data sources used and the conceptual aspects of the variables employed in this study. The two variables of concern in this study are Multiple Instance Learning in a hybrid model and an uncertain metric on the decision-making. Following this, the first dataset will contribute to simulate a real-life example, furthermore, to decide on the best setup. The second dataset represents a real-life histopathology example of colon cancer.

4.1.1 The Modified NIST (MNIST) set

The MNIST database comprised of 70,000 handwritten digits in total that is constructed from the National Institute of Standards and Technology's (NIST) Special Database (SD) 3 and 1. 60,000 of the examples appears in the training set, and the rest, 10,000 of them are located in the test set. It is a subset dataset of a larger available from NIST that is originally determined SD-3 as its training set and SD-1 as its test set. However, the quality of the training and the test sets are not the same since SD-3 was collected among Census Bureau employees, whereas SD-1 among high-school students. Therefore, the modified version was constructed by mixing the SD-3 and 1 to get reliable learning results from the NIST.

The original black and white image sizes were normalized to preserve their aspect ratio in a 20x20 pixel box. Consequently, the resulting images contain greyscale levels according to the image interpolation technique that the normalization algorithm uses. There are three different versions of the set. In the first version of the dataset, the images were centered in a 28x28 field. In the following version, it was cropped down to 20x20 pixels. Lastly, the image size was reduced to 16x16 pixels. In this study, the initial version of 28x28 pixels is chosen in the interest of the comparability over Ilse et al. (2018) results. In addition, the pixels are binarized.

4.1.2 MNIST-BAGS for the MIL problem

The MNIST-BAGS preparation were done in the same approach as Ilse et al. (2018) did. An MNIST bag is constituted of a random number of Gaussian-distributed 28x28 grayscale handwritten digit images. The bag label is set as positive if it contains the number *9*, otherwise negative. The reason for picking the number *9* is considering a fair comparison environment with Attention-based MIL classifier (Ilse et al., 2018) alone. Furthermore, Ilse et al. (2018) investigate the influence of the number of bags and the number of instances in a bag. Thus, there is a fixed number of bags and instances in a bag, 1000 and 10, respectively. In addition, 50 bags were reserved as the validation set to observe the progress over the training session. Lastly, the number of test bags was set as 1000 during the evaluation.

4.1.3 The colon cancer dataset

This dataset consists of 100 Hematoxylin and Eosin (H&E) stained colorectal adenocarcinomas histology images. All images originally comprised of 500x500 pixels, then were cropped from the areas that are not overlap. Cropping images also helps to identify the areas with artifacts, overstaining, unsuccessful auto-focusing to display the outliers that are typically found in real scenarios. An Omnyx VL120 scanner obtained the whole-slide images at a pixel resolution of 0.55 µm/pixel. The images represent a variety of tissue displays from both normal and malignant regions.

Annotation of nuclei was conducted manually by a pathologist and a graduate student. 22,444 nuclei out of 29,756 were available for the classification task with class labels, namely epithelial, inflammatory, fibroblast, and miscellaneous. Colorectal cancer originates from the epithelial cells (Lee and Yun, 2010), as a result, it plays a significant role in identifying cancer. Another indicator is the inflammatory cells that emerge with a high chance (Terzić et al., 2010) in case of colorectal tumors. Lastly, a cancer-associated fibroblast (CAF) promotes tumorigenic features (Cirri and Chiarugi, 2011) that can be driven from normal fibroblasts. The rest of the labels represent adipocyte, endothelium, mitotic figure, a nucleus of necrotic (a dead cell) under miscellaneous class. The label distribution over the data is as follows; 7,722 epithelial, 5,712 fibroblast, 6,971 inflammatory, and 2,039 miscellaneous nuclei.

4.1.4 The colon cancer Bags for the MIL problem

A bag involves 27x27 image patches. Moreover, a bag is labeled as *positive* if it contains at least one nucleus from the epithelial class since colon cancer arises from the epithelial cells. The proposed transformations in Sirinukunwattana et al. (2016) are applied to the dataset before bag constructions. Moreover, additional changes are conveyed as in Ilse et al. (2018) followed by random adjustments to the amount of H&E by decomposing the RGB color of the tissue into the HE color space (Ruifrok et al., 2001), furthermore, multiplying the amount of HE for a pixel by two independently and identically distributed (i.i.d.) Gaussian random variables, then random rotations and mirrors every patch, lastly color normalization over patches.

4.2 Experimental Results

4.2.1 The MNIST-BAGS

During the experiments, the number of instances in a bag of MNIST are kept as 10 while the number of labeled bag size varies, namely 50, 1000, 3500, and 6000. 1000 labeled bags are taken as the main consideration among the experiments in terms of choosing the best approach and parameter search. Then, the same parameters are used for the rest of the labeled bag numbers. So, the proposed methods (see Section 3.3.4), are compared in 1000 labeled bags results. Additionally, a further experiment is conducted by including 50 labeled bags to illustrate a challenging environment for the model comparison.

In order to find the best parameters among the different approaches, an exhaustive search is made with a wide range of parameters. More detailed results may be examined in Table A1.1 in Appendix. To examine the contribution of the hybrid integration of the VAE and AD-MIL models, the loss rates of each component are separately investigated for the rest of the experiments. The results of the best runs from the parameter search are presented in the Figures 4.1 and 4.2. The figures include ELBO ($\log p(x)$) and ELBO + MIL ($\log p(x, y)$) of the hybrid approach along with the accuracy results, respectively. Additionally, they include the baseline models, specifically the plain VAE and AD-MIL rates in loss and accuracy rate results, correspondingly. As the first Figure 4.1 displays the $\log p(x, y)$ rate, the latter demonstrates the ELBO ($\log p(x)$) results. Each approach converges at a different round to its peak rates within the 100 epochs. The test results show that the ssMILVAE$_p$ approach performs best in terms of loss rates. The three model types except the ssMILVAE$_a$ achieve 95% accuracy following the Deep Attention MIL classifier. Nonetheless, it is worth mentioning that these results represent only one-fold test results with a single run. More detailed outcomes are shown in the following experiments.

Figure 4.2 displays the ELBO results of the best grid search outcome for the four approaches along with baselines. The VAE results are integrated with the upper part of the figure that displays the ELBO rates during the training and validation sessions. The below part compares the training and the validation accuracies by including the AD-MIL accuracy rates. Again, the ssMILVAE$_p$ among the suggested methods performs the best among

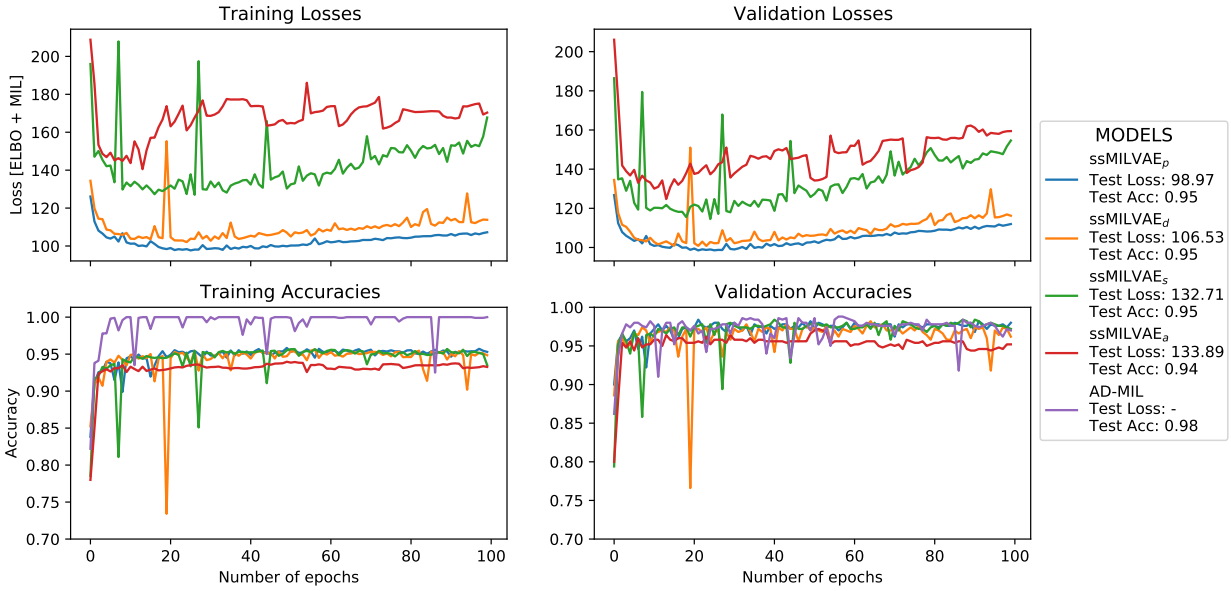


Figure 4.1: The training/validation loss rates from the best results of the parameter search of four proposed approaches including the baseline models.

the other type of approaches. It displays the closest results to the VAE and the AD-MIL classifier. For this reason, the ssMILVAE$_p$ approach is suggested for the rest of the experimental setups for the latter dataset.

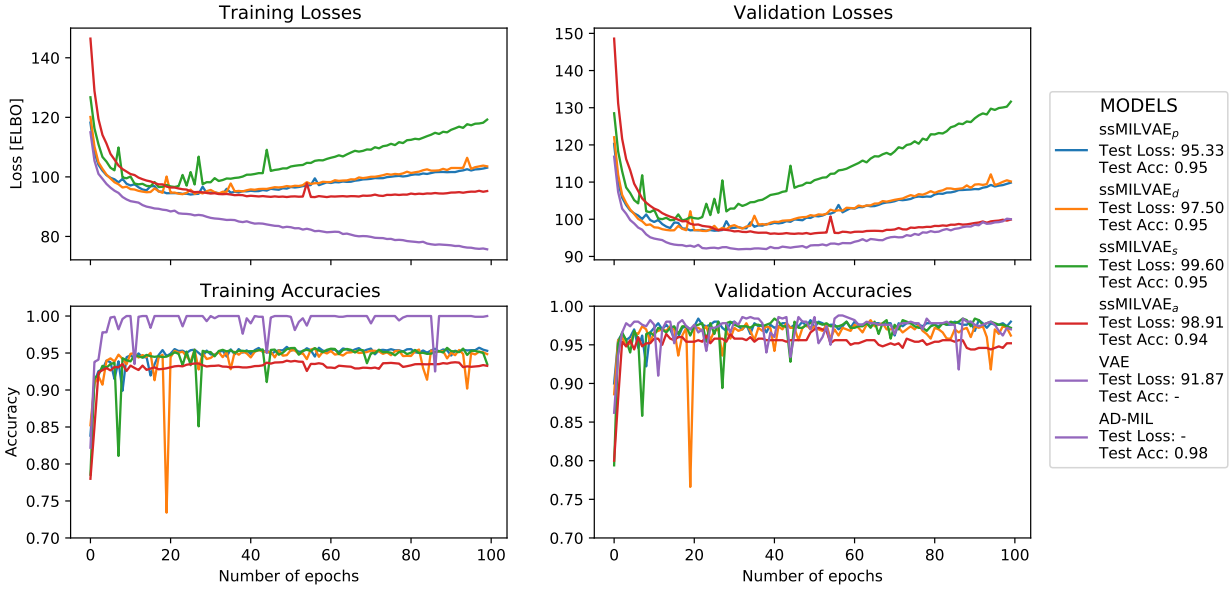


Figure 4.2: The training/validation ELBO rates from the best results of the parameter search of four proposed approaches including the baseline models.

Figure 4.3 compares the ROC (Receiver Operating Characteristic) and the AUC (Area under the ROC Curve) results for a bag size 10. The ROC analyses are examined with 50 labeled bags meanwhile the AUC results demonstrate each achievement with 50, 1000, 3500, and 6000 labeled bags. More detailed results can be found about the ROC tables for the rest of the number of labeled bags and the AUC findings in Figure A1.1 and Table A1.6 in Appendix, respectively. The discoveries from the analysis are as follows: Initially, the ssMILVAE$_p$ approach produces significantly higher AUC than the AD-MIL alone when an insufficient number of labeled data is shown to the models. The ROC chart confirms that. Furthermore, the AD-MIL classifier catches the proposed approach as the number of labeled bags increases as expected. As a result, these findings can be summarised that the integration of the VAE and the AD-MIL architectures cooperate adequately and serve moderately in the direction of the first research question. Further, the hybrid structure stands out particularly in cases where the number of labeled data is not adequate. Additionally, the results of this experiment show promising findings for developing a productive framework for MIL, which is the primary goal of this study, that is worth exploring further in this area.

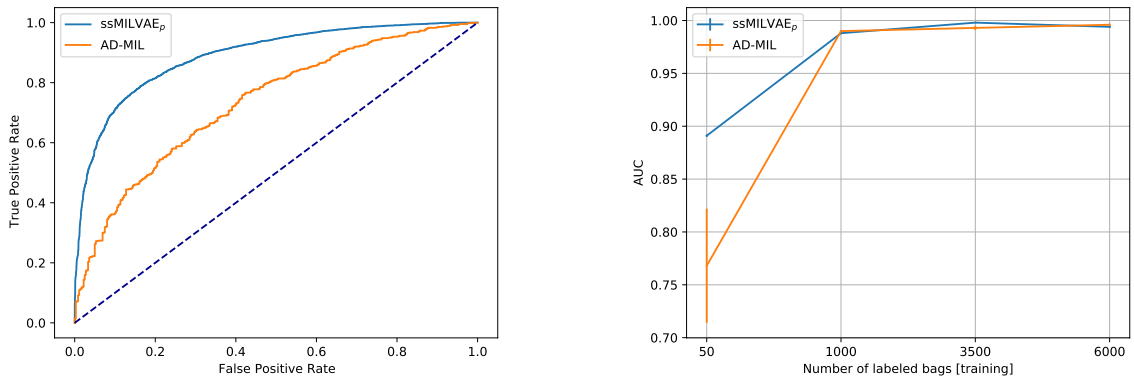
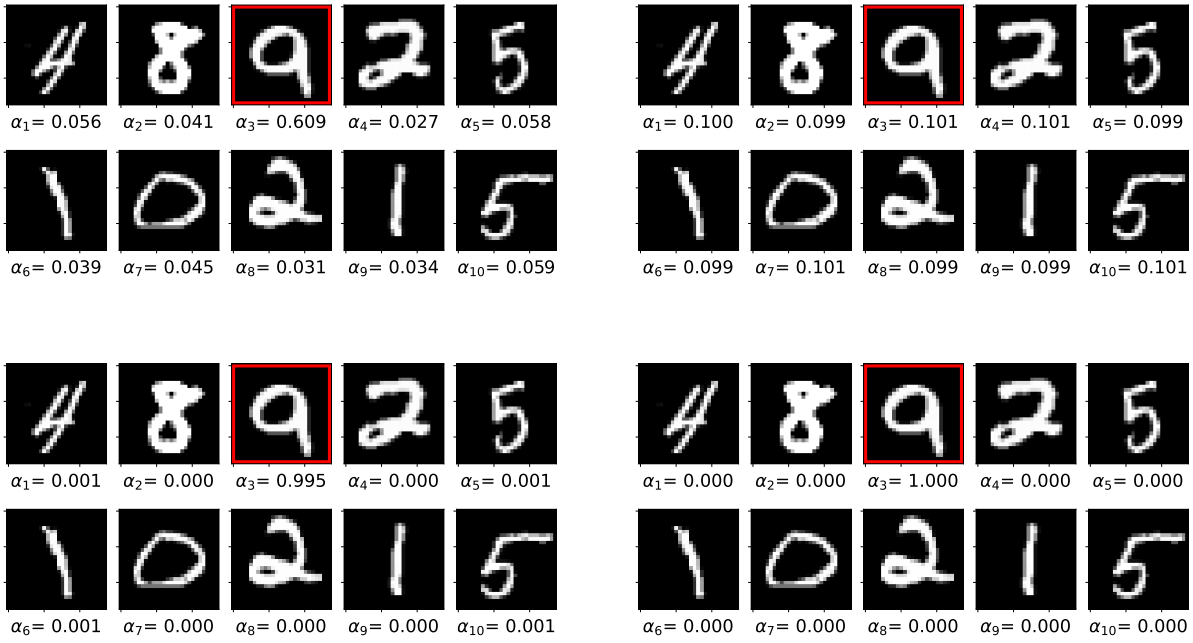


Figure 4.3: The ROC and AUC results for MNIST-BAGS from test data with 10 instances per training bag under an experiment conducted for 5 times.

The findings of this experiment reveal that the proposed technique is superior to others under a small sample size. Due to the fact that attention acts as a gradient update filter during backpropagation (Wang et al., 2017), instances with greater weights contribute more to learning the encoder network of instances. This is particularly essential given the few numbers of cases with medical imaging issues.

The following Figure 4.4 presents an exemplary result of the attention mechanism both from the ssMILVAE and the AD-MIL with the 50 and 1000 labeled bags at the upper and below part of the Figure, respectively. The bag consists of a single target value. The discriminative network assigns the corresponding attention weight to each digit. The bag is accurately anticipated as positive except by the AD-MIL that is trained with 50 labeled bag, and the single target value is highlighted. In addition, the attention mechanism performances differ among the models and training of a different number of bags, notably in the first example where the number of labeled bags is 50. The ssMILVAE$_p$ can give more accurate attention weights upon the instances in the bag. Better assignments also affect the prediction of the label of the bag, specifically in few labeled bags.



(a) ssMILVAE$_p$ **(b)** AD-MIL

Figure 4.4: Attention weights of a **positive** bag by the AD-MIL and the ssMILVAE$_p$, respectively. Upper bags represents the attention weight results from 50 labeled bags training while the bottom shows 1000 labeled bags results.

Figure 4.5 presents an example of a bag that consists of multiple targets values. The trained network assigns the corresponding attention weight to each digit. The bag is accurately anticipated as positive, and all nines are highlighted. It can be concluded that the ssMILVAE$_p$ provides better attention weight assignments among the multiples nines. The evidence implies that multiple target values can help in prediction performances even though very low attention weight assignments as seen in the AD-MIL when the number

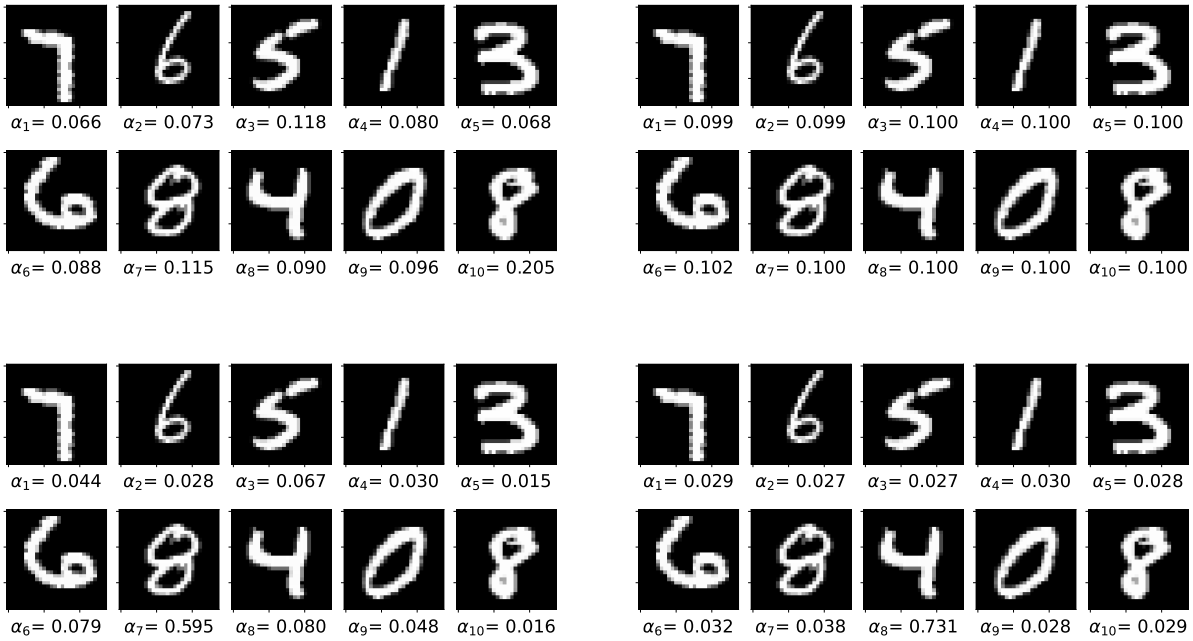
of labeled bags is equal to 50.

(a) ssMILVAE$_p$ **(b)** AD-MIL

Figure 4.5: Attention weights of a true classified **positive** bag by the AD-MIL and the ssMILVAE$_p$, respectively. Upper bags represents the attention weight results from 50 labeled bags training while the bottom shows 1000 labeled bags results.

Lastly, an example of a negative bag is represented by the Figure 4.6. The trained networks assign the corresponding attention weight to each digit. The bag is accurately anticipated as negative except by the AD-MIL that is trained with 50 labeled bag. A recurrent theme in the performance differences among the proposed and discriminative methods is a sense that the ssMILVAE$_p$ provides more accurate attention weights than the AD-MIL. More examples of how the attention mechanism works in a different number of labeled bags may be found in the Appendix.

Together these results provide important insights into the proposed approach consistency in satisfying the expectations with regards to the results in Ilse et al. (2018). According to the findings, it can be seen that the ssMILVAE$_p$ architecture specially deviates in comparison with the AD-MIL under a small amount of labeled MNIST-BAGS. The semi-supervised MIL VAE approach produces promising results with a few labeled bags as expected by utilizing unlabeled data. The following experiments were concerned with demonstrating how the ssMILVAE learns with only $\log p(x)$ and how the hybrid architecture $\log p(x, y)$ puts value on the decision-making. Additionally, a single VAE loss rate is included in



(a) ssMILVAE$_p$ **(b)** AD-MIL

Figure 4.6: Attention weights of a **negative** bag by the AD-MIL and the ssMILVAE$_p$, respectively. Upper bags represents the attention weight results from 50 labeled bags training while the bottom shows 1000 labeled bags results.

the experiment while comparing the positive and negative bag loss distribution on the histogram.

The next Figure 4.7 touchs upon a number of findings and exhibits them in histogram tables. The experiment aims to demonstrate how the classification results can change by the use of $\log p(x, y)$ in a hybrid architecture. The goal is to illustrate that the ssMILVAE$_p$ is capable to distinguish the true and false classified bags, yet $\log p(x)$ does not yield an adequate performance on it. The Subfigure 4.7a and 4.7b include the differences of $\log p(x)$ estimations between ssMILVAE$_p$ and VAE on the positive and the negative bags. Following, the Subfigures 4.7c and 4.7d display the difference between $\log p(x)$ and $\log p(x, y)$ both acquired from the ssMILVAE$_p$ approach including the true and false classified bags. Lastly, the Subfigures 4.7e and 4.7f focus on the false classified bags in detail.

The Subfigures 4.7a and 4.7b exhibit the ELBO distribution between the positive and negative MNIST-BAGS test data by the ssMILVAE$_p$ and VAE excluding the classification results. The results now provide evidence that the positive bags accumulate at 97.47 with 8.47 standard deviation, while the negative bags have a mean of 99.39 with 8.11

standard deviation in the hybrid MIL setup. Furthermore, the VAE results highlight similar results with a mean of 94.64, and a standard deviation of 7.92 in the positive bags. The negative bags gather at 96.43 with a 7.54 deviation. Although these findings yield the difference between positive and negative bags, they do not include the classifier predictions. However, the findings show that there is not any significant difference between the $\log p(x)$ approximations upon the positive and the negative bags.

The next Subfigures 4.7c and 4.7d present the $\log p(x)$ and the $\log p(x, y)$ estimations on the true and false classified test bags by the ssMILVAE$_p$ approach on histogram. The true classification range of $\log p(x)$ values takes up at 98.11 with 8.38 deviation while the false classified bags ELBO mean follows at a close range of 99.40 with 8.47 standard deviation. Due to the high classification rate of the model, a closer look of the false classified example rates are also shared in Subfigure 4.7e. However, it is not likely to infer a certain outcome just with the ELBO rates. Accordingly, the classification losses are also included in the following figures to investigate the certainty effect of the hybrid architecture.

The Subfigure 4.7d displays the loss distribution over the true and false classified test bags in MNIST-BAGS. The mean of the true classified bags equals 144.33 with a standard deviation of 336.74 whereas the false classification rates take up a mean of 3011.71. It is worth discussing these interesting facts revealed by the results that the ssMILVAE$_p$ architecture can help to infer the uncertainty on the decision-making procedure by $p(y|x)p(x)$. The proposed structure provides higher loss for false classified examples for the majority of the time that shows the certainty of the decision. Additionally, the difference can be used in two directions. First, to interpret the predictions as an uncertain decision. Secondly, as a suggestion system to incorporate the false classified example into a true prediction. For instance, the loss rate higher than 440 (mean plus the standard deviation) can be a sign for an uncertain decision according to the MNIST-BAGS test results. On the contrary, there might be a trial for turning the false classification into a true prediction. The findings provide promising results that this kind of support may boost the classification accuracy in a positive direction.

Taken together, the ssMILVAE$_p$ model shows higher performance than the other three approaches. This approach also competes against the AD-MIL head-to-head in accuracy,

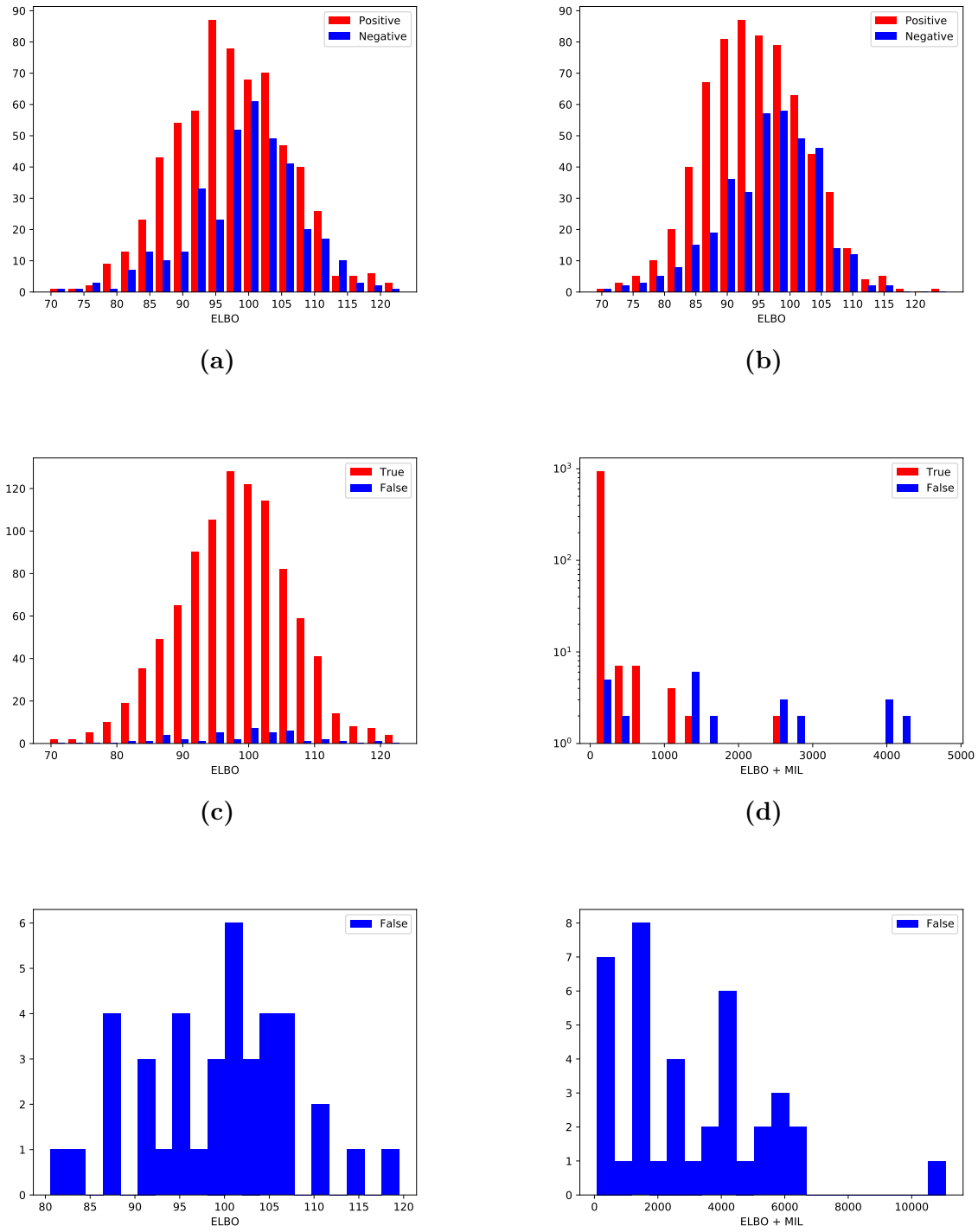


Figure 4.7: Experiment results to show the difference between $\log p(x)$ and $\log p(x, y)$ on the learning procedure. Subfigures 4.7a and 4.7b represent the VAE loss from both ssMILVAE$_p$ and VAE models over the positive and negative bags, respectively. Furthermore, Subfigures 4.7c and 4.7d show the VAE loss of the true and false classified bags from the ssMILVAE$_p$ model. And, Subfigures 4.7e and 4.7f focus on the false classified bag results.

and AUC when the labeled data consists of 1000, 3500, and 6000 bags. But the hybrid approach especially shows its capacity when the model only sees 50 labeled bags. The recommended model proceeds a better performance under a few numbers of labeled bags. The performance under few labeled data also reflects upon the attention weights to each instance in a bag. When the number of the labeled training set is low, the semi-supervised MIL VAE can handle assigning more reliable values to the target (nine in the analysis) and the rest of the values in a bag. The number of the labeled set also affects the bag label decision-making in identifying the target values and non-related values in the bag. When such circumstances occur that the training samples are small, the suggested approach makes a true decision for a positive bag whereas the AD-MIL cannot make the right decision for the bag label. Yet, the AD-MIL starts catching up with the semi-supervised MIL VAE and passes it at a small rate as the number of labeled data increases.

The results of this section indicate that the ssMILVAE$_p$ model architecture shows a strong signal about the research question of this study that is a healthy cooperation possibility between the integration of a VAE and AD-MIL in MNIST-BAGS. Furthermore, the hybrid structure also states that the model can give a certainty level on the decision-making to an extent by considering the cooperation between the loss functions of each architecture. It can be seen that false classified examples can be detected by considering the ELBO + MIL rate $\log p(x, y)$. In addition, the loss rate can identify (un)certain decisions of the classifier under given data. The following section, therefore, moves on to discuss the results of the COLON CANCER dataset.

4.2.2 The COLON CANCER

According to the findings in MNIST experiments, the ssMILVAE$_p$ (see Section 3.3.4.1 for detailed architecture) is settled for the rest of the COLON CANCER analysis. All experiments ran for a different number of labeled training sets, particularly 22, 92, and 162. Again, the goal is to seek to answer the research questions for the latter dataset. In this section, 22 labeled bags of training results will be mainly discussed. Further details and results about the rest of the numbers of labeled bags can be found in Appendix.

The Figure 4.8 displays the AUC results for COLON CANCER test data. The analysis shows that the ssMILVAE$_p$ performs better than the AD-MIL when the insufficient number

of labeled bags are shown to the models. The recommended design gets 6.4% higher accuracy, and higher AUC than the AD-MIL. Moreover, a lower variance is observed in the experiment results. The proposed method also receives the highest recall among the trials. The high recall is particularly crucial in the medical domain because false-negative results may have a serious impact on the decision-making of the experts, including patient death. From the chart, a trend can be seen that the AD-MIL performs better than the proposed approach as the number of labeled data increases. It was seen that a similar tendency in performance resembles in the MNIST set. However, a decline emerges in the ssMILVAE$_p$ performance. The evidence of the results may be related to the data complexity comes in with the COLON CANCER dataset. Unfortunately, a justification does not exist on this assumption. For this reason, another parameter search may be needed to have closer results to the AD-MIL by using the full dataset as labeled. This was not done for two reasons; the first argument is that competing with the classifier in full data exceeds the area of this research. The second idea regards a fair comparison between the models. For more detailed underlying results of the experiment containing accuracy, precision, recall, and AUC results, Table A2.3 may be visited. Further findings for the rest of the number of labeled bags are presented in Figure A2.1 in Appendix.

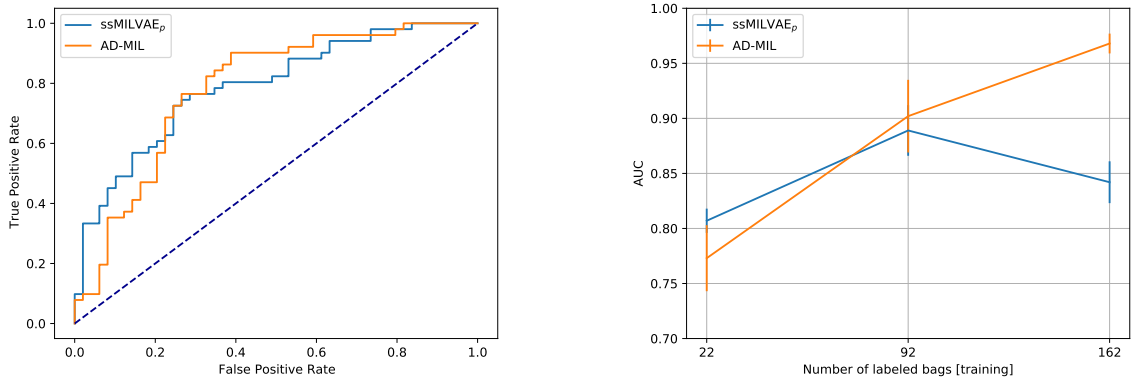
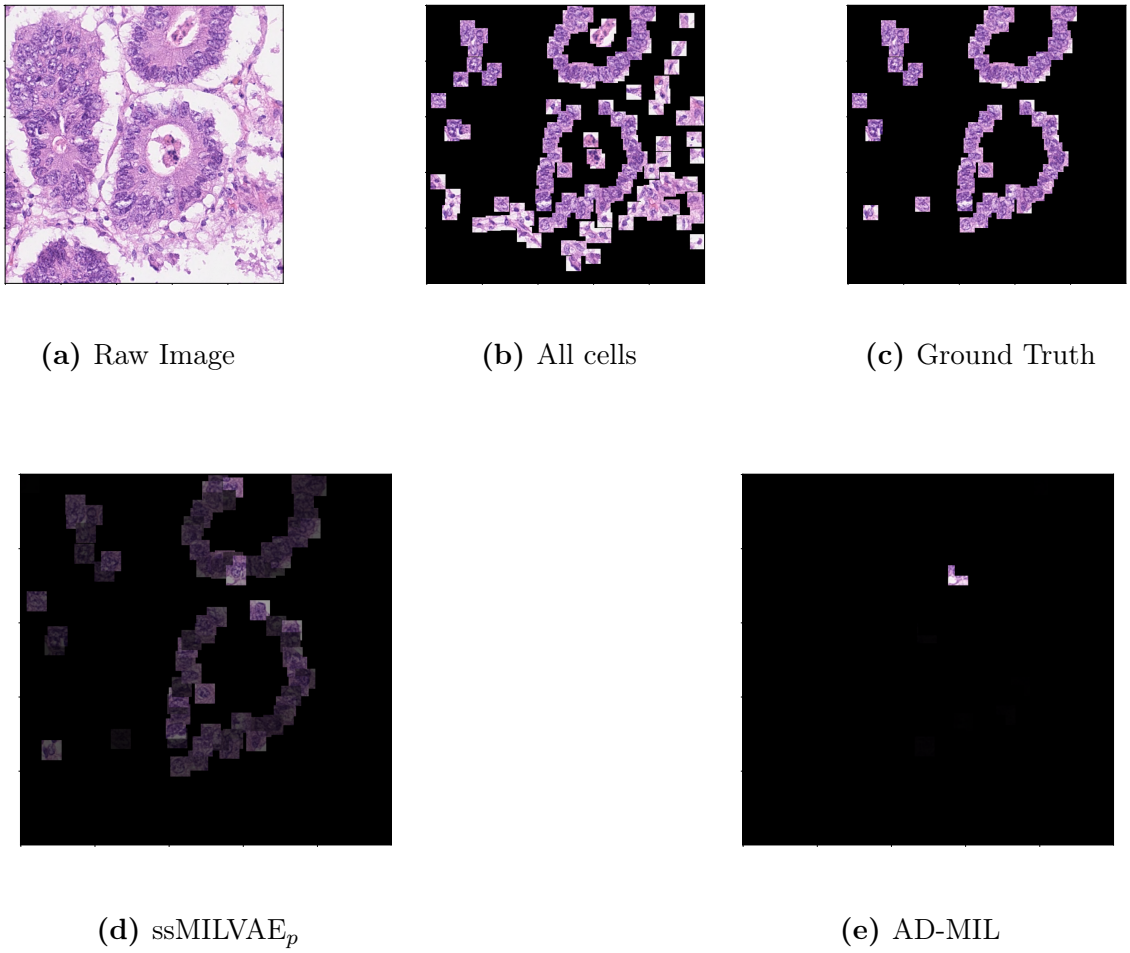


Figure 4.8: The test ROC and AUC for the COLON CANCER dataset under an experiment conducted for 5 times by 10-fold cross validation.

The following Figure 4.9 shows how the attention mechanism may be used to provide Region of Interests (ROIs). It depicts a histopathological picture splits into smaller patches containing single cells. A heatmap is generated by multiplying whole-cell images by their respective attention weight. Specifically, the attention weights are rescaled as $a'_k = (a_k - \min(\mathbf{a}))\,/(\max(\mathbf{a}) - \min(\mathbf{a}))$. Even though only image-level annotations are

utilized along with the small number of labeled data during training, there is an agreement between the heatmap in Figure 4.9d that is provided by the proposed method and the ground truth in Figure 4.9c. On the contrary, it can be observed that the only use of $p(y|x)$ in the discriminative model is not enough to imitate the general pattern of the ground truth.



(a) Raw Image **(b)** All cells **(c)** Ground Truth

(d) ssMILVAE$_p$ **(e)** AD-MIL

Figure 4.9: (a) H&E stained histology image. (b) 27x27 patches centered around all marked nuclei. (c) Ground truth: Patches that belong to the class epithelial. (d) Heatmap: Every patch from (b) multiplied by its corresponding attention weight, the attention weights are rescaled by using $a'_k = (a_k - \min(\mathbf{a}))\,/(\max(\mathbf{a}) - \min(\mathbf{a}))$

The following Figure 4.10 revisits the findings for $\log p(x, y)$ (as in the MNIST-BAGS that the hybrid structure provides. The Subfigure 4.10a and 4.10b include the differences of $\log p(x)$ estimations between ssMILVAE$_p$ and VAE on the positive and the negative bags. The subsequent Subfigures 4.10c and 4.10d display the difference between $\log p(x)$ and $\log p(x, y)$ both acquired from the ssMILVAE$_p$ approach including the true and false classified bags. Due to the few number of test bags in the COLON CANCER dataset, a

detailed visualization was not needed for misclassified examples.

The Subfigures 4.10a and 4.10b exhibit the ELBO distribution between the positive and negative COLON CANCER test data by the ssMILVAE$_p$ and VAE without considering the classification results. The chart illustrates that the positive bags accumulate at 9605.38 with a 31.87 standard deviation, while the position of the negative bag is around 9780.68 with a 238.66 standard deviation. Furthermore, the VAE results display similar results with a mean of 8510.92 and a standard deviation of 90.46 in the positive bags. The negative bags gather at 8651.56 with a 487.43 deviation. Although the Subfigure 4.10b shows that the VAE provides lower $\log p(x)$ for the positive and negative test bags, the low rates do not yield anything significant to distinguish between the models.

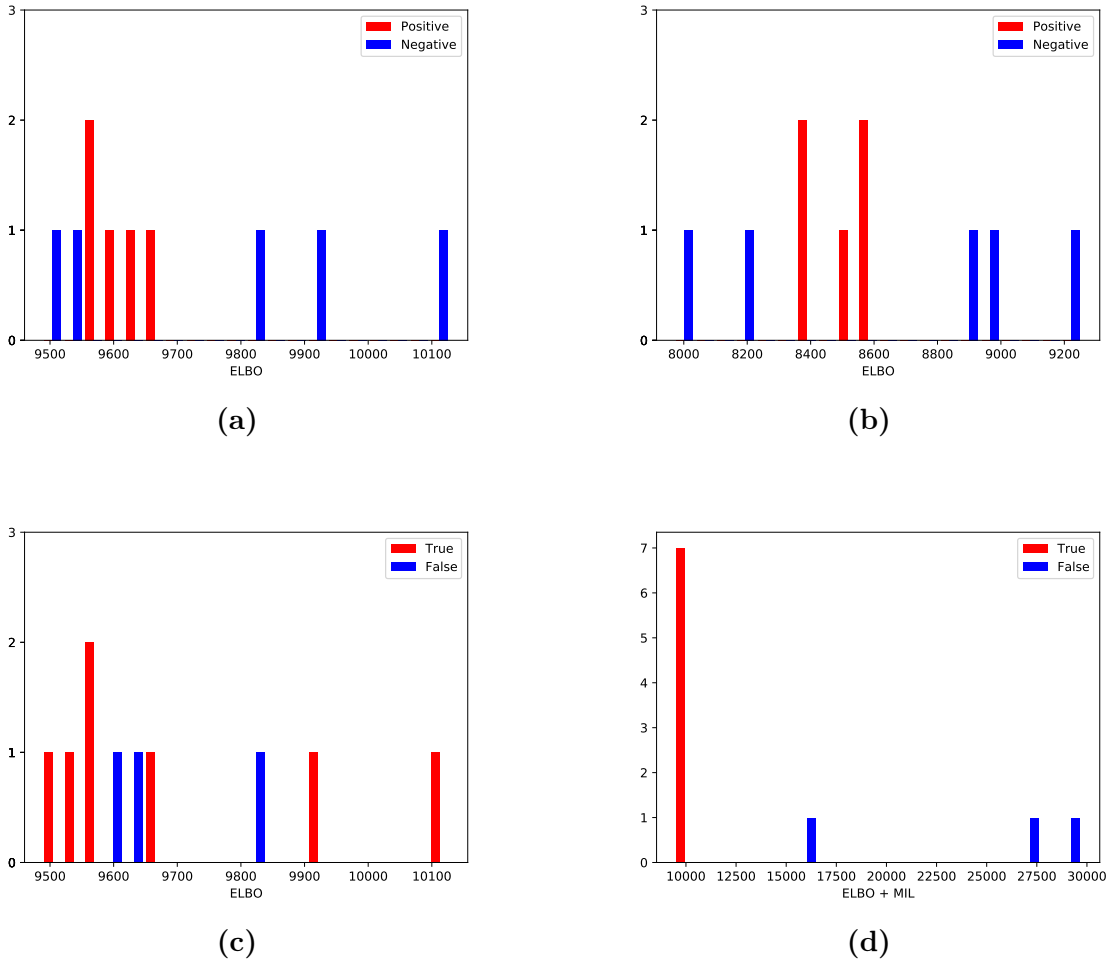


Figure 4.10: Experiment results to show the difference between $\log p(x)$ and $\log p(x, y)$ on the learning procedure. Subfigures 4.10a and 4.10b represent the VAE loss from both ssMILVAE$_p$ and VAE models over the positive and negative bags, respectively. Furthermore, Subfigures 4.10c and 4.10d show the VAE loss of the true and false classified bags from the ssMILVAE$_p$ model.

The next Subfigures 4.10c and 4.10d present the $\log p(x)$ and the $\log p(x, y)$ estimations on the true and false classified test bags by the ssMILVAE$_p$ approach on histogram. The Subfigure 4.10c illustrates that the true classification range of the ELBO values takes up at $9,696.24$ with a 218.72 deviation, while the false classified bags ELBO mean follows closely $9,685.55$ with a 102.61 standard deviation. However, from the chart, it can be seen that by far the $\log p(x)$ does not infer a certainty on the decision-making. Furthermore, there exists no reliable distinction between the distributions. For that reason, the classification losses are also included in the following figure (see Figure 4.10d) to investigate the certainty effect of $p(x, y)$. The mean of the true classified bags falls into $9,692.42$ with a standard deviation of 231.55 whereas the false classification rates take up a mean of $24,141.33$ with a high standard deviation of $5,951.70$. These results demonstrate that the hybrid architecture can assume uncertainty by the unlabeled data in carrying out the decision procedure. According to the Subfigure 4.10c, the densities of the test set are overlapped. Subfigure 4.10d provides a conspicuous separation between the true and false classified examples. The findings suggest that the ssMILVAE$_p$ loss higher than $9,730$ can be set as a threshold for illustrating uncertainty on the decision-making of the model. Again, the results are encouraging that the classification accuracy may be improved if such a threshold is implemented in the decision procedure.

In summary, these results show that the proposed method proves its performance under the few numbers of labeled data by demonstrating better achievements and consistent results about the competitor model architecture. The ssMILVAE$_p$ achieves higher accuracy in both datasets in the direction of the goal of this dissertation. Thereby, it is worth mentioning that the proposed method provides promising results to reduce the workload of the pathologists as data preparation requires devotion of their significant amount of time from their daily routines. The hybrid design signs a strong signal for a reliable certainty on the decision. Further, the recommended model performs adequately in assigning attention weights in providing ROIs. Together, these results provide insights into the success of the hybrid combination of a VAE and AD-MIL in multiple instance learning problems.

5 Discussion

As mentioned in the literature review, there played little attention so far to hybrid modeling in a MIL problem. Prior studies have noted the importance of the hybrid approach since the combination of generative and discriminative modeling utilizes the unlabeled data to center the knowledge of the input. An initial objective of the study was to develop a deep generative framework for a MIL problem. Although various studies sought to analyze the hybrid modeling and attention mechanism, very little was found in the literature on the question of the MIL problem. The majority of the researches were made for Out-of-Distribution Detection (OOD). Yet, there was not any significant questioning on the uncertainty for in-distribution samples in a MIL setup.

This study set out with the aim of assessing the importance of three crucial research goals. The first question in this study sought to implement a generative framework for multiple instance learning. The second aim of this research was to investigate whether the hybrid learning approach can provide state-of-the-art performance in the semi-supervised setting. Lastly, it was hypothesized that the proposed model evaluation is superior to the baseline models when the number of data is low. This research presents a hybrid model created by a variational autoencoder (VAE) and an Attention-based Deep MIL classifier (AD-MIL) in a MIL problem.

The current study found that an integration of the AD-MIL and the VAE cooperate adequately. The most impressive appealing finding was that the proposed method learns the input data better than a discriminative model can when the labeled data lacks. Meaning, a clear distinction is observed in the loss rate between true and false classified examples. Comparing the ELBO and the loss rates shows that the hybrid approach plays a critical role in the decision-making in terms of uncertainty against a false classification. So, the loss value can be set as a threshold to provide an uncertainty after exceeding this rate aside from the prediction of the category. What is surprising is that the probability that comes from the generative model, $p(x)$, contributes to the classifier $p(y|x)$ by exploiting the unlabeled data in a MIL problem. The last finding in this research is that the better performance of the proposed method than a single classifier in both datasets. The semi-supervised MIL VAE demonstrates significantly better achievements than the predictive

part alone when it has come to a few available labeled data. Although the MNIST-BAGS results are more successful than the COLON CANCER results, with a lack of justification, it is thought that this discrepancy could be attributed to the insufficient number of examples in the test set in the latter dataset. But it is worth noting that these accuracy comparisons do not involve the uncertainty rate. The loss rate as a threshold will open the gap between the performances. One unanticipated finding was that the better performance on the attention weights. In cases of the number of labeled data is not sufficient, the ssMILVAE$_p$ still provides qualified attention scores. Moreover, notable instances in a bag, especially in a positive example, are highlighted in this way. Consequently, the attention weights can be used to create supportive and more interpretable visualizations next to the results.

According to the findings, the ssMILVAE$_p$ produces ELBO rates close to a single VAE and better accuracy rates than an attention-based MIL classifier alone. These results show that two components of the hybrid model demonstrate a fruitful relationship. Furthermore, the loss criterion of the proposed model (ELBO + MIL) gives greater values for false classified examples. Greater values indicate the ability of the model in terms of sharing an uncertainty rate of the decision. This uncertainty may make a tremendous contribution to the various domain such as medical. Because an expert needs to trust the decision-making of the model whether a tumor or a similar disease occurs in a cell or not. For this purpose, the model has to be aware of whether the input fits the learned distribution. Moreover, one of the main focuses of this research was classification accuracy against the low number of labeled data. This resulted successively as the proposed approach shows an improved performance in both datasets when the number of labeled data is inadequate. On the other hand, the proposed method can still compete with a plain discriminative model as the number of labeled data increases. But the intention of this research does not concern such circumstances since it is unlikely to find a sufficient number of labeled data in a real-world example.

These results match those observed in earlier studies. The cooperation of a VAE and the AD-MIL in a hybrid structure satisfies the analysis conducted by Tulyakov et al. (2017). The findings belong to the uncertainty research corroborates the ideas of Nalisnick et al. (2019), who suggested that a hybrid architecture can make certain decisions between observed and unseen data. Moreover, there are similarities between the attitudes expressed

by the accuracy scores and attention weights in this study and those described by Ilse et al. (2018).

This study has provided here should be pursued further. The research concentrated on a generative framework in a binary MIL problem. But there is still abundant room for further progress in determining an integration of a multi-class MIL problem since in some tasks such as text classification, there may need to categorize a document based on multiple labels (Pham et al., 2015; Feng and Zhou, 2017). Furthermore, the data is assumed to be independently and identically distributed (i.i.d). But in a recent work, a non-i.i.d setup (Zhang, 2021) also gives promising results regarding the achieved AUC rates in comparison with the AD-MIL. From structural side, different strategies in architectural design may contribute in various ways. For instance, Burda et al. (2015) introduces importance weighting to the VAE that the method helps to learn richer latent space representations. Moreover, Higgins et al. (2016) augments a hyperparameter β that forces the model to learn a more efficient latent representation of the data which leads to more interpretable latent variables. A step further, a combination of a VAE and GAN (Zhao et al., 2020) can produce interesting results with regards to the comparison tables containing the Attention-based Deep MIL classifier. Lastly, only a single loss function is defined for each dataset in this study. In future investigations, it might be possible to use different loss functions and training strategies (Barron, 2019; Song et al., 2020) to observe the behavior of the model. Research questions that could be asked may include how to generate more qualified samples without any sacrification due to the α value in loss criterion or perform in OOD. Those issues and ideas are left for future research.

6 Conclusion

The present study was designed to determine the influence of integrating a generative framework for a MIL problem in a semi-supervised setting. This study contributes to the literature with a flexible hybrid modeling approach in a MIL setup that is fully parameterized by neural networks. Returning to the questions posed at the beginning of this study, it is now possible to state that the ssMILVAE architecture satisfies all experiments that are conducted to verify the results. This study has shown that the AD-MIL classifier performs in agreement with a VAE. Multiple experimental runs revealed that the ELBO rates of the ssMILVAE model follow the VAE (Kingma and Welling, 2013) closely. After the fulfilling results, four different approaches (see Section 3.3.4) were tested. Among the offered techniques, the results suggest that the ssMILVAE$_p$ approach has the best rates in both accuracy and loss criteria. One of the more significant findings to emerge from this study is that the hybrid integration $p(y|x)p(x)$ in a MIL problem provides greater loss that it makes possible to reincorporate the false classified examples. The histogram results of the expected loss rates and ELBO alone clearly show that the proposed method can provide an uncertainty assessment between true and false predictions. The current finding adds to a growing body of literature on contributing to such areas as medical domains. The ssMILVAE$_p$ can provide additional insight about the given example with a threshold agreement in the loss rate. Lastly and most importantly, the performance of the model against insufficient training data was the final research direction in this study. The experimental results show that the ssMILVAE$_p$ achieves notable accuracy rates in comparison to the predictive model performance alone. In addition to the research questions, this study also conducted substantial work by providing a type of interpretation of the model decision-making by displaying ROIs. This kind of additional output can be highly essential in a wide range of practical applications as it can be produced together with the final diagnosis to an expert.

This project was undertaken to overcome several problems in image processing in a MIL problem (see Section 1.3 for details) by developing a generative framework with an integration of a VAE and an AD-MIL. This work contributes to existing knowledge of the hybrid approach by providing an untested combination in a MIL setup. The results indicate that the ssMILVAE$_p$ approach can deal with the aforementioned issues.

Additionally, this study has demonstrated, for the first time, that a hybrid setup learns the data distribution adequately with the use of unlabeled data in a MIL problem. Another contribution comes from the high recall results that should increase the reliability of the model's decision-making. Because it is crucial to distinguish the positive example from the negative ones, especially in medical domains. In addition, to predict a patient as negative while s/he carries the disease is not tolerable, specifically if it includes fatal variants.

The investigation of developing a generative framework for multiple-instance learning has contributed to three primary research goals through this thesis. First, this study found out that the integration of a VAE and an Attention-based MIL classifier adequately. This architecture also lets exploiting both the generative side of the approach and the predictive part including the attention mechanism. Second, this dissertation has revealed that hybrid learning can provide state-of-the-art performance in semi-supervised settings according to the comparison with the baseline models. Additionally, the hybrid structure provides a better understanding of the data distribution such that the loss rates can be observed with a clear separation between true and false classified examples. Meaning that this architecture can yield higher loss values which can be used as a threshold to feedback as a (un)certain decision-making to the user. Third, the findings suggest that the proposed approach achieves higher classification rates on the semi-supervised scenario than the baselines. Taken together, this study confirms the development of a deep generative framework for the MIL.

Finally, a number of limitations need to be examined. Considering the limited time given throughout this dissertation, the out-of-distribution detection (OOD) could not have been fully experimented with during the investigation of the uncertainty of given examples. Furthermore, the data diversity was only limited by two, namely the MNIST-BAGS and the COLON CANCER datasets. There may need to vary the diversity of the datasets to entirely be confident about the superiority of the proposed method. Another substantial limitation includes the hardware qualification for such model architectures. The GPU model that was utilized during the experiments was NVIDIA GEFORCE RTX 2060 6GB. Due to the lack of GPU, the number of components in DMoL calculation (see Equation 3.22 was limited to 5. It is believed that a higher number of components might lead to better achievements (Salimans et al., 2017) in the COLON CANCER dataset. Similarly, the

number of channels in parts of VAE could not be set to bigger values that also might help in the accuracy.

This dissertation has thrown up interesting questions in need of further investigation. It would be interesting to assess the effects of multi-class MIL problem. Further research might explore other areas than the medical domain by changing the architectures of VAE and the AD-MIL.

References

Adel, T., Smith, B., Urner, R., Stashuk, D., and Lizotte, D. J. (2013). Generative multiple-instance learning models for quantitative electromyography. *arXiv preprint arXiv:1309.6811*.

Amores, J. (2013). Multiple instance classification: Review, taxonomy and comparative study. *Artificial intelligence*, 201:81–105.

Andrews, S., Tsochantaridis, I., and Hofmann, T. (2002). Support vector machines for multiple-instance learning. In *NIPS*, volume 2, pages 561–568. Citeseer.

Athiwaratkun, B. and Kang, K. (2015). Feature representation in convolutional neural networks. *arXiv preprint arXiv:1507.02313*.

Ballé, J., Laparra, V., and Simoncelli, E. P. (2016). End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*.

Barron, J. T. (2019). A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bellman, R. (1960). Dynamic programming, princeton, 1957. *BellmanDynamic Programming1957*.

Bellman, R. E. (2015). *Adaptive control processes: a guided tour*. Princeton university press.

Belongie, S., Malik, J., and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE transactions on pattern analysis and machine intelligence*, 24(4):509–522.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Bunescu, R. C. and Mooney, R. J. (2007). Multiple instance learning for sparse positive bags. In *Proceedings of the 24th international conference on Machine learning*, pages 105–112.

Burda, Y., Grosse, R., and Salakhutdinov, R. (2015). Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168.

Chapelle, O., Schölkopf, B., and Zien, A. (2006). A discussion of semi-supervised learning and transduction. In *Semi-supervised learning*, pages 473–478. MIT Press.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.

Chen, X., Li, Y., Yao, L., Adeli, E., and Zhang, Y. (2021). Generative adversarial u-net for domain-free medical image augmentation. *arXiv preprint arXiv:2101.04793*.

Ciompi, F., Geessink, O., Bejnordi, B. E., De Souza, G. S., Baidoshvili, A., Litjens, G., Van Ginneken, B., Nagtegaal, I., and Van Der Laak, J. (2017). The importance of stain normalization in colorectal tissue classification with convolutional networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 160–163. IEEE.

Cireşan, D. C., Giusti, A., Gambardella, L. M., and Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks. In *International conference on medical image computing and computer-assisted intervention*, pages 411–418. Springer.

Cirri, P. and Chiarugi, P. (2011). Cancer associated fibroblasts: the dark side of the coin. *American journal of cancer research*, 1(4):482.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65.

Cui, H., Heidorn, P. B., and Zhang, H. (2002). An approach to automatic classification of text for information retrieval. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 96–97.

Diaz, C. A. E. and Morales-Menendez, R. (2018). Parsimonious modeling for binary classification of quality in a high conformance manufacturing environment. *Transactions on Machine Learning and Data Mining*, 11(1):27–41.

Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71.

Ding, C., He, X., Zha, H., and Simon, H. D. (2002). Adaptive dimension reduction for clustering high dimensional data. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 147–154. IEEE.

Doran, G. and Ray, S. (2016). Multiple-instance learning from distributions. *The Journal of Machine Learning Research*, 17(1):4384–4433.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892–898.

Escobar, C. A. and Morales-Menendez, R. (2018). Machine learning techniques for quality control in high conformance manufacturing environment. *Advances in Mechanical Engineering*, 10(2):1687814018755519.

Feng, J. and Zhou, Z.-H. (2017). Deep miml network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Fix, E. and Hodges, J. (1951). An important contribution to nonparametric discriminant analysis and density estimation. *International Statistical Review*, 3(57):233–238.

Foulds, J. and Smyth, P. (2011). Multi-instance mixture models and semi-supervised learning. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 606–617. SIAM.

Gammerman, A., Vovk, V., and Vapnik, V. (2013). Learning by transduction. *arXiv preprint arXiv:1301.7375*.

Gärtner, T., Flach, P. A., Kowalczyk, A., and Smola, A. J. (2002). Multi-instance kernels. In *ICML*, volume 2, page 7.

Ghaffarzadegan, S. (2018). Deep multiple instance feature learning via variational autoencoder.

Goetz, M., Weber, C., Binczyk, F., Polanska, J., Tarnawski, R., Bobek-Billewicz, B., Koethe, U., Kleesiek, J., Stieltjes, B., and Maier-Hein, K. H. (2015). Dalsa: domain adaptation for supervised learning from sparsely annotated mr images. *IEEE transactions on medical imaging*, 35(1):184–196.

Gondara, L. (2016). Medical image denoising using convolutional denoising autoencoders. In *2016 IEEE 16th international conference on data mining workshops (ICDMW)*, pages 241–246. IEEE.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gordon, J. and Hernández-Lobato, J. M. (2017). Bayesian semisupervised learning with deep generative models. *arXiv preprint arXiv:1706.09751*.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

He, Y., Guo, J., and Zheng, X. (2018). From surveillance to digital twin: Challenges and recent advances of signal processing for industrial internet of things. *IEEE Signal Processing Magazine*, 35(5):120–129.

Herrera, F., Ventura, S., Bello, R., Cornelis, C., Zafra, A., Sánchez-Tarragó, D., and Vluymans, S. (2016). Multiple instance learning. In *Multiple instance learning*, pages 17–33. Springer.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2016). beta-vae: Learning basic visual concepts with a constrained variational framework.

Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. (1995). The" wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14(5).

Hou, L., Samaras, D., Kurc, T. M., Gao, Y., Davis, J. E., and Saltz, J. H. (2016). Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ilse, M., Tomczak, J., and Welling, M. (2018). Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR.

Ilse, M., Tomczak, J. M., and Welling, M. (2020). Deep multiple instance learning for digital histopathology. In *Handbook of Medical Image Computing and Computer Assisted Intervention*, pages 521–546. Elsevier.

Jaakkola, T. S., Haussler, D., et al. (1999). Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, pages 487–493.

Javadi, G., Samadi, S., Bayat, S., Pesteie, M., Jafari, M. H., Sojoudi, S., Kesch, C., Hurtado, A., Chang, S., Mousavi, P., et al. (2020). Multiple instance learning combined with label invariant synthetic data for guiding systematic prostate biopsy: a feasibility study. *International journal of computer assisted radiology and surgery*, 15(6):1023–1031.

Jiao, C. and Zare, A. (2017). Multiple instance hybrid estimator for learning target signatures. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 988–991. IEEE.

Jimenez-del Toro, O., Otálora, S., Andersson, M., Eurén, K., Hedlund, M., Rousson, M., Müller, H., and Atzori, M. (2017). Analysis of histopathology images: From traditional machine learning to deep learning. In *Biomedical Texture Analysis*, pages 281–314. Elsevier.

Keeler, J. D., Rumelhart, D. E., and Leow, W.-K. (1991). *Integrated segmentation and recognition of hand-printed numerals*. Microelectronics and Computer Technology Corporation.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. (2014). Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improving variational inference with inverse autoregressive flow. *arXiv preprint arXiv:1606.04934*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243.

Kraus, O. Z., Ba, J. L., and Frey, B. J. (2016). Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59.

Kuleshov, V. and Ermon, S. (2017). Deep hybrid models: Bridging discriminative and generative approaches. In *Proceedings of the Conference on Uncertainty in AI (UAI)*.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Lafarge, M. W., Pluim, J. P., Eppenhof, K. A., Moeskops, P., and Veta, M. (2017). Domain-adversarial neural networks to address the appearance variability of histopathology images. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 83–91. Springer.

Lakshminarayanan, B., Roy, D. M., and Teh, Y. W. (2016). Mondrian forests for large-scale regression when uncertainty matters. In *Artificial intelligence and statistics*, pages 1478–1487. PMLR.

Lasserre, J. A., Bishop, C. M., and Minka, T. P. (2006). Principled hybrids of generative and discriminative models. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 87–94. IEEE.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lee, C. H. and Yoon, H.-J. (2017). Medical big data: promise and challenges. *Kidney research and clinical practice*, 36(1):3.

Lee, S.-J. and Yun, C. C. (2010). Colorectal cancer cells–proliferation, survival and invasion by lysophosphatidic acid. *The international journal of biochemistry & cell biology*, 42(12):1907–1910.

Li, C. H., Gondra, I., and Liu, L. (2012). An efficient parallel neural network-based multi-instance learning algorithm. *The Journal of Supercomputing*, 62(2):724–740.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

Liu, K.-L., Li, W.-J., and Guo, M. (2012). Emoticon smoothed language models for twitter sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26.

Liu, W., Qin, C., Gao, K., Li, H., Qin, Z., Cao, Y., and Si, W. (2019). Research on medical data feature extraction and intelligent recognition technology based on convolutional neural network. *IEEE Access*, 7:150157–150167.

Lu, M. Y., Chen, R. J., Wang, J., Dillon, D., and Mahmood, F. (2019). Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. *arXiv preprint arXiv:1910.10825*.

Maaløe, L., Sønderby, C. K., Sønderby, S. K., and Winther, O. (2016). Auxiliary deep generative models. In *International conference on machine learning*, pages 1445–1453. PMLR.

Maron, O. and Lozano-Pérez, T. (1998). A framework for multiple-instance learning. *Advances in neural information processing systems*, pages 570–576.

Nahid, A.-A., Mehrabi, M. A., and Kong, Y. (2018). Histopathological breast cancer image classification by deep neural network techniques guided by local clustering. *BioMed research international*, 2018.

Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., and Lakshminarayanan, B. (2019). Hybrid models with deep and invertible features. In *International Conference on Machine Learning*, pages 4723–4732. PMLR.

Nallapati, R. (2004). Discriminative models for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 64–71.

Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848.

Oh, K.-S. and Jung, K. (2004). Gpu implementation of neural networks. *Pattern Recognition*, 37(6):1311–1314.

Oppenheim, A. V. (1999). *Discrete-time signal processing*. Pearson Education India.

Oquab, M., Bottou, L., Laptev, I., Sivic, J., et al. (2014). Weakly supervised object recognition with convolutional neural networks. In *Proc. of NIPS*, pages 1545–5963. Citeseer.

O'Shea, K. and Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Paisley, J., Blei, D., and Jordan, M. (2012). Variational bayesian inference with stochastic search. *arXiv preprint arXiv:1206.6430*.

Papandreou, G., Chen, L.-C., Murphy, K. P., and Yuille, A. L. (2015). Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Pappas, N. and Popescu-Belis, A. (2014). Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *Proceedings of the 2014 Conference on Empirical Methods In Natural Language Processing (EMNLP)*, pages 455–466.

Pappas, N. and Popescu-Belis, A. (2017). Explicit document modeling through weighted multiple-instance learning. *Journal of Artificial Intelligence Research*, 58:591–626.

Pathak, D., Shelhamer, E., Long, J., and Darrell, T. (2014). Fully convolutional multi-class multiple instance learning. *arXiv preprint arXiv:1412.7144*.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572.

Pham, A., Raich, R., Fern, X., and Arriaga, J. P. (2015). Multi-instance multi-label learning in the presence of novel class instances. In *International Conference on Machine Learning*, pages 2427–2435. PMLR.

Pinheiro, P. O. and Collobert, R. (2015). Weakly supervised semantic segmentation with convolutional networks. In *CVPR*, volume 2, page 6. Citeseer.

Raina, R., Madhavan, A., and Ng, A. Y. (2009). Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th annual international conference on machine learning*, pages 873–880.

Raina, R., Shen, Y., Ng, A. Y., and McCallum, A. (2003). Classification with hybrid generative/discriminative models. In *NIPS*, volume 3, pages 545–552. Citeseer.

Rakhlin, A., Shvets, A., Iglovikov, V., and Kalinin, A. A. (2018). Deep convolutional neural networks for breast cancer histology image analysis. In *International conference image analysis and recognition*, pages 737–744. Springer.

Ramon, J. and De Raedt, L. (2000). Multi instance neural networks. In *Proceedings of the ICML-2000 workshop on attribute-value and relational learning*, pages 53–60.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and variational inference in deep latent gaussian models. In *International Conference on Machine Learning*, volume 2, page 2. Citeseer.

Ruifrok, A. C., Johnston, D. A., et al. (2001). Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology*, 23(4):291–299.

Salakhutdinov, R. and Hinton, G. (2009). Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978.

Salimans, T., Karpathy, A., Chen, X., and Kingma, D. P. (2017). Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*.

Salimans, T. and Knowles, D. A. (2013). Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882.

Scudder, H. (1965). Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371.

Shi, X., Xing, F., Xie, Y., Zhang, Z., Cui, L., and Yang, L. (2020). Loss-based attention for deep multiple instance learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5742–5749.

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., and Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298.

Sirinukunwattana, K., Raza, S. E. A., Tsang, Y.-W., Snead, D. R., Cree, I. A., and Rajpoot, N. M. (2016). Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE transactions on medical imaging*, 35(5):1196–1206.

Song, Y., Yu, Z., Zhou, T., Teoh, J. Y.-C., Lei, B., Choi, K.-S., and Qin, J. (2020). Cnn in ct image segmentation: Beyond loss function for exploiting ground truth images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 325–328. IEEE.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312.

Terzić, J., Grivennikov, S., Karin, E., and Karin, M. (2010). Inflammation and colon cancer. *Gastroenterology*, 138(6):2101–2114.

Theis, L., Shi, W., Cunningham, A., and Huszár, F. (2017). Lossy image compression with compressive autoencoders. *arXiv preprint arXiv:1703.00395*.

Thompson, N. C., Greenewald, K., Lee, K., and Manso, G. F. (2020). The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*.

Tufail, A. B., Ma, Y.-K., and Zhang, Q.-N. (2020). Binary classification of alzheimer's disease using smri imaging modality and deep learning. *Journal of digital imaging*, 33(5):1073–1090.

Tulyakov, S., Fitzgibbon, A., and Nowozin, S. (2017). Hybrid vae: Improving deep generative models using partial observations. *arXiv preprint arXiv:1711.11566*.

Uspensky, J. V. (1937). Introduction to mathematical probability.

Van Der Maaten, L., Postma, E., Van den Herik, J., et al. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71):13.

Van Engelen, J. E. and Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440.

Vapnik, V. and Chervonenkis, A. (1974). Theory of pattern recognition.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., and Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12).

Vu, Q. D., Graham, S., Kurc, T., To, M. N. N., Shaban, M., Qaiser, T., Koohbanani, N. A., Khurram, S. A., Kalpathy-Cramer, J., Zhao, T., et al. (2019). Methods for segmentation and classification of digital microscopy tissue images. *Frontiers in bioengineering and biotechnology*, 7:53.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., and Tang, X. (2017). Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Wang, J. and Zucker, J.-D. (2000). Solving multiple-instance problem: A lazy learning approach.

Wang, X., Yan, Y., Tang, P., Bai, X., and Liu, W. (2018). Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24.

Wiens, J. J. (2003). Missing data, incomplete taxa, and phylogenetic accuracy. *Systematic biology*, 52(4):528–538.

Xu, J., Li, H., and Zhou, S. (2015a). An overview of deep generative models. *IETE Technical Review*, 32(2):131–139.

Xu, J., Luo, X., Wang, G., Gilmore, H., and Madabhushi, A. (2016). A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. *Neurocomputing*, 191:214–223.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015b). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Xu, K., Liu, S., Zhang, G., Sun, M., Zhao, P., Fan, Q., Gan, C., and Lin, X. (2019). Interpreting adversarial examples by activation promotion and suppression. *arXiv preprint arXiv:1904.02057*.

Xu, K., Liu, S., Zhao, P., Chen, P.-Y., Zhang, H., Fan, Q., Erdogmus, D., Wang, Y., and Lin, X. (2018). Structured adversarial attack: Towards general implementation and better interpretability. *arXiv preprint arXiv:1808.01664*.

Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M., Eric, I., and Chang, C. (2014). Deep learning of feature representation with multiple instance learning for medical image analysis. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1626–1630. IEEE.

Yetisgen-Yildiz, M. and Pratt, W. (2005). The effect of feature representation on medline document classification. In *AMIA annual symposium proceedings*, volume 2005, page 849. American Medical Informatics Association.

Zafra, A., Pechenizkiy, M., and Ventura, S. (2013). Hydr-mi: A hybrid algorithm to reduce dimensionality in multiple instance learning. *Information sciences*, 222:282–301.

Zhang, C., Platt, J., and Viola, P. (2005). Multiple instance boosting for object detection. *Advances in neural information processing systems*, 18:1417–1424.

Zhang, J., Marszałek, M., Lazebnik, S., and Schmid, C. (2007). Local features and kernels for classification of texture and object categories: A comprehensive study. *International journal of computer vision*, 73(2):213–238.

Zhang, Q. and Goldman, S. A. (2001). Em-dd: An improved multiple-instance learning technique. In *Advances in neural information processing systems*, pages 1073–1080.

Zhang, W. (2021). Non-iid multi-instance learning for predicting instance and bag labels using variational auto-encoder. *arXiv preprint arXiv:2105.01276*.

Zhang, Z., Chen, P., McGough, M., Xing, F., Wang, C., Bui, M., Xie, Y., Sapkota, M., Cui, L., Dhillon, J., et al. (2019). Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence*, 1(5):236–245.

Zhao, Y., Yang, F., Fang, Y., Liu, H., Zhou, N., Zhang, J., Sun, J., Yang, S., Menze, B., Fan, X., and 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)*.

Zhou, Y., Sun, X., Liu, D., Zha, Z., and Zeng, W. (2017). Adaptive pooling in multi-instance learning for web video annotation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 318–327.

Zhou, Z.-H. (2006). Learning with unlabeled data and its application to image retrieval. In *Pacific Rim International Conference on Artificial Intelligence*, pages 5–10. Springer.

Zhu, X. J. (2005). Semi-supervised learning literature survey.

Zupan, J. (1994). Introduction to artificial neural network (ann) methods: what they are and how to use them. *Acta Chimica Slovenica*, 41:327–327.

Appendix

A1 The MNIST-BAGS

Table A1.1: Grid search parameters for MNIST dataset.

Type	Parameters
Number of hidden layers	[3, 4, 5]
Number of hidden units	[32, 64, 128, 256, 512, 1024]
Latent dimensions	[16, 32, 64]
Alpha Multiplier	[1, 3, 5, 32, 50, 100]
Learning Rate	[1e-6, 1e-5, 3e-5, 1e-4, 3e-4, 1e-3]
Weight Decay	[1e-5, 1e-4, 1e-3, 0]

Table A1.2: VAE Architecture for MNIST dataset.

	Encoder		**Decoder**
Layer	Type	Layer	Type
1	Conv(3, 2, 1)-32 + LeakyReLU(0.2)	1	ConvTranspose(1, 1, 0)-512 + LeakyReLU(0.2)
2	Conv(3, 2, 1)-64 + LeakyReLU(0.2)	2	ConvTranspose(4, 1, 0)-256 + LeakyReLU(0.2)
3	Conv(3, 2, 1)-128 + LeakyReLU(0.2)	3	ConvTranspose(4, 2, 0)-128 + LeakyReLU(0.2)
4	Conv(3, 2, 1)-256 + LeakyReLU(0.2)	4	ConvTranspose(4, 2, 0)-64 + LeakyReLU(0.2)
5	Conv(3, 2, 1)-512 + LeakyReLU(0.2)	5	ConvTranspose(4, 1, 0)-32 + LeakyReLU(0.2)
6	Conv(1, 1, 0)-32	6	ConvTranspose(4, 1, 0)-1 + Sigmoid

Table A1.3: The only part that differs from the original model (Ilse et al., 2018) is the first convolutional block according to the size of the inputs.

Layer	Type
1	Conv(3, 1, 2)-20 + ReLU
2	MaxPool(2, 2, 1)
3	Conv(3, 1, 2)-50 + ReLU
4	MaxPool(2, 2, 1)

Table A1.4: The auxiliary network of the auxiliary model type for MNIST-BAGS.

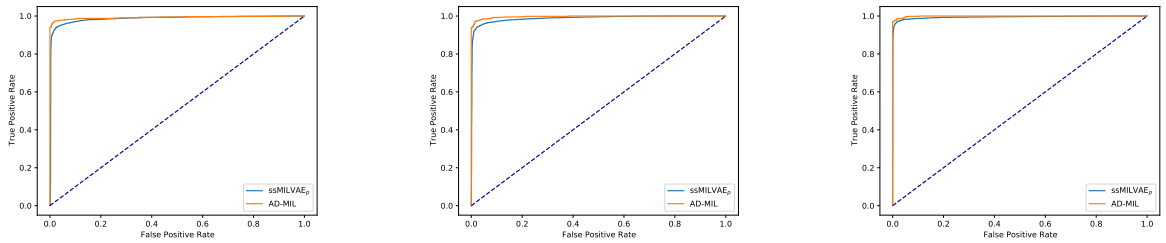
Layer	Type
1	Conv(7, 1, 0)-64 + LeakyReLU(0.2)
2	MaxPool(2, 2)
3	Conv(5, 1, 0)-256 + LeakyReLU(0.2)
4	MaxPool(2, 2)
5	Conv(3, 1, 0)-16

Table A1.5: MNIST-BAGS: The optimization procedure details for the ssMILVAE$_s$.

Model / Parameters	Optimizer	β_1, β_2	Learning rate	Weight decay
AD-MIL	Adam	0.9, 0.999	5e-4	1e-4
VAE	Adam	0.9, 0.999	1e-4	1e-3

Table A1.6: Results on MNIST-BAGS. Experiments were run 5 times and an average (± a standard error of the mean) is reported.

# LABELED			50		
METHOD	**ACCURACY**	**PRECISION**	**RECALL**	**F-SCORE**	**AUC**
SS MIL VAE	**0.815** ± 0.001	**0.858** ± 0.001	**0.851** ± 0.001	**0.854** ± 0.000	**0.891** ± 0.001
Attention MIL	0.735 ± 0.035	0.828 ± 0.014	0.738 ± 0.050	0.777 ± 0.035	0.768 ± 0.054
# LABELED			1000		
METHOD	**ACCURACY**	**PRECISION**	**RECALL**	**F-SCORE**	**AUC**
SS MIL VAE	0.952 ± 0.000	0.972 ± 0.000	0.952 ± 0.000	0.962 ± 0.000	0.988 ± 0.000
Attention MIL	**0.965** ± 0.006	**0.981** ± 0.004	**0.963** ± 0.005	**0.972** ± 0.004	**0.990** ± 0.001
# LABELED			3500		
METHOD	**ACCURACY**	**PRECISION**	**RECALL**	**F-SCORE**	**AUC**
SS MIL VAE	0.954 ± 0.001	0.967 ± 0.001	0.961 ± 0.000	0.964 ± 0.001	0.988 ± 0.000
Attention MIL	**0.965** ± 0.005	**0.978** ± 0.005	**0.967** ± 0.003	**0.973** ± 0.004	**0.993** ± 0.002
# LABELED			6000		
METHOD	**ACCURACY**	**PRECISION**	**RECALL**	**F-SCORE**	**AUC**
SS MIL VAE	0.970 ± 0.000	0.989 ± 0.001	0.963 ± 0.000	0.976 ± 0.000	0.994 ± 0.000
Attention MIL	**0.976** ± 0.003	**0.991** ± 0.001	**0.971** ± 0.003	**0.981** ± 0.002	**0.996** ± 0.001



(a) 1000 **(b)** 3500 **(c)** 6000

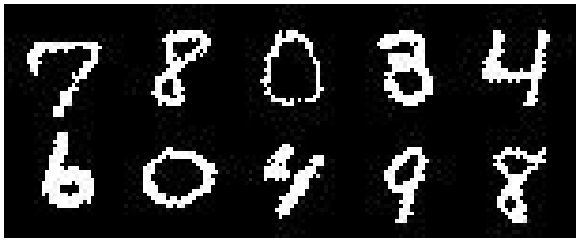
Figure A1.1: The ROC results of (a) 1000, (b) 3500 and (c) 6000 labeled bags from the test MNIST-BAGS with 10 instances per training bag under an experiment conducted for 5 times.

Table A1.7: The loss rates of the ssMILVAE$_p$ training and validation MNIST-BAGS during the training.

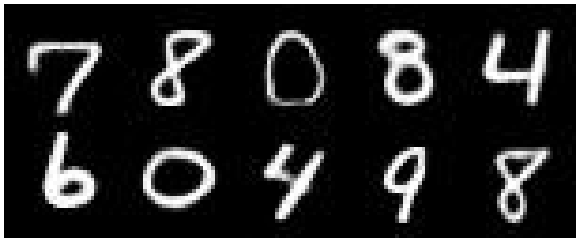
Loss / Epoch	**20**	**40**	**60**	**80**	**100**
Training	1525.59	663.90	679.78	738.30	686.52
Validation	1481.94	657.71	696.90	800.08	703.67

Table A1.8: The loss rates of the AD-MIL training and validation MNIST-BAGS during the training.

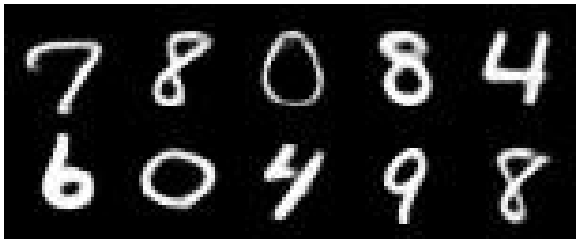
Model / Epoch		20	40	60	80	100
SS MIL VAE	Training	0.668	0.776	0.776	0.771	0.782
	Validation	0.666	0.780	0.780	0.772	0.784
Attention	Training	1.00	1.00	1.00	1.00	1.00
	Validation	0.698	0.698	0.706	0.718	0.716



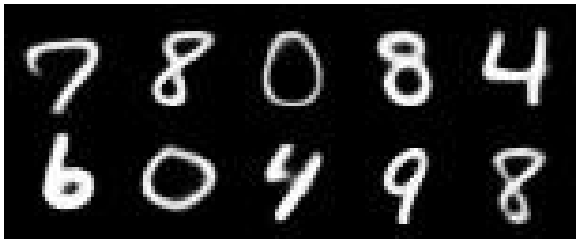
(a) Original Image



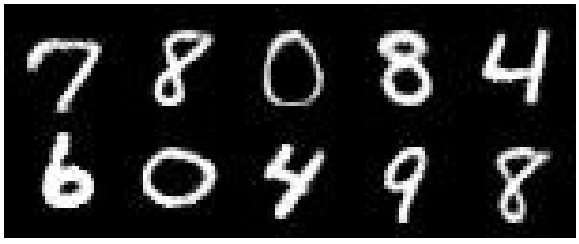
(b) VAE



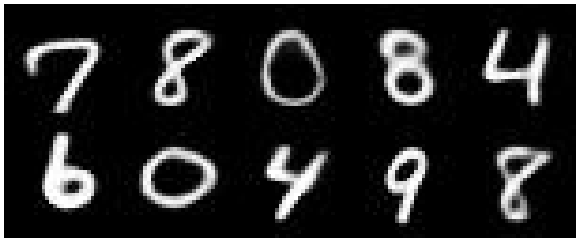
(c) ssMILVAE$_p$



(d) ssMILVAE$_d$



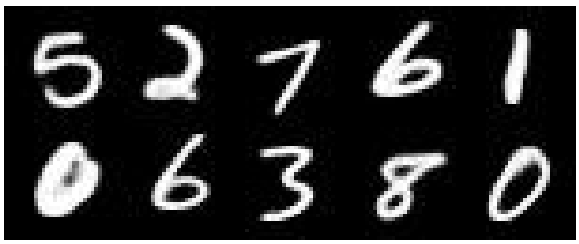
(e) ssMILVAE$_s$



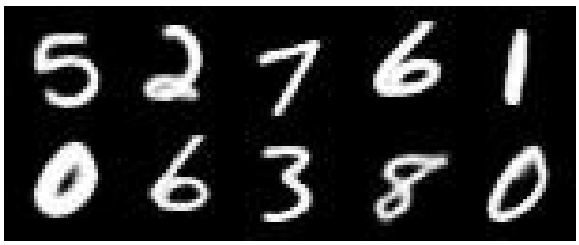
(f) ssMILVAE$_a$

Figure A1.2: Reconstruction examples of a **positive** bag from four different approaches tested for the MNIST dataset. To compare the results, an example by the VAE is displayed. The VAE model was trained with the full data.

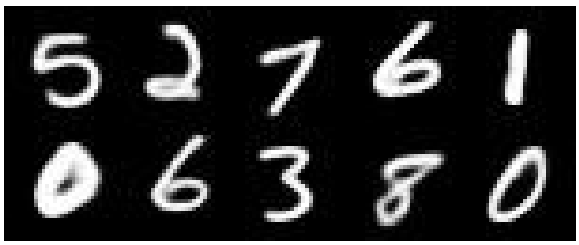
(a) Original Image



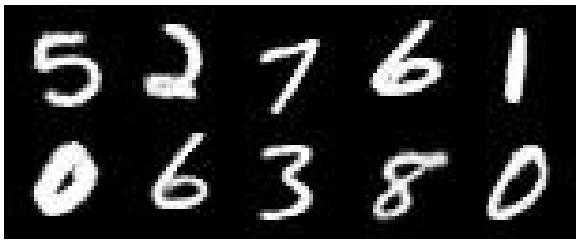
(b) VAE



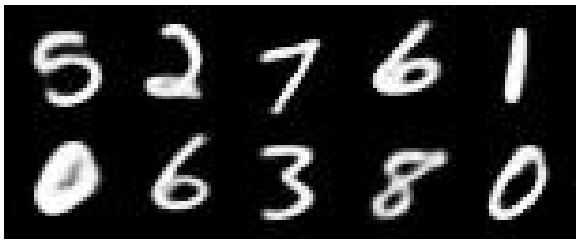
(c) ssMILVAE$_p$



(d) ssMILVAE$_d$

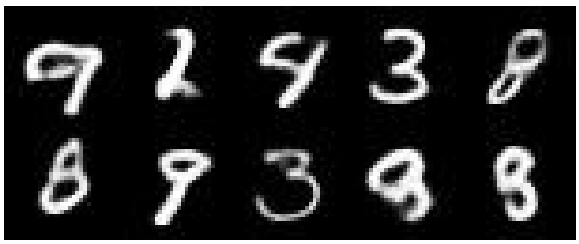


(e) ssMILVAE$_s$

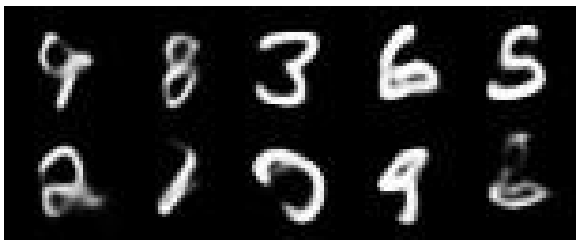


(f) ssMILVAE$_a$

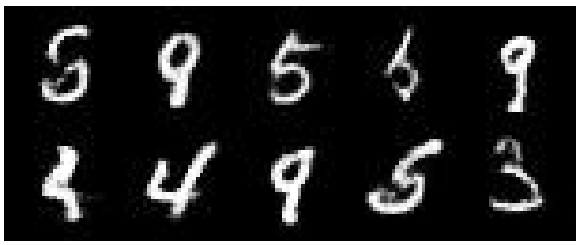
Figure A1.3: Reconstruction examples of a **negative** bag from four different approaches tested for the MNIST set. To compare the results, an example by the VAE is displayed. The VAE model was trained with the full data.



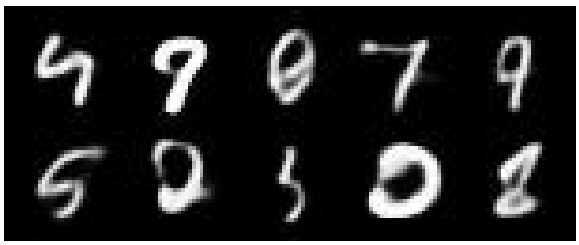
(a) ssMILVAE$_p$



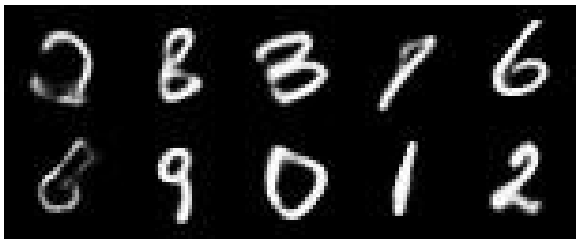
(b) ssMILVAE$_d$



(c) ssMILVAE$_s$

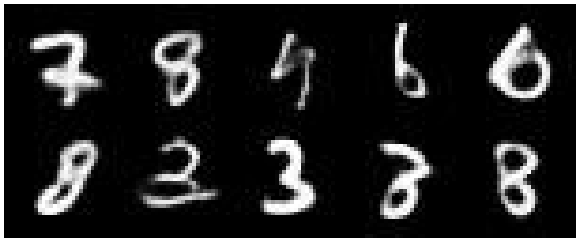


(d) ssMILVAE$_a$

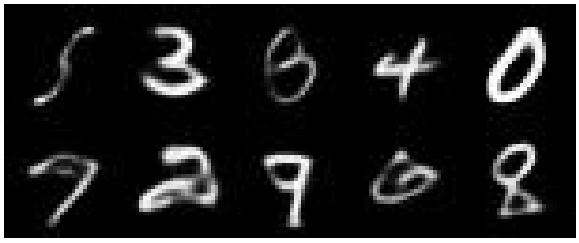


(e) VAE

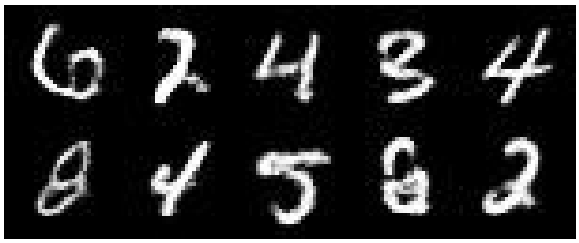
Figure A1.4: Sample examples that are classified as **positive** from four different approaches tested for the MNIST set. To compare the results, an example by the VAE is displayed without prediction. The VAE model was trained with the full data.



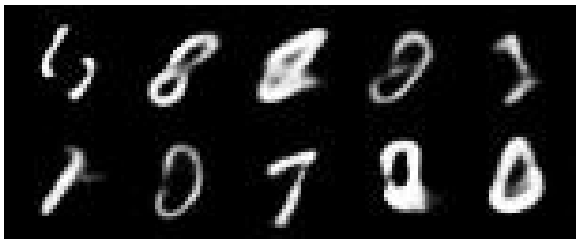
(a) ssMILVAE$_p$



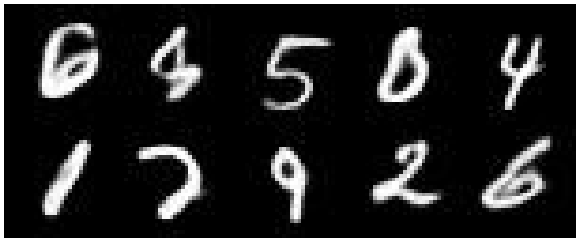
(b) ssMILVAE$_d$



(c) ssMILVAE$_s$

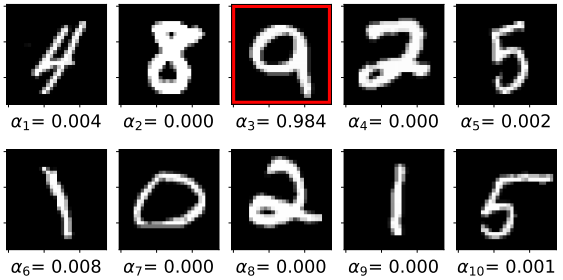


(d) ssMILVAE$_a$

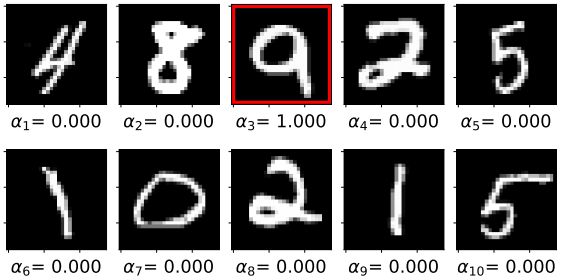


(e) VAE

Figure A1.5: Sample examples that are classified as **negative** from four different approaches tested for the MNIST set. To compare the results, an example by the VAE is displayed without prediction. The VAE model was trained with the full data.

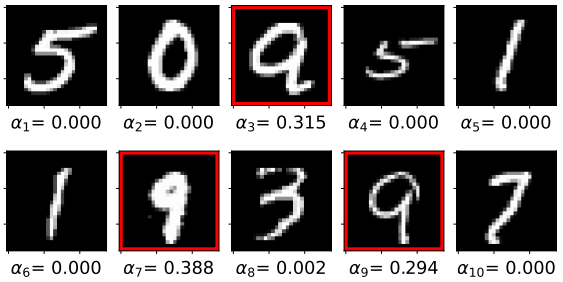


(a) ssMILVAE$_p$



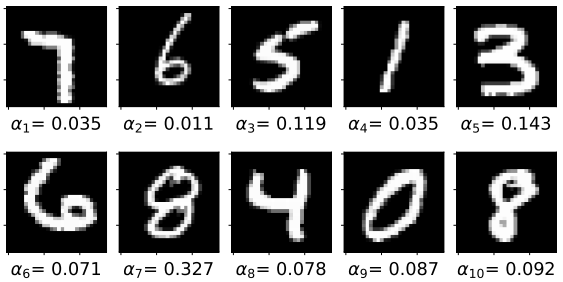
(b) AD-MIL

Figure A1.6: Attention weights of a true classified **positive** bag, that contains only a single instance of the target value, by (a) the ssMILVAE$_p$ and (b) the AD-MIL, respectively.



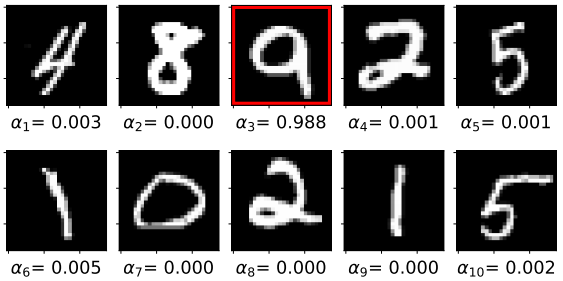
(a) Semi-Supervised MIL VAE **(b)** Attention MIL

Figure A1.7: Attention weights of a true classified **positive** bag, that contains more than a single instance of the target value, by (a) the ssMILVAE$_p$ and (b) the AD-MIL, respectively.



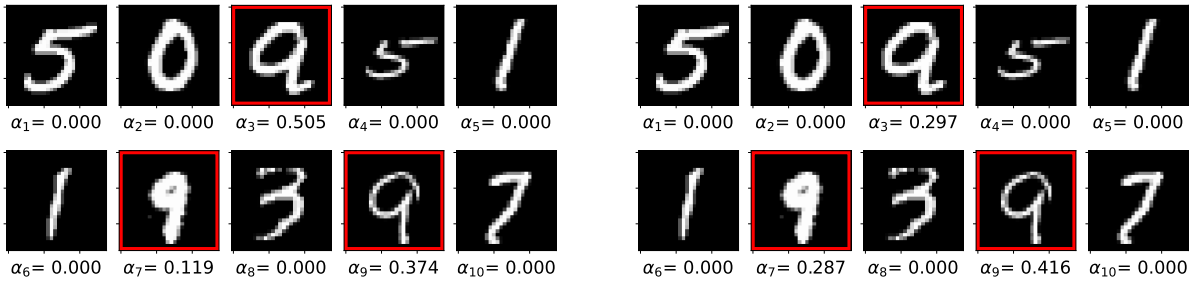
(a) ssMILVAE$_p$ **(b)** AD-MIL

Figure A1.8: Attention weights of a true classified **negative** bag, that does not contain any instance of the target value, by (a) the ssMILVAE$_p$ and (b) the AD-MIL, respectively.



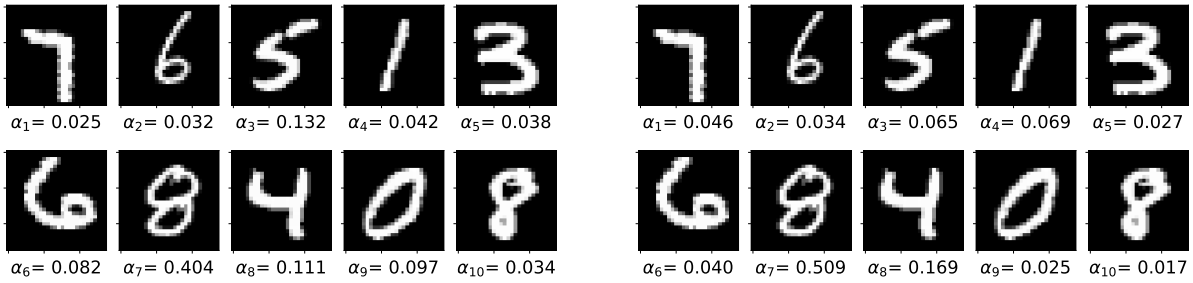
(a) ssMILVAE$_p$ **(b)** AD-MIL

Figure A1.9: Attention weights of a true classified **positive** bag, that contains only a single instance of the target value, by (a) the ssMILVAE$_p$ and (b) the AD-MIL, respectively.



(a) ssMILVAE$_p$ **(b)** AD-MIL

Figure A1.10: Attention weights of a true classified **positive** bag, that contains more than a single instance of the target value, by (a) the ssMILVAE$_p$ and (b) the AD-MIL, respectively.



(a) ssMILVAE$_p$ **(b)** AD-MIL

Figure A1.11: Attention weights of a true classified **negative** bag, that does not contain any instance of the target value, by (a) the ssMILVAE$_p$ and (b) the AD-MIL, respectively.

A2 The COLON CANCER

Table A2.1: Grid search parameters for COLON CANCER dataset.

Number of hidden layers	[3, 4, 5]
Number of hidden units	[32, 64, 128, 256, 512, 1024]
Latent dimensions	[16, 32, 64]
Alpha	[100, 1000, 10000]
Learning Rate	[1e-6, 1e-5, 3e-5, 1e-4, 3e-4, 1e-3]
Weight Decay	[1e-5, 1e-4, 1e-3, 0]

Table A2.2: VAE Structure for the COLON CANCER dataset.

Encoder		Decoder	
Layer	Type	Layer	Type
1	Conv(5, 1, 0)-64 + LeakyReLU(0.2)	1	ConvTranspose(3, 1, 0)-128 + LeakyReLU(0.2)
2	MaxPool(2, 2)	2	Upsample(2)
3	Conv(4, 1, 0)-128 + LeakyReLU(0.2)	3	ConvTranspose(4, 1, 0)-64 + LeakyReLU(0.2)
4	MaxPool(2, 2)	4	Upsample(2)
5	Conv(3, 1, 0)-64	5	ConvTranspose(7, 1, 0)-100

Table A2.3: Results on COLON CANCER. Experiments were run 5 times and an average (± a standard error of the mean) is reported.

# LABELED			22		
METHOD	**ACCURACY**	**PRECISION**	**RECALL**	**F-SCORE**	**AUC**
SS MIL VAE	**0.727** ± 0.027	0.679 ± 0.027	**0.915** ± 0.013	**0.779** ± 0.021	**0.787** ± 0.027
Attention MIL	0.663 ± 0.024	**0.790** ± 0.065	0.583 ± 0.115	0.611 ± 0.061	0.773 ± 0.030
# LABELED			92		
METHOD	**ACCURACY**	**PRECISION**	**RECALL**	**F-SCORE**	**AUC**
SS MIL VAE	**0.806** ± 0.019	0.859 ± 0.024	0.757 ± 0.013	0.804 ± 0.015	0.889 ± 0.023
Attention MIL	0.805 ± 0.032	**0.872** ± 0.059	**0.759** ± 0.024	**0.805** ± 0.028	**0.902** ± 0.033
# LABELED			162		
METHOD	**ACCURACY**	**PRECISION**	**RECALL**	**F-SCORE**	**AUC**
SS MIL VAE	0.775 ± 0.010	0.743 ± 0.018	**0.870** ± 0.016	0.801 ± 0.011	0.842 ± 0.019
Attention MIL	**0.904** ± 0.011	**0.953** ± 0.014	0.855 ± 0.017	**0.901** ± 0.011	**0.968** ± 0.009

(a) 92 **(b)** 162

Figure A2.1: The ROC results of (a) 92, and (b) 162 labeled bags from the COLON CANCER test bags under an experiment conducted for 5 times.

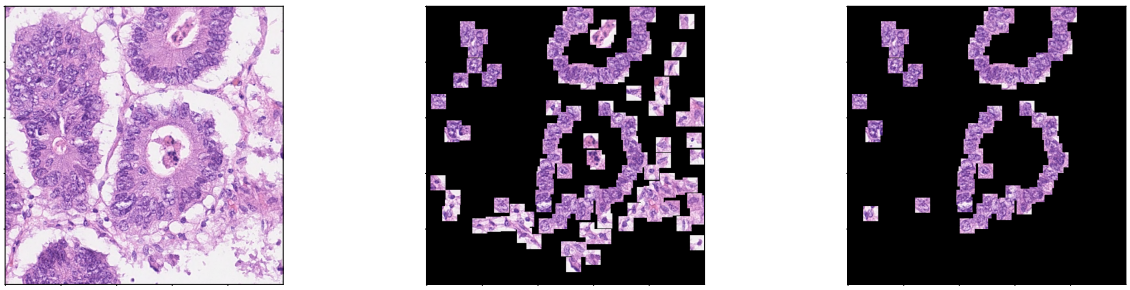
Table A2.4: The loss and accuracy rates from the ssMILVAE$_p$ training and validation COLON CANCER sets during the training.

Loss / Epoch	**20**	**40**	**60**	**80**	**100**
Training	10842.68	10936.53	**10697.46**	10894.22	11187.12
Validation	10849.36	10755.72	**9887.75**	9901.99	10143.65

Accuracy / Epoch					
Training	0.598	0.666	**0.765**	0.716	0.777
Validation	0.777	0.777	**0.944**	0.888	0.944

Table A2.5: The accuracy rates from the AD-MIL training and validation COLON CANCER sets during the training.

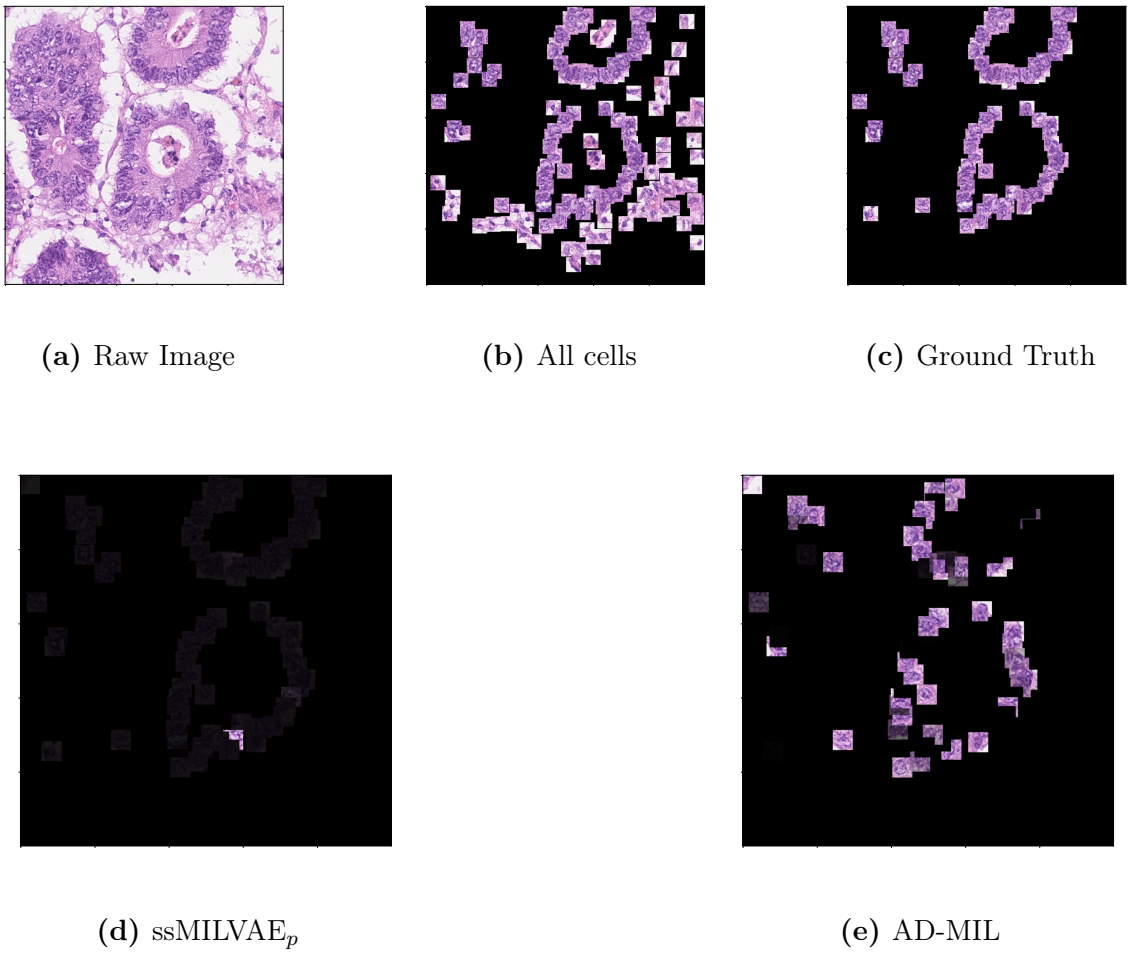
Accuracy / Epoch	**20**	**40**	**60**	**80**	**100**
Training	0.636	0.636	0.681	**1.00**	0.818
Validation	0.444	0.444	0.555	**0.66**	0.555



(a) Raw Image **(b)** All cells **(c)** Ground Truth

(d) ssMILVAE$_p$ **(e)** AD-MIL

Figure A2.2: (a) H&E stained histology image. (b) 27x27 patches centered around all marked nuclei. (c) Ground truth: Patches that belong to the class epithelial. (d) Heatmap: Every patch from (b) multiplied by its corresponding attention weight by 92 labeled bags of training set, the attention weights are rescaled by using $a'_k = (a_k - \min(\mathbf{a}))\,/(\max(\mathbf{a}) - \min(\mathbf{a}))$.



(a) Raw Image **(b)** All cells **(c)** Ground Truth

(d) ssMILVAE$_p$ **(e)** AD-MIL

Figure A2.3: (a) H&E stained histology image. (b) 27x27 patches centered around all marked nuclei. (c) Ground truth: Patches that belong to the class epithelial. (d) Heatmap: Every patch from (b) multiplied by its corresponding attention weight by 162 labeled bags of training set, the attention weights are rescaled by using $a'_k = (a_k - \min(\mathbf{a}))\,/(\max(\mathbf{a}) - \min(\mathbf{a}))$.