UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Deep Learning Strategies for Overcoming Diagnosis Challenges with Limited Annotations

October, 2023

ELECTRONIC ENGINEERING DEPARTMENT

Author: Rocío del Amor del Amor

Supervisors: Prof. Valery Naranjo Ornedo
Dr. Adrián Colomer Granero

The work of Rocío del Amor to carry out this research and to elaborate this dissertation has been supported by the Spanish Ministry of Universities under the FPU grant FPU20/05263.

Acknowledgements

La presente tesis representa el resultado de un arduo camino, marcado por el esfuerzo y la dedicación incansable. Afortunadamente, he estado rodeada de personas excepcionales y admirables. Aunque expresar mi gratitud por todo el apoyo y la confianza depositada en mí resulta desafiante, estas palabras están especialmente dedicadas a aquellos que han sido mi sostén en cada etapa de este trayecto.

Mis primeras líneas van dedicadas a Valery, mi directora de tesis, o mejor dicho, la persona que ha sido mi guía y apoyo incondicional en este camino. Como dijo Isaac Newton a Descartes y Hooke, "Si he visto más, es poniéndome sobre los hombros de Gigantes". Cada logro, cada avance y cada descubrimiento en este trabajo se deben a ti, a esa gigante que me tendió su mano hace varios años y me ha sostenido con sabiduría y paciencia a lo largo de este proceso. Tu influencia en mi vida académica y personal ha sido invaluable, y deseo profundamente que nuestros caminos nunca se separen.

Gracias, Adri, por ser además de mi codirector, mi guía y el mejor consejero que se puede tener. Tu dedicación y preocupación por nuestro bienestar son ejemplares, y valoro enormemente el apoyo incondicional que nos brindas. Trabajar contigo mano a mano es un verdadero placer, y me siento muy afortunada de seguir tus pasos en este apasionante campo. Por muchos más logros juntos y por seguir compartiendo risas y experiencias en nuestro camino.

También me gustaría agradecer a todos mis compañeros y amigos de CVBLAB por compartir su día a día conmigo. Sois un gran impuso para afrontar cada nuevo día en la ofi. Gracias, Fer, Sandra, Silvia, Elena, Cristian, Pablo, Alejandro, Ilán, Claudio, María, Natalia, Fernando, Óscar, Jesús, Fran, Juan y, especialmente, Leti. Gracias, Leti, por haberte convertido en esa amiga que todo el mundo desea tener, por no dejarme caer en mis peores momentos y por ser esa persona en la que siempre puedo confiar. Sigamos construyendo recuerdos y superando desafíos juntas.

También quiero darle las gracias a Alex Frangi, mi supervisor durante la estancia de investigación. Gracias por tu sabiduría y por abrirme las puertas de Leeds, una ciudad donde conocí a personas que guardaré en mi corazón para siempre.

Gracias a Mónica, Lucía y Carmen. Emprendimos este camino juntas allá por 2015 y hasta ahora hemos seguido apoyándonos como el primer día. Nuestras quedadas llenas de quejas, salseo y risas han sido un regalo durante este periodo. Espero que se sigan repitiendo muchos años más, os quiero.

Siempre lo he dicho, tengo a las mejores amigas del mundo. Gracias Fasiles Chasiles por compartir toda una vida juntas, por ser mi fuente de energía y por ser un apoyo incondicional. Juntas hemos aprendido a superar cualquier adversidad, más que amigas sois parte de mi familia, os quiero.

Gracias a mi familia, dado que soy lo que soy gracias a ella. Gracias a mis abuelos por el amor infinito que me dan. A mis tías, por siempre estar ahí regalándome su apoyo. A mis padres, por ser el pilar fundamental de mi vida. Gracias por cuidarme como lo hacéis, por vuestro sacrificio y esfuerzo para que tanto mi hermano como yo podamos alcanzar nuestras metas. Os estaré eternamente agradecida por todo lo que me dais. Y no puedo olvidar a mi increíble hermano, mi cómplice de aventuras y mi mejor amigo. Gracias por escucharme, entenderme en todo momento y por estar siempre a mi lado, compartiendo risas, aventuras y momentos. Gracias por ser mi familia y por hacer de mi vida una experiencia llena de amor y felicidad.

Por último, me gustaría dar las gracias a mi compañero de vida, mi Ricis. Desde el momento en que nuestros caminos se cruzaron, supe que había encontrado a alguien muy especial. Has sido y eres mi apoyo incondicional, mi confidente y mi mayor fuente de inspiración. A tu lado, he encontrado un amor puro y sincero que me ha ayudado a crecer y a ser la mejor versión de mí misma. Cada momento compartido a tu lado ha sido una aventura llena de risas, complicidad y amor. Agradezco a la vida por haberte puesto en mi camino y por permitirme disfrutar de tu amor incondicional. Te quiero más de lo que las palabras pueden expresar y espero continuar construyendo un futuro maravilloso juntos.

Abstract

In recent years, deep learning (DL) has become one of the main areas of artificial intelligence (AI), driven mainly by the advancement in processing power. DL-based algorithms have achieved amazing results in understanding and manipulating various types of data, including images, speech signals and text.

The digital revolution in the healthcare sector has enabled the generation of new databases, facilitating the implementation of DL models under the supervised learning paradigm. Incorporating these methods promises to improve and automate the detection and diagnosis of diseases, allowing the prediction of their evolution and facilitating the application of clinical interventions with higher efficacy.

One of the main limitations in the application of supervised DL algorithms is the need for large databases annotated by experts, which is a major barrier in the medical field. To overcome this problem, a new field of developing unsupervised or weakly supervised learning strategies using the available unannotated or weakly annotated data is opening up. These approaches make the best use of existing data and overcome the limitations of reliance on precise annotations.

To demonstrate that weakly supervised learning can offer optimal solutions, this thesis has focused on developing different paradigms that allow training models with weakly annotated or non-expert annotated databases. In this regard, two data modalities widely used in the literature to study various types of cancer and inflammatory diseases have been used: omics data and histological images. In the study on omics data, methods based on deep clustering have been developed to deal with the high dimensions inherent to this type of data, developing a predictive model without requiring annotations. In comparison, the results of the proposed method outperform other existing clustering methods.

Regarding histological imaging studies, the detection of different diseases has been addressed in this thesis, including skin cancer (spitzoid melanoma and spindle cell neoplasms) and ulcerative colitis. In this context, the multiple instance learning (MIL) paradigm has been employed as the baseline in all developed frameworks to deal with the large size of histological images. Furthermore, diverse learning methodologies have been implemented, tailored to the specific problems being addressed. For the detection of spitzoid melanoma, an inductive learning approach has been used, which requires a smaller volume of annotations. To address the diagnosis of ulcerative colitis, which involves the identification of neutrophils as biomarkers, a constraint learning approach has been utilized. With this method, the annotation cost has been significantly reduced while achieving substantial improvements in the obtained results. Finally, considering the limited number of experts in the field of spindle cell neoplasms, a novel annotation protocol for non-experts has been designed and validated. In this context, deep learning models that work with the uncertainty associated with such annotations have been developed.

In conclusion, this thesis has developed cutting-edge techniques to address the medical sector's challenge of precise data annotation. Using weakly annotated or non-expert annotated data, novel paradigms and methodologies based on deep learning have been proposed to tackle disease detection and diagnosis in omics data and histological images. These innovations can improve effectiveness and automation in early disease detection and monitoring.

Resumen

En los últimos años, el aprendizaje profundo (DL) se ha convertido en una de las principales áreas de la inteligencia artificial (IA), impulsado principalmente por el avance en la capacidad de procesamiento. Los algoritmos basados en DL han logrado resultados asombrosos en la comprensión y manipulación de diversos tipos de datos, incluyendo imágenes, señales de habla y texto.

La revolución digital del sector sanitario ha permitido la generación de nuevas bases de datos, lo que ha facilitado la implementación de modelos de DL bajo el paradigma de aprendizaje supervisado. La incorporación de estos métodos promete mejorar y automatizar la detección y el diagnóstico de enfermedades, permitiendo pronosticar su evolución y facilitar la aplicación de intervenciones clínicas de manera más efectiva.

Una de las principales limitaciones de la aplicación de algoritmos de DL supervisados es la necesidad de grandes bases de datos anotadas por expertos, lo que supone una barrera importante en el ámbito médico. Para superar este problema, se está abriendo un nuevo campo de desarrollo de estrategias de aprendizaje no supervisado o débilmente supervisado que utilizan los datos disponibles no anotados o débilmente anotados. Estos enfoques permiten aprovechar al máximo los datos existentes y superar las limitaciones de la dependencia de anotaciones precisas.

Para poner de manifiesto que el aprendizaje débilmente supervisado puede ofrecer soluciones óptimas, esta tesis se ha enfocado en el desarrollado de diferentes paradigmas que permiten entrenar modelos con bases de datos débilmente anotadas o anotadas por médicos no expertos. En este sentido, se han utilizado dos modalidades de datos ampliamente empleadas en la literatura para estudiar diversos tipos de cáncer y enfermedades inflamatorias: datos ómicos e imágenes histológicas. En el estudio sobre datos ómicos, se han desarrollado métodos basados en *deep clustering* que permiten lidiar con las altas dimensiones inherentes a este tipo de datos, desarrollando un modelo predictivo sin la necesidad de anotaciones. Al comparar el método propuesto con otros métodos de clustering presentes en la literatura, se ha observado una mejora en los resultados obtenidos.

En cuanto a los estudios con imagen histológica, en esta tesis se ha abordado la detección de diferentes enfermedades, incluyendo cáncer de piel (melanoma spitzoide y neoplasias de células fusocelulares) y colitis ulcerosa. En este contexto, se ha empleado el paradigma de *multiple instance learning* (MIL) como línea base en todos los marcos desarrollados para hacer frente al gran tamaño de las imágenes histológicas. Además, se han implementado diversas metodologías de aprendizaje, adaptadas a los problemas específicos que se abordan. Para la detección de melanoma spitzoide, se ha utilizado un enfoque de aprendizaje inductivo que requiere un menor volumen de anotaciones. Para abordar el diagnóstico de colitis ulcerosa, que implica la identificación de neutrófilos como biomarcadores, se ha utilizado un enfoque de aprendizaje restrictivo. Con este método, el coste de anotación se ha reducido significativamente al tiempo que se han conseguido mejoras sustanciales en los resultados obtenidos. Finalmente, considerando el limitado número de expertos en el campo de las neoplasias de células fusiformes, se ha diseñado y validado un novedoso protocolo de anotación para anotaciones no expertas. En este contexto, se han desarrollado modelos de aprendizaje profundo que trabajan con la incertidumbre asociada a dichas anotaciones.

En conclusión, esta tesis ha desarrollado técnicas de vanguardia para abordar el reto de la necesidad de anotaciones precisas que requiere el sector médico. A partir de datos débilmente anotados o anotados por no expertos, se han propuesto novedosos paradigmas y metodologías basados en deep learning para abordar la detección y diagnóstico de enfermedades utilizando datos ómicos e imágenes histológicas. Estas innovaciones pueden mejorar la eficacia y la automatización en la detección temprana y el seguimiento de enfermedades.

Resum

En els últims anys, l'aprenentatge profund (DL) s'ha convertit en una de les principals àrees de la intel·ligència artificial (IA), impulsat principalment per l'avanç en la capacitat de processament. Els algorismes basats en DL han aconseguit resultats sorprenents en la comprensió i manipulació de diversos tipus de dades, incloent-hi imatges, senyals de parla i text.

La revolució digital del sector sanitari ha permés la generació de noves bases de dades, la qual cosa ha facilitat la implementació de models de DL sota el paradigma d'aprenentatge supervisat. La incorporació d'aquests mètodes promet millorar i automatitzar la detecció i el diagnòstic de malalties, permetent pronosticar la seua evolució i facilitar l'aplicació d'intervencions clíniques de manera més efectiva.

Una de les principals limitacions de l'aplicació d'algorismes de DL supervisats és la necessitat de grans bases de dades anotades per experts, la qual cosa suposa una barrera important en l'àmbit mèdic. Per a superar aquest problema, s'està obrint un nou camp de desenvolupament d'estratègies d'aprenentatge no supervisat o feblement supervisat que utilitzen les dades disponibles no anotades o feblement anotats. Aquests enfocaments permeten aprofitar al màxim les dades existents i superar les limitacions de la dependència d'anotacions precises.

Per a posar de manifest que l'aprenentatge feblement supervisat pot oferir solucions òptimes, aquesta tesi s'ha enfocat en el desenvolupat de diferents paradigmes que permeten entrenar models amb bases de dades feblement anotades o anotades per metges no experts. En aquest sentit, s'han utilitzat dues modalitats de dades àmpliament emprades en la literatura per a estudiar diversos tipus de càncer i malalties inflamatòries: dades ómicos i imatges histològiques. En l'estudi sobre dades ómicos, s'han desenvolupat mètodes basats en *deep clustering* que permeten bregar amb les altes dimensions inherents a aquesta mena de dades, desenvolupant un model predictiu sense la necessitat d'anotacions. En comparar el mètode proposat amb altres mètodes de clustering presents en la literatura, s'ha observat una millora en els resultats obtinguts.

Quant als estudis amb imatge histològica, en aquesta tesi s'ha abordat la detecció de diferents malalties, incloent-hi càncer de pell (melanoma spitzoide i neoplàsies de cèl·lules fusocelulares) i colitis ulcerosa. En aquest context, s'ha emprat el paradigma de *multiple instance learning* (MIL) com a línia base en tots els marcs desenvolupats per a fer front a la gran grandària de les imatges histològiques. A més, s'han implementat diverses metodologies d'aprenentatge, adaptades als problemes específics que s'aborden. Per a la detecció de melanoma spitzoide, s'ha utilitzat un enfocament d'aprenentatge inductiu que requereix un menor volum d'anotacions. Per a abordar el diagnòstic de colitis ulcerosa, que implica la identificació de neutròfils com biomarcadores, s'ha utilitzat un enfocament d'aprenentatge restrictiu. Amb aquest mètode, el cost d'anotació s'ha reduït significativament al mateix temps que s'han aconseguit millores substancials en els resultats obtinguts. Finalment, considerant el limitat nombre d'experts en el camp de les neoplàsies de cèl·lules fusiformes, s'ha dissenyat i validat un nou protocol d'anotació per a anotacions no expertes. En aquest context, s'han desenvolupat models d'aprenentatge profund que treballen amb la incertesa associada a aquestes anotacions.

En conclusió, aquesta tesi ha desenvolupat tècniques d'avantguarda per a abordar el repte de la necessitat d'anotacions precises que requereix el sector mèdic. A partir de dades feblement anotades o anotats per no experts, s'han proposat nous paradigmes i metodologies basats en deep learning per a abordar la detecció i diagnòstic de malalties utilitzant dades *ómicos i imatges histològiques. Aquestes innovacions poden millorar l'eficàcia i l'automatització en la detecció precoç i el seguiment de malalties.

Contents

List of Figures xvii

List of Tables xxiii

List of Algorithms xxvii

1 Introduction 1

1.1 Motivation . . . 3

1.2 Objectives . . . 6

1.3 Main contributions . . . 7

1.3.1 Contribution to epigenetic data . . . 7

1.3.2 Contribution to histological data . . . 9

1.4 Framework . . . 13

1.5 Outline . . . 14

2 Deep Embedded Clustering for Breast Cancer Distinction 17

2.1 Introduction . . . 19

2.2 Material . . . 23

2.3 Methods . . . 24

2.3.1 Statistical analysis . . . 24

2.3.2 Dimensionality reduction . . . 24

2.3.3 Proposed method: Deep embedded refined clustering . . . 28

2.4 Experimental Results . . . 30

2.4.1 GSE32393 Series: Performance evaluation . . . 31

2.4.2 GSE50220 Series: Generalization ability of the DERC algorithm . . . 37

2.4.3 Comparison with the state of the art . . . 37

2.5 Discussion . . . 38

2.6 Conclusion . . . 40

3 Weakly Supervised framework for Spitzoid Cancer Diagnosis 41

3.1 Introduction . . . 43
3.2 Related work . . . 47
3.3 Material . . . 49
3.4 Methods . . . 51
3.4.1 Source model: ROI selection . . . 51
3.4.2 Target model: WSI prediction . . . 54
3.5 Ablation Experiments . . . 55
3.5.1 Database partitioning . . . 55
3.5.2 Source model selection . . . 56
3.5.3 Target model selection . . . 58
3.6 Prediction Results . . . 60
3.7 Discussion . . . 62
3.7.1 Source model: ROI selection . . . 62
3.7.2 Target model: WSI prediction . . . 65
3.8 Conclusion . . . 66

4 Constrained Multiple Instance Learning 67
4.1 Introduction . . . 69
4.2 Related work . . . 71
4.2.1 Multiple instance learning . . . 71
4.2.2 Constrained CNNs . . . 72
4.3 Methodology . . . 73
4.3.1 Problem formulation . . . 73
4.3.2 MIL backbone with location constraints . . . 74
4.3.3 MIL attention-embedding weights . . . 76
4.4 Experiments and Results . . . 76
4.4.1 Implementation . . . 76
4.4.2 Ablation experiments . . . 78
4.4.3 Comparison to the literature . . . 81
4.5 Conclusion . . . 82

5 Uncertainty-aware histology image classification 85
5.1 Introduction . . . 87
5.2 Related work . . . 88
5.2.1 Skin WSIs . . . 88
5.2.2 Uncertainty estimation . . . 89
5.3 Methods . . . 90
5.3.1 Labeling uncertainty . . . 92
5.3.2 Dual-branch uncertainty calibration . . . 93
5.4 Experimental setting . . . 95
5.4.1 Dataset . . . 95

5.4.2 ROI extraction 96
5.4.3 Implementation details 97
5.4.4 Evaluation metrics 97
5.5 Results 97
5.5.1 Comparison to the literature 97
5.5.2 Ablation studies 101
5.6 Conclusion 104

6 Final conclusions **107**
6.1 Global remarks 109
6.2 Specific remarks 109
6.3 Future work 112

Merits **113**

Bibliography **123**

List of Figures

2.1 DNA methylation process. Methylation at the 5’ position of the cytosine catalyzed by DNMT (DNA methyltransferases) in the presence of S-adenosyl methionine (SAM). 19

2.2 Architecture of the proposed conventional autoencoder used for the non-supervised dimensionality reduction. 25

2.3 Main differences between a conventional and a variational autoencoder. Instead of just learning a function representing the data (a compressed representation) like conventional autoencoders, variational autoencoders learn the parameters of a probability distribution representing the input data. 26

2.4 Architecture of a variational autoencoder. The proposed algorithm is optimized by minimizing two loss functions. One of them corresponding to the latent space regularisation and the other one corresponding to the input data reconstruction. 27

2.5 Architecture of the proposed method (DERC) to detect breast cancer using DNA methylation data. The proposed algorithm is trained to minimize both, clustering and reconstruction loss. . . 29

2.6 Final dimensionality reduction architectures. (a) Conventional autoencoder architecture composed of 4 stacks. (b) Variational autoencoder architecture composed of 4 stacks. Note that with the variational autoencoder algorithm, the latent space is obtained in two stages, two dense layers of 10 neurons representing the mean and the logarithm of the variance of the latent distribution and a sampling layer to obtain the points of the latent space. 33

2.7 Latent space of the dimensionality reduction algorithms. (a) Visualization of 10-dimensional features extracted by the latent space of the pre-trained variational autoencoder. (b) Visualization of 10-dimensional features extracted by the latent space of the pre-trained conventional autoencoder. 34

3.1 Representative patches extracted from WSIs presenting different spitzoid melanocytic lesions; (a)-(b): Benign spitzoid nevus containing well-defined melanocytic nests in an organized fashion; (c): Malignant lesion representative of the cellular disorder with ill-defined large tumor nest; (d): Malignant lesion with pagetoid spread, very common in this type of lesions; (e): Typical mitosis; (f): Atypical mitosis. 45
3.2 Annotation of a benign and a malignant spitzoid lesion. Patches (a) and (c) show characteristic patterns of the tumor region, benign and malignant respectively. Although patches (b) and (d) are inside the interest region annotated by the pathologists, these patches correspond to reactive stroma and do not contain tumor cells. 50
3.3 Overview of the proposed source model to conduct the tumor region detection. Blue and orange frames correspond to the base encoder network consisting of feature extraction and refinement. Note that VGG16 has been used as the feature extractor. After that, a projection head (green frame) maps the embedded representations in a lower-dimensional space to maximize the agreement in the classification stage (cyan frame). 51
3.4 Pipeline showing the embedded-level approach for spitzoid melanocytic lesion classification. The weights of the pre-trained feature extractor and feature refinement of the source model (σ^s and δ^s) are used to initialize this approach. After that, we use the output of the projection head and tile-level attention to weight the patches in the prediction of a whole biopsy. Using an aggregated bag-level feature vector we classify the entire biopsy. 52
3.5 Architecture of the Squeeze-and-Excitation blocks used to exploit the dependencies between feature channels. 53
3.6 Class activation maps (CAMs) for images correctly classified as tumor region or ROI (first row) and non-tumor regions (second row). First column: original images; Second column: CAMs obtained using the VGG16 model. Third column: CAMs using Squeeze and excitation network (SeaNet) with VGG16 as the backbone. SeaNet model focuses on the most distinctive features and, in this case, pays attention to the pagetoid spread to define a patch as tumorous and to the healthy stromal region for the non-tumoral region. 57

3.7 Original images (first column) and Class Activation Maps (CAMs) obtained with the VGG16 model (second column) and the Squeeze and excitation network (SeaNet) with VGG16 (third column). (b) and (e): Patches misclassified by the VGG16 model predicted as no ROI and ROI, respectively; (c) and (f): Patches well classified, ROI and No ROI respectively. 58
3.8 Whole slide image-level prediction for the source model (ROI estimation). (a) Manual annotation by experts; (b) System prediction completely in line with the annotation of (a); (c) Manual annotation by experts with expansion of areas with melanocytic nests characteristic of the lesion; (d) System prediction with certain areas annotated by the pathologists predicted as non-tumor regions. The expansion of the areas where there are no activations demonstrate that there are no melanocytic nests characteristic of the lesion; (e) Manual annotation by experts with expansion of area where melanocytic cells with melanosomes are found; (d) System prediction with expansion in the regions not annotated by the pathologist to demonstrate the presence of tumor cells. 61
3.9 Visualization of the attention weights of the bag aggregation function in heat maps. (a) Benign sample; (b) Malignant sample. 62

4.1 The larger image corresponds to a Whole-Slide Image (WSI) of a patient suffering from ulcerative colitis. The patches marked with colours denote different interest structures. Specifically: (a) lamina propia, (b) surface epithelium, (c) cryptal epithelium and (d) cryptal lumen. The black mark indicates the presence of a neutrophil. 71
4.2 Pipeline showing the embedded-level approach for ulcerative colitis detection. By incorporating the proposed location constraints, we force the backbone to extract more significant features from each patch belonging to a given bag. After that, we classify the entire biopsy using an aggregated bag-level feature vector weighted by the proposed attention-embedding weights. 74
4.3 Ablation studies on MIL formulation. Hyperparameters study for λ_{ac} are performed for bag-level accuracy on validation set. Confidence intervals are shown at 95%. 79

4.4 Class activation maps (CAMs) of some regions where neuthophils are found. First column: original images with pathologist annotation (green and red annotations); Second column: CAMs obtained using the normal MIL model. Third column: CAMs using the proposed location constraints. 80

4.5 Distribution of embedding weights across the instances that comprise a WSI. (a) Proposed attention embeddings. (b) Attention weights proposed in [124]. 82

5.1 **Method overview**. In this work, we address weakly supervised histology image classification on skin WSIs by quantifying the uncertainty of the individual annotators during labeling. Concretely, we train an embedding-based Multiple Instance Learning (MIL) model to predict up to six different categories using standard cross-entropy loss. We propose to quantify annotator-specific uncertainty by following a soft labels annotations protocol, such that $Y_k^{sl} = [0, 1]$, and $\sum_k Y_k^{sl} = 1$. In this fashion, our model captures information regarding inter-category dependencies and avoids over-fitting to uncertain, noisy annotations. Then, we propose a dual-branch min-max uncertainty calibration (DBEC) based on the annotated soft labels. Based on uncertainty calibration using Shannon entropy regularization (see Eq. 5.3), we propose to (i) maximize the entropy on high-confidence labeled samples, by entropy maximization (H^+), and (ii) to minimize the entropy on samples labeled with low-confidence (H^-). Thus, entropy minimization encourages the network to produce confident outputs on uncertain cases, based on the features of the sample, and thus diminishing noise propagation. A threshold τ is empirically fixed to differentiate low and high certain labels, and the dual-branch min-max uncertainty is combined with cross-entropy loss (see Eq. 5.5). Circles in bag-level predictions and references indicate soft-max scores. The more intense the color, the higher the score. 91

5.2 In-depth study of the soft labels annotated by in-training pathologists. (a) Quality of the labels, in terms of F1-score, in the training subset. Reference labels are the expert ground truth. (b) Percentage of samples with maximum confidence above the threshold $\tau = 0.7$. (c) Average confidence per each class, on positive samples. Dashed, red lines indicate average values. . . . 100

5.3 Normalized confusion matrices, averaged among non-expert annotators, obtained using (a) raw hard labels, (b) the model trained using hard labels, and (c) the model trained using the dual-branch entropy calibration proposed in Eq. 5.5. Reference labels are the expert ground truth. 101
5.4 Ablation study on the use of artificial model calibration of hard labels (HL) or annotated soft labels (SL). For the first approach, label smoothing (LS) and entropy maximization (H^+) are used. F1-score is presented for each method and non-expert annotator. 102
5.5 Ablation study on the effect of the confidence threshold τ on the proposed dual-branch entropy calibration (DBEC) based on annotated soft labels. 104

List of Tables

2.1 Reconstruction error of the proposed dimensionality reduction algorithms. Conventional autoencoder (AE) and variational autoencoder (VAE). 32
2.2 Comparison of the K-means clustering effect based on different feature extraction. 35
2.3 Architecture of the proposed deep embedded refined clustering model. 36
2.4 Comparison of the clustering effect of the proposed DERC based on different β values. 36
2.5 Results obtained over the external dataset. Note that in this case, the input data corresponds to the GSE50220 series after selecting the CpG sites extracted with the statistical analysis in the GSE32393 series. 37
2.6 Results reached by the state-of-the-art approaches in comparison with the proposed method when predicting the DNAm databases. Note that S1 refers to the primary set (GSE32393 series) and S2 to the external database (GSE50220 series). . . . 37

3.1 Main histological features of normal melanocytes and spitzoid lesions. 44
3.2 CLARIFYv1 database description. Amount of whole slide images with their respective biopsy label (first row), number of patches of each tumor region (second row) and number of non-interest region (third row). 50
3.3 Classification results reached during the validation stage with the proposed fine-tuned architectures. SeaNet: Squeeze-and-Excitation network. 56

3.4 Classification results reached during the validation stage using different projection head modules. SeaNet: Squeeze-and-Excitation network (with VGG16 as backbone), MLP: multi-layer perceptron, GMP: global max-pooling, GAP: global average-pooling. 59
3.5 Classification results reached during the validation stage using different aggregation functions. BGMP: batch global max-pooling; BGAP: batch global average-pooling; BGAS: batch global attention summary. 59
3.6 Classification results reached during the prediction stage. SM: source model; TM: target model. The proposed source model (SM) was composed of the SeaNet (with VGG16) + global max-pooling (GMP). The proposed target model (TM) used the batch global attention summary (BGAS) layer as an aggregation function. 60

4.1 PICaSSO Histologic Remission Index (PHRI) to predict histological remmision. 70
4.2 Database description. Amount of whole-slide images (first row), number of patches (second row) and percentage of slides with PHRI>0, ulcerative colitis (third row). 77
4.3 Comparison of the different attention embedding weights on the validation set. BGAP: batch global average pooling, BGMP: batch global max-pooling, LCMIL: neutrophil constrained weak supervision (proposed). Note that in all cases the location constraint proposed is integrated into the backbone. 81
4.4 Comparison of the different baseline frameworks in the test cohort. Note that for the test cohort only the global bag label are available. 82

5.1 Database distribution. DSV: database from Valencia; DSG: database from Granada. Lm:leiomyomas; lms: leiomyosarcomas; df:dermatofibromas; dfs: dermatofibrosarcomas; fxa: fibroxanthomas; spindle cell melanomas; cef: squamous cell carcinoma. . 96
5.2 Number of images used for training, validation and testing the models of each non-expert annotator (ten in total). Note that for the validation and test set the same samples labeled by all non-experts were used. 96

5.3 Quantitative comparison to prior literature. The metrics presented are the accuracy and micro-averaged F1-score (ACC/F1-score). The model trained with expert labels (second column) is used as the upper bound of the non-expert-based models. Colored values indicate the relative improvement of each method concerning the baseline using hard labels from non-expert in terms of the F1-score. Green indicates improvement and red a worsening lack. HL: hard labels; SL: soft labels; H^+: entropy maximization. 99

5.4 Ablation experiment on the components of the proposed calibration formulation. 103

List of Algorithms

1 Proposed methodology for the DERC approach. 31

Chapter 1

Introduction

This chapter introduces the motivation and the objectives pursued in this thesis, as well as the main contributions. It also includes the thesis framework and the thesis outline.

Contents

1.1 Motivation **3**
1.2 Objectives **6**
1.3 Main contributions **7**
1.3.1 Contribution to epigenetic data 7
1.3.2 Contribution to histological data 9
1.4 Framework **13**
1.5 Outline **14**

1.1 Motivation

Evolution of artificial intelligence

In the 1950s, Professor John McCarthy from Stanford University coined artificial intelligence as “the science and engineering of making intelligent machines” [1]. Artificial intelligence (AI) enables machines to imitate human cognitive functions such as problem-solving and learning. Machine learning (ML), a branch of AI, leverages data to develop computer systems that can learn and improve from experience without being explicitly programmed [2]. ML is a type of hand-driven learning since a feature engineering process should be carried out before the classification or regression stage. Therefore, a considerable understanding and expertise for representation, i.e., selection of features, is required [3]. Among others, outstanding approaches to feature extraction include texture analysis via local binary patterns (LBPs) [4], HoG [5] or SIFT [6]. The features extracted by these descriptors subsequently feed ML algorithms. There is a wide range of ML algorithms, and the choice of a particular approach is often informed by several factors, such as the type, size and complexity of the data and the task to be solved. Common ML methods include support vector machines (SVM) [7], ensemble-based methods such as random forests (RF) [8], K-means clustering [9], and others. ML has revolutionized various industries, including healthcare, by enabling the development of intelligent systems capable of extracting insights and making predictions from vast amounts of data.

Deep learning (DL) has rapidly gained popularity, emerging as a powerful subfield within ML [10]. One of the main advantages of DL is its ability to automatically extract features from raw data, eliminating the need for manual feature engineering. Unlike traditional ML approaches, DL models can process raw data directly, enabling them to uncover complex patterns and make decisions based on the learned representations. DL algorithms, inspired by the structure and function of the human brain, have demonstrated remarkable success in addressing complex classification tasks [11]. This success can be attributed, in part, to the availability of powerful computing machines equipped with graphical processing units (GPUs) and the availability of large-scale datasets [12]. DL techniques, such as artificial neural networks (ANN) with multiple layers, exhibit exceptional capabilities in analyzing diverse healthcare data types, encompassing medical images, genomic data and electronic health records [13]. By automatically learning intricate patterns and representations from these data, DL models have the potential to provide valuable insights, enhance diagnosis accuracy and support personalized treatment decisions [14].

The advances in DL have led to significant progress in the computer vision (CV) domain. Convolutional neural networks (CNNs)-based methods have significantly advanced the CV field, particularly in medical image analysis and classification [15]. Although these methods have been used since 1980, they are now considered the fundamental component of various vision tasks due to increased computation power and algorithmic development. CNNs can capture the underlying representation of the images using several convolution layers, followed by activation functions and pooling layers. The repeated application of filters (kernels) to the input image generates activation maps, often referred to as feature maps, highlighting salient regions of the image.

The problem of supervised frameworks

In DL-based systems, supervised learning is the predominant approach, where the model is trained using labeled data. In this process, each data sample is associated with a corresponding class or label of interest, which participates in the training process. The objective is to establish a relationship within the learning system that maps input data from the training set to its output labels. In medicine, input data can include medical acquisitions (e.g., images. clinical data, physiological signals, etc.), while the output label can be, for example, the disease diagnosis, the patient condition (e.g., the disease stage at a given follow-up time) or the outcome after therapy (e.g., recurrence, survival). Once this relationship is learned during the training phase, it can subsequently be applied to classify new input data with unknown labels into the predefined classes established during training.

Nevertheless, supervised DL models require large and curated datasets with high-quality annotations to perform appropriately. Achieving large and fully annotated datasets in real-world applications can be challenging. The annotation process is expensive and prone to subjectivity, making the task of obtaining such curated labeled datasets a cumbersome and complex task in practice. To overcome this problem, techniques such as transfer learning have been applied. In short, transfer learning is a common technique that involves leveraging knowledge gained from one task and applying it to a different but related task [16]. One of the main advantages of transfer learning is its ability to extract useful features from large and diverse datasets. Models trained on massive datasets, such as ImageNet [17], CIFAR [18] or COCO [19], which contain millions of labeled images, have learned to recognize general visual patterns and features. These pre-trained models serve as a valuable starting point for various computer vision tasks, including image classification, object detection, and image segmentation.

Towards a less-supervised perspective

While transfer learning can be effective in numerous scenarios, it may not be sufficient with domains that exhibit substantial dissimilarities. In this sense, there is a need to explore and develop novel deep-learning techniques that can perform well in label-poor scenarios where standard supervised learning approaches become infeasible or impractical [20]. This includes algorithms capable of incorporating any type of knowledge into learning that are easily accessible and also models able to learn on scarce, imbalanced datasets that lack precise labels. These challenges are particularly evident in the medical sector, where the complexities inherent in the medical data present a pressing demand for such algorithms. Unsupervised learning plays a crucial role in addressing these challenges by operating on unlabeled training and discovering hidden patterns or structures that can differentiate the data into subsets of similar samples by leveraging the inherent structure within the data. Unsupervised learning algorithms uncover meaningful representations and relationships without explicit labels. This approach has been widely applied in various domains, demonstrating its utility in data clustering, dimensionality reduction, anomaly detection and generative modeling. In the middle ground before the unsupervised and supervised scenarios, other learning strategies aim to build predictive models with weak supervision. Typically, there are three types of weak supervision. The first is incomplete supervision, i.e., only a (usually small) subset of training data is given with labels while the other data remain unlabeled. The second type is inexact supervision, i.e., only coarse-grained labels are given and the third type is inaccurate supervision, i.e., the given labels are not always ground truth [21]. These approaches offer promising avenues for effectively learning from partially labeled data, augmenting the capabilities of traditional supervised learning methods.

As mentioned earlier, the need for less supervised learning paradigms to improve diagnosis is particularly pronounced in the medical sector. The complexity of the data, combined with the high cost of annotation, presents significant hurdles for the development of accurate and efficient diagnostic systems. In this Ph.D. thesis, our motivation lies in developing novel learning methodologies that can be applied to the medical sector, extracting discoveries to advance medical research and enhance patient diagnosis. We overcome the limitations of scarce and imprecise labels in medical datasets by exploring various families of learning methods, including unsupervised and weakly supervised approaches. Concretely, our research aims to design new methodologies that can be applied to two cutting-edge research areas within diagnosis: genomic data analysis and digital pathology. Genomic data holds immense potential for personalized medicine, offering insights into individual genetic profiles and disease risk factors. In addition, we are entering the field of digital pathology, which digitizes microscopic tissue samples for analysis, representing the gold standard for a definitive diagnosis of numerous diseases. By applying DL techniques to these digital pathology images, we aim to develop advanced algorithms capable of accurate disease detection. In summary, this Ph.D. thesis aims to bridge the gap between the challenges faced in the medical sector regarding supervision information and the potential of AI-powered solutions to deal with the lack of accurate and amount of labels.

1.2 Objectives

The main objective of this Ph.D. thesis is dual, to make novel discoveries in the field of medicine and simultaneously develop cutting-edge techniques to address the challenge of precise data annotation, reducing the burden on medical professionals. In this sense, this thesis focuses on leveraging deep learning methodologies to create advanced diagnostic-aid systems capable of effectively analyzing weakly annotated medical databases. We aim to address several learning methods to cover different data domains and problems, e.g., unsupervised learning, self-training, weakly supervised learning, inductive learning, calibration and constraint learning. Each chapter of this thesis presents distinct methodologies tailored to specific data types and problem domains. In concrete, this thesis is focused on genomic and histological data. For genomic data, our primary focus is developing algorithms capable of effectively processing high-dimensional data to improve breast cancer diagnosis. Regarding histological data, our research focuses on diagnosing various types of cancer that have not been extensively studied in the literature, such as spitzoid melanoma and fusocellular skin cancer. Additionally, we aim to investigate a well-known inflammatory bowel disease, ulcerative colitis. To achieve the main purpose, all chapters share these specific objectives:

- Collecting, processing and conditioning the databases used and understanding the disease patterns that need to be detected.
- Designing and developing DL-based predictive algorithms tailored to analyze the medical data under study. The learning paradigm employed should be capable of handling the features of the disease and data modality. A comprehensive exploration of state-of-the-art methods is

necessary to propose cutting-edge solutions for automatic data-driven diagnostics.

- Conducting quantitative and qualitative evaluations of the proposed models and trying to improve the interpretability of the developed models. If possible, a comparative analysis with other state-of-the-art models should be performed to demonstrate the effectiveness of the proposed solutions.

1.3 Main contributions

As was mentioned earlier, this thesis incorporates outstanding contributions to the medical community and deep learning fields, introducing innovative discoveries in medicine and developing cutting-edge techniques to address the challenges of limited sample availability and the difficulties medical professionals face in accurately annotating data. These advancements represent a significant stride towards facilitating efficient and precise medical data analysis while reducing the workload on medical professionals.

1.3.1 Contribution to Genomic Data Analysis: Advanced Computing for Insightful Discoveries

Epigenetic mechanisms, one specific subset of genomic studies, play a crucial role in the normal development and maintenance of tissue-specific gene expression profiles. In mammalian cells, DNA methylation (DNAm) is based on the selective addition of a methyl group to the cytosine nucleotide under the action of DNA methyltransferases. Specifically, DNAm takes place in cytosines that precede guanines, known as CpG dinucleotides. There are CpG-rich areas (CpG islands) often located in the gene-promoting regions. The methylation of these CpG sites silences the promoter activity and correlates negatively with the gene expression. The methylation of the promoter regions in some vital genes and, therefore, their inactivation has been firmly established as one of the most common mechanisms for cancer [22, 23] and autoimmune/inflammatory disorders development [24]. Because the methylation patterns can be observed in the early stages, DNA methylation analysis becomes a powerful tool in the early diagnosis, treatment and prognosis of several disorders, such us cancer. Nowadays, DNA methylation is made at a molecular level generating large amount of data. The extremely high dimensions of the methylation data compared to the generally small number of available samples leads to the so-called curse of dimensionality problem, the main limitation in developing appropriate methods for DNAm data analysis. To overcome this challenge, it is crucial to devise effective approaches to transform the high-dimensional data space into a more meaningful and lower-dimensional representation while addressing the constraints imposed by the limited sample size. Therefore, one of the main contributions of this thesis is the development of new models that effectively address the challenges posed by DNAm data to enhance cancer diagnosis. Specifically, to tackle the issue of overfitting in scenarios with a limited number of samples, an innovative unsupervised algorithm is presented in this work. Notably, a novel deep clustering approach that combines self-training through autoencoders with clustering techniques is proposed (see Chapter 2).

Curse of dimensionality & Deep clustering

Several approaches based on high dimensional data reduction have been proposed to deal with the curse of the dimensionality problem. In this context, PCA and Fisher Criterion [25], non-negative matrix factorisation (NMF) [26], Random Boltzmann Machine (RBM) [27], deep autoencoder [28] and variational autoencoders (VAEs) [29] have been proposed. Several state-of-the-art methods propose different unsupervised and supervised classification algorithms for cancer identification after performing a dimensionality reduction of the DNA methylation data [26–28, 30]. However, no previous studies have been focused on the development of both tasks simultaneously. In this thesis, we propose a novel deep clustering algorithm for dimensionality reduction followed by a soft-assignment algorithm to perform an unsupervised classification, see Chapter 2. As the main novelty, the method is optimized through a weighted loss function in an end-to-end way. This loss function comprises two terms: (1) a reconstruction term in charge of optimizing the latent features provided by the autoencoder algorithm and (2) a clustering term used to improve the classification based on the latent features of the autoencoder.

Using this new methodology, this thesis brings noticeable contributions to the diagnosis of breast cancer [31] and spitzoid melanoma detection [32]. Regarding breast cancer, the proposed algorithm outperforms other state-of-the-art methods evaluated under the same conditions [26, 27]. Regarding the work proposed in [27], they obtain an error rate of 2.94 using a deep neural network (DNN) following a self-organizing feature map (SOM) compared to 0.73 obtained by the proposed method. Furthermore, their algorithm predicts four cancer examples as healthy, while our method only misclassifies one cancer sample. In the work proposed in [26], when they reduce the number of features to 540, the accuracy drops to 97.85 %, lower than achieved with the proposed method. Additionally, our latent space has only ten features, reducing the dimensionality to 99.9637 %. Therefore, this work could contribute to a faster and more effective diagnosis of breast cancer, improving cancer care and advancing the future of breast cancer research technologies. See Chapter 2 for more details.

In the case of spitzoid melanoma detection [32], the proposed method achieves approximately 0.90 accuracy, outperforming other supervised methods with which we made various comparisons. To conduct these comparisons, we employed an AE and VAE, along with a multi-layer perceptron classifier, resulting in accuracies of 0.85 and 0.65, respectively. Consequently, the unsupervised end-to-end training approach propsed not only minimizes data reconstruction but also enhances the classification process, enabling a more effective differentiation between benign and malignant spitzoid melanoma. This work represents a significant contribution as, in collaboration with pathologists from Hospital Clínico of Valencia, we have developed the first approach to diagnosing these challenging neoplasms.

1.3.2 Contribution to Whole Slide Image Analysis: Empowering Diagnostics with Cutting-Edge Techniques

Digital Pathology (DP) has experienced significant growth in recent years, becoming essential for the diagnosis and prognosis of tumors. DP involves capturing, storing, and analyzing high-resolution digital images of tissues, known as Whole Slide Images (WSIs) [33]. A WSI is a digital scanning technology that captures and converts glass slides for use in pathology, histology, and other healthcare fields into high-resolution digital images. A digital image can be viewed, analyzed, and shared electronically, which makes diagnosis, research, and collaboration between healthcare professionals more efficient and accurate [34]. The development of CAD systems based on WSI analysis presents important hardware limitations because of their large size. For this reason, the typical approach generally involves extracting small patches from larger WSIs, resulting in thousands of patches per image. Unfortunately, for these deep learning techniques to perform effectively, they require large and diverse annotated datasets [35]. To address this limitation, this thesis introduces notable contributions encompassing the development and validation of annotation tools, creating and publishing an extensive annotated database and designing innovative deep learning methods that do not require accurate annotations [36–42]. Specifically, an inductive training approach based on multiple instance learning is introduced to address complex neoplasms, such as spitzoid tumors. Additionally, a constraint-based convolutional neural network (CNN) is proposed to enhance the prediction of ulcerative colitis and reduce the workload associated with annotating small elements. Techniques for calibration in deep learning models are also explored to improve their accuracy when trained with annotations from non-expert pathologists. These contributions aim to advance the diagnosis and treatment in the medical field by addressing specific challenges and optimizing the performance of machine learning models.

Multiple Instance Learning

Multiple Instance Learning (MIL) is a weakly supervised paradigm that works to group the data on bags of instances, and only bag-level labels are known during training. In this setting, instances are independent of each other and the global label is positive if at least one of the instances belongs to the given category. Under an embedded-based method, a combination of transformed instances using a symmetric (permutation-invariant) function are aggregated to produce a bag-level representation that produces a global classification. In this thesis, we focus on embedded-based methods in the context of gigapixel histology WSI classification. In this application, each WSI is considered a bag, and extracted patches constitute instances. This method allows for incorporating information from all patches to perform the final diagnosis. However, all patches may not be equally important in reaching the final diagnosis. Therefore, in Chapter 3, we propose attention-based inductive learning to solve this problem. This learning paradigm is composed of two models: source and target models. In particular, the source model is trained to predict tumor regions by a patch-based CNN using inaccurate annotations. An improved CNN called SeaNeT, which effectively refines the extracted features is developed. After that, the backbone of the source model is retrained to classify nevus and malignant biopsies resulting in a target model characterized by a reduced number of labels as this model is retrained at the biopsy level. The target model is trained under an attention MIL paradigm. Each bag contains the tumor region pseudo-labeled by the source model, which facilitates the model training loop since the number of available biopsies is limited. Additionally, we employ a learnable weighting scheme to assign higher importance to instances that the model considers more important.

The application of this methodology in spitzoid detection, as presented in this thesis, is a significant contribution to the field. This lesion has been studied with deep learning techniques for the first time, obtaining an accuracy of 0.9231 and 0.80 for tumor region detection and melanoma diagnosis, respectively. In addition, after visualizing saliency maps of representative samples indicating the presence of tumor regions in WSIs, we can conclude that the developed algorithm could help the decision-making in cases of ambiguity for pathologists. At times, due to the large amount of different patterns in a lesion, pathologists can overlook some tumor areas and the developed method enhances the detection of these regions. Since there are no public databases for this type of neoplasm, the database and the relevant demographic data will be published in the article "A Spitzoid Tumor dataset with clinical metadata and Whole Slide Images for Deep Learning models" which has already been accepted in the journal Scientific Data.

The methodology based on MIL is also the backbone of the methods proposed in Chapter 4 and Chapter 5, incorporating the essential components and employing the most appropriate learning strategy to effectively solve the given problems.

Constraint formulations on weak supervision

Constrained classification aims to guide the training of an artificial neural network towards a solution that satisfies a given condition, which takes advantage of additional knowledge of the global labels. This learning paradigm has gained popularity in weakly-supervised scenarios (e.g., weakly supervised segmentation or MIL) since it allows to incorporate local information for improving the final task. Recent works have tackled weakly-supervised segmentation by imposing constraints on deep CNNs [43–46]. In particular, an L2 penalty term was proposed in [44] to impose equality constraints on the size of the target regions in the context of histopathology image segmentation. Additionally, the authors showed in [45] that imposing inequality constraints on size directly in gradient-based optimization, via also L2 penalty term, provided better accuracy and stability when few pixels of an image are labeled. In this thesis, Chapter 4 further along this line of research. In particular, we propose a constrained formulation that leverages prior knowledge of relative tissue location by imposing constraints on the activation maps of the feature extractor at the bag (WSI)-level. Including an L2 penalty in the loss function, we strict the expansion of positive instances during training. Additionally, under the MIL paradigm, we propose a new weighted average of instances where weights are obtained from the constrained activation maps.

The proposed constrained MIL approach is validated in the context of ulcerative colitis (UC), a chronic inflammatory bowel disease (IBD) affecting the colon and the rectum. The treatment of UC aims to extinguish bowel inflammation and prevent complications. Histological assessment plays a critical role in determining inflammatory activity. To identify UC activity effectively, neutrophil detection has proven to be an accurate indicator.

Therefore, the aim is to distinguish histological remission (favorable clinical outcomes) from activity based on the detection of neutrophils.

In most MIL-based works, the WSIs employed have broad features that determine a positive bag. However, in this case, small cells (neutrophils) with features very similar to others in the tissue differentiate whether a bag is positive, which poses a significant challenge. Therefore, the typical MIL approach is not useful as the extracted activations are degraded and do not allow satisfactory classification. This is demonstrated in Chapter 4 when comparing the proposed method with state-of-the-art MIL methods. In collaboration with pathologists from 7 centers in UK, Germany, Belgium, Italy, Canada, and the USA, the new index, PICaSSO Histologic Remission Index (PHRI), has been designed and validated using artificial intelligence for improved prognosis of ulcerative colitis [41]. The application of new AI algorithms developed during this thesis has had a significant impact on the detection of neutrophils [40, 42] and ulcerative colitis, achieving an accuracy of 87% when tested on a large cohort of 375 WSI [39]. These achievements have prompted our research group to become a leader in ulcerative colitis detection.

Uncertainty estimation in weakly annotations

Deep learning models require large, curated datasets with high-quality annotations to perform properly. In many cases, recruiting expert pathologists to annotate large databases is not feasible. Unfortunately, without sufficient labels, the data-hungry learning-based methods often struggle with overfitting, leading to inferior performance [47]. To alleviate this issue, collecting additional labeled data with varying label qualities, e.g., pathologists-in-training (henceforth, non-expert annotators) or using machine-generated labels, is a common practice. In this sense, one of the main contributions of this thesis (see Chapter 5) is the design and validation of a new annotation protocol for non-expert annotators. However, directly introducing data with low-quality (noisy labels) may confuse the network training, which easily leads to performance degradation [48, 49]. A main body of literature exploits multiple annotators in a crowdsourcing scenario to extract the underlying noise-free label distribution. Nevertheless, gathering multiple annotators in the medical context may be unrealistic. To tackle this issue, Chapter 5 of this thesis introduces a novel approach: an uncertainty-aware pipeline designed to handle the inherent uncertainty in the annotation process, which may not require multiple label sources. In concrete, we proposed a novel formulation based on dual-branch entropy calibration (DBEC) to calibrate overconfident outputs and uncertain soft labels. We use a set of 10 non-expert annotators to validate the proposed methodology. It is worth mentioning that, we use 10 annotators to validate the model and demonstrate that it is invariant to the level of experience of the non-expert pathologist, the hospital they belong to, etc., as well as to study its limitations. However, it should be noted that in all cases, the model has only been trained with the labels of one non-expert annotator.

The proposed uncertainty-aware method presented in Chapter 5 is applied to cutaneous spindle cell neoplasm detection, one of the most challenging skin neoplasms not studied in previous studies. We develop a new annotation tool for WSI labeling and validate it with ten non-expert pathologists, showing that this new tool can improve the accuracy of non-expert annotation. Using these annotations, we evaluate the proposed method finding average improvements of nearly $\sim 4.0\%$ in averaged F1-score using the baseline methods, which increases up to $\sim 6.6\%$ using the proposed dual-branch calibration. Additionally, in collaboration with pathologists of Hospital Clínico of Valencia, we have prepared a large WSI dataset containing both global biopsy-level labels and pixel-level annotations by expert and non-expert pathologists that will be published in the article "Annotation Protocol and Crowdsourcing Multiple Instance Learning Classification of Skin Histological Images: the CR-AI4SkIN Dataset" currently under review in the journal Artificial Intelligence in Medicine.

1.4 Framework

This Ph.D. thesis is part of four different research projects, as detailed below:

- *SAMUEL* – Artificial Intelligence System for Molecular and Morphological Skin Cancer Characterization. This is a regional project that aims to improve the diagnosis of melanocytic lesions by exploring new diagnosis aid systems based on artificial focus on digitized histological images, epigenetic information and clinical data. *SAMUEL* project was funded by *Agencia Valenciana de la Innovación (AVI)* (INNEST/2021/321). Chapter 2 contributes to this project by developing new diagnostic methods using epigenetic data.

- *CLARIFY* – Cloud artificial intelligence for pathology. This is an European project that proposes the creation of a research infrastructure based on AI and cloud-oriented data algorithms to facilitate the interpretation and diagnosis of triple-negative breast cancer (TNBC), high-risk non-muscle-invasive bladder cancer (HR-NMIBC) and spitzoid melanocytic lesions (SML) from histopathological images. *CLARIFY* project was funded by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska Curie grant agreement (No 860627). Chapter 3 contributes to this project giving rise to a weakly supervised framework for assessing spitzoid melanocytic lesions using histological images.

- *PICASSO* – Paddington International Virtual ChromoendoScopy ScOre. This multi-center consortium aims to improve the detection of ulcerative colitis using histological and endoscopic data. *PICASSO* was funded through a contract with Eli Lilly and Company. Chapter 4 contributes to this project in the design and development of a classification system capable of identifying ulcerative colitis from specific histological structures.

- *AI4SKIN* – Artificial intelligence for cutaneouS spindle cell neoplasm hIstopathological diagNosis. This is a national project whose objective is to develop an artificial intelligence-based diagnostic aid system for detecting spindle cell melanoma and non-melanoma skin cancer. *AI4SKIN* project was funded by the *Ministerio de Economía, Industria y Competitividad* (PID2019-105142RB-C21). Chapter 5 contributes to this project in designing and developing a classification system capable of identifying spindel cell cancer from non-expert labels.

1.5 Outline

This thesis is divided into 6 chapters. The current chapter introduces the motivation behind the research involved in this thesis, the proposed objectives and the main contributions. Subsequently, this chapter also details the framework and the thesis outline.

Chapter 2 corresponds to the paper: "A Deep Embedded Refined Clustering Approach for Breast Cancer Distinction based on DNA Methylation" [31]. It was published in the journal *Neural Computing and Applications* belonging to the editorial *Springer*. *Neural Computing and Applications* had an impact factor of 5.102 when the article was published in 2021, an impact score of 5.60 and an h-index of 94. The best rank was in the category *computer science, artificial intelligence* with a percentile of 69.31 (Q2).

Chapter 3 corresponds to the paper: "An Attention-based Weakly Supervised framework for Spitzoid Melanocytic Lesion Diagnosis in Whole Slide Images" [50]. It was published in the journal *Artificial Intelligence in Medicine* (AIIM) belonging to the editorial *ELSEVIER*. AIIM journal had an impact factor of 7.011 when the article was published in 2021, an impact score of 8.30 and an h-index of 93 in 2021. The top ranking was in the category *Engineering and biomedical* with a percentile of 79.08 (Q1).

Chapter 4 corresponds to the paper "Constrained Multiple Instance Learning for Ulcerative Colitis prediction using Histological Images" [42], published in the journal *Computer Methods and Programs in Biomedicine* (CMPB). The paper was published in 2022, but the publication data is from 2021. CMPB journal had an impact factor of 7.027, an impact score of 7.64 and an h-index of 115 in 2021. The best rank was in the category *computer science, theory & methods* with a percentile of 89.55 (Q1).

Chapter 5 corresponds to the paper "Labeling confidence for uncertainty-aware histology image classification" [38], published in the journal *Computerized Medical Imaging and Graphics* (CMIG). The paper was published in 2023, but the publication details date from 2021. CMIG journal had an impact factor of 7.422, an impact score of 8.40 and an h-index of 87 in 2021. The top ranking was in the category *radiology, nuclear medicine medical imaging* with a percentile of 90.07 (Q1).

Note that Chapters 2, 3, 4 and 5 are based on the same communication structure. First, they present an abstract followed by an introduction containing a review of the literature and the contribution of the proposed work. Next, the material section explains the datasets used to train and evaluate the developed ML algorithms, which are explained in the following methodology part. Then, the performance reached by the proposed methods is presented and discussed in the results and discussion sections, respectively. At the end, a brief conclusion recapitulates each work's main results and contributions and also establishes some future research directions.

In Chapter 6, the findings from each paper along with the global aim of this Ph.D. thesis are presented. Final remarks from a global perspective are exposed and future research lines are suggested. Then, in Merits, the journal publications and international conferences, as well as research awards derived from this thesis are included. Note that this thesis has generated additional scientific publications beyond the papers included in the manuscript. Finally, the Bibliography is displayed.

Chapter 2

A Deep Embedded Refined Clustering Approach for Breast Cancer Distinction based on DNA Methylation

The content of this chapter corresponds to the author version of the following published paper: del Amor, R., Colomer, A., Monteagudo, C. & Naranjo, V. A Deep Embedded Refined Clustering Approach for Breast Cancer Distinction based on DNA Methylation. Neural Computing and Applications, 34, 10243–10255 (2021).

Contents

2.1 Introduction . . . **19**
2.2 Material . . . **23**
2.3 Methods . . . **24**
2.3.1 Statistical analysis . . . 24
2.3.2 Dimensionality reduction . . . 24
2.3.3 Proposed method: Deep embedded refined clustering . . . 28
2.4 Experimental Results . . . **30**
2.4.1 GSE32393 Series: Performance evaluation . . . 31
2.4.2 GSE50220 Series: Generalization ability of the DERC algorithm . . . 37
2.4.3 Comparison with the state of the art . . . 37
2.5 Discussion . . . **38**
2.6 Conclusion . . . **40**

A Deep Embedded Refined Clustering Approach for Breast Cancer Distinction based on DNA Methylation

Rocío del Amor[1], Adrián Colomer[1], Carlos Monteagudo [2] and Valery Naranjo[1]

[1]Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano, Universitat Politècnica de València, 46022, Valencia, Spain
[2]Pathology Department. Hospital Clínico Universitario de Valencia, Universidad de Valencia, Valencia, Spain

Abstract

Epigenetic alterations have an important role in the development of several types of cancer. Epigenetic studies generate a large amount of data, which makes it essential to develop novel models capable of dealing with large-scale data. In this work, we propose a deep embedded refined clustering method for breast cancer differentiation based on DNA methylation. In concrete, the deep learning system presented here uses the levels of CpG island methylation between 0 and 1. The proposed approach is composed of two main stages. The first stage consists in the dimensionality reduction of the methylation data based on an autoencoder. The second stage is a clustering algorithm based on the soft-assignment of the latent space provided by the autoencoder. The whole method is optimized through a weighted loss function composed of two terms: reconstruction and classification terms. To the best of the authors' knowledge, no previous studies have focused on the dimensionality reduction algorithms linked to classification trained end-to-end for DNA methylation analysis. The proposed method achieves an unsupervised clustering accuracy of 0.9927 and an error rate (%) of 0.73 on 137 breast tissue samples. After a second test of the deep-learning-based method using a different methylation database, an accuracy of 0.9343 and an error rate (%) of 6.57 on 45 breast tissue samples is obtained. Based on these results, the proposed algorithm outperforms other state-of-the-art methods evaluated under the same conditions for breast cancer classification based on DNA methylation data.

2.1 Introduction

Epigenetic mechanisms are crucial for the normal development and maintenance of tissue-specific gene expression profiles in mammals. Recent advances in the field of cancer epigenetics have shown extensive reprogramming of every component of the epigenetic machinery, including DNA methylation, histone modifications, nucleosome positioning and non-coding RNAs [51]. In concrete, several studies demonstrate that DNA methylation (DNAm) plays a crucial role in the tumorigenesis process [22, 23].

In mammalian cells, DNA methylation is based on the selective addition of a methyl group to the cytosine nucleotide under the action of DNA methyltransferases [52], Figure 2.1. Specifically, DNAm takes place in cytosines that precede guanines, known as CpG dinucleotides [53].

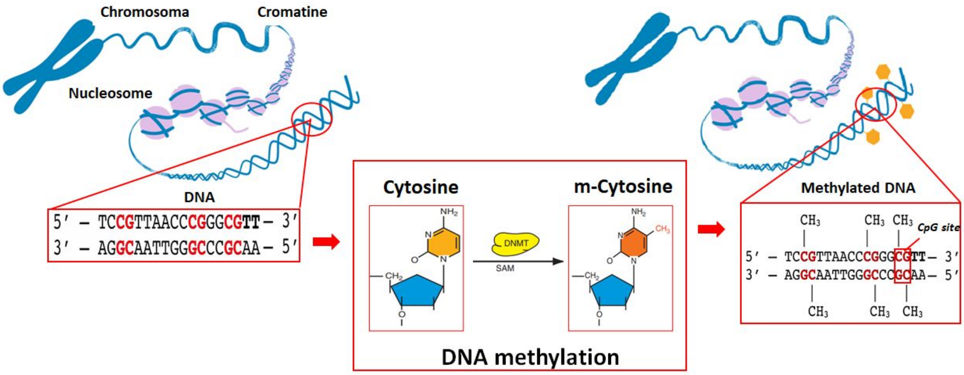

Chromosoma
Cromatine
Nucleosome
DNA
5′ – TCCGTTAACCCGGGCGTT– 3′
3′ – AGGCAATTGGGCCCGCAA– 5′
Cytosine
m-Cytosine
NH2
N
O
CH2
DNMT
SAM
CH3
DNA methylation
Methylated DNA
CH3
CpG site
5′ – TCCGTTAACCCGGGCGTT– 3′
3′ – AGGCAATTGGGCCCGCAA– 5′

Figure 2.1: DNA methylation process. Methylation at the 5' position of the cytosine catalyzed by DNMT (DNA methyltransferases) in the presence of S-adenosyl methionine (SAM).

CpG sites are not randomly distributed throughout the genome but there are CpG-rich areas known as CpG islands often located in the gene promoting regions. CpG islands are usually largely unmethylated in normal cells. The methylation of these CpG sites silences the promoter activity and correlates negatively with the gene expression. The methylation of the promoter regions in some vital genes, such as tumor suppressor genes, and therefore their inactivation, has been firmly established as one of the most common mechanisms for cancer development [30, 54]. Because the methylation patterns can be observed in the early stages of cancer [51], DNA methylation analysis becomes a powerful tool in the early diagnosis, treatment and prognosis of cancer.

The DNA methylation analysis has experienced a revolution during the last decade, especially due to the adaptation of microarray technology to the study of methylation and the emergence of Next-Generation Sequencing (NGS) [55, 56]. These technological advances combined with the development of techniques such as reduced representation bisulfite sequencing (RRBS), which is an efficient and high-throughput approach for analyzing the genome-wide methylation profiles, have allowed the DNA methylation analysis at the molecular level [57]. This is the reason why, current methylation studies generate a large amount of data. Additionally, since the study of DNA methylation is still a bit expensive, the number of available samples is relatively low. The extremely high dimensions of the methylation data compared to the generally small number of available samples leads to the so-called curse of dimensionality problem, the main limitation in the development of appropriate methods for DNAm data analysis.

The curse of dimensionality (COD) was introduced by Belman in 1957 [58] and refers to the difficulty of finding hidden structures when the number of variables is large. The high data dimensionality has different adverse effects: increased computational effort, large waste of space, overfitting and poor visualization [59]. In most cases, a dimensional increase has no significant benefit, since a lower data dimensionality might contain more relevant information. In machine learning problems, a small increase in data dimensionality requires a large increase in data volume to maintain a similar level of performance on tasks such as clustering, regression, etc. A well-established settlement to mitigate the curse of dimensionality is to transform the data from a higher-dimensional space to a more useful lower-dimensional space [59]. There are different types of dimensionality reduction algorithms. Some of them, such as Principal component analysis (PCA) or Manifold learning, use linear or nonlinear combinations of existing features to create new features. Others, such as Forward selection or Random forests, only keep the most important features in the dataset and removes redundant ones.

More recently, several state-of-the-art approaches based on high dimensional data clustering have been proposed to deal with the curse of dimensionality problem. Among these approaches, stand out the methods based on divisive hierarchical clustering [60, 61] and subspace clustering algorithms [62, 63]. Tasoulis et al. introduced a new approach to divisive hierarchical clustering identifying clusters in nonlinear manifolds. This approach uses isometric mapping (Isomap) to recursively embed subsets of data in one dimension and then performs a binary partition designed to avoid the splitting of clusters [60]. In [61], the authors proposed a new divisive hierarchical clustering method in which each partition in the hierarchy is induced by a hyperplane separator. Subspace clustering splits the data samples into groups such that each group contains only data samples lying in the same low-dimensional subspace of the given high-dimensional feature space. In [62], the author proposed a local extension of the well-known iterative subspace clustering algorithms in which the entire cluster is approximated with a single linear/affine subspace. Araújo et al. introduced a soft-subspace clustering algorithm, a Self-organizing Map (SOM) with a time-varying structure, to cluster data without any prior knowledge of the number of categories or of the neural network topology, both determined during the training process [63].

To mitigate the curse of dimensionality problem existing in the DNAm data, a dimensionality reduction is necessary before implementing any algorithm that identifies the presence of cancer using methylation profiles [25]. In this context, Yuvaraj et al. presented different algorithms for dimensionality reduction based on PCA and Fisher Criterion [25]. However, the DNA methylation datasets cannot be efficiently described by these dimensionality reduction methods due to its non-Gaussian character. Jazayer et al. used a non-negative matrix factorization (NMF) for the dimensionality reduction of breast methylation data, followed by ELM and SVM classifiers for cancer identification. However, with the NMF algorithm, it is not possible to directly transfer the input to a smaller dimensional space than the number of samples because this method transfers the data to an output space with a dimension equal to the minimum of {samples, DNAm dimension}. That is the reason why, in this study, the authors use a column-splitting method to overcome the curse of the dimensionality problem [26].

Recent advances in the field of artificial intelligence have allowed the development of deep learning algorithms that perform an embedding of CpG methylation states to extract biologically significant lower-dimensional features [27–29]. Zhongwei et al. presented a stack of Random Boltzmann Machine (RBM) layers with the aim of reducing the dimensionality of a breast DNAm set composed of cancer and non-cancer samples. The proposed model first selected the best 5,000 features based on variance from over 27,000 features and subsequently used four RBM layers to reduce the number of features to 30. After reducing the data dimensionality, they carried out a binary classification of the generated features using unsupervised methods [27]. Khwaja et al. proposed a deep autoencoder system for differentiation of several cancer types (breast cancer, lung carcinoma, lymphoblastic leukemia and Urological tumors) based on the DNA methylation states. After a statistical analysis, in which the features providing non-useful information for differentiation between cancer classes are eliminated, the authors used a Deep Belief Network for dimensionality reduction with a posterior supervised classification [28]. Titus et al. proposed an unsupervised deep learning framework with variational autoencoders (VAEs) to learn latent representations of DNA methylation from three independent breast tumor datasets. They demonstrate the feasibility of VAEs to track representative differential methylation patterns among clinical sub-types of breast tumors but they do not perform any classification with the extracted characteristics [29].

Several state-of-the-art methods propose different unsupervised and supervised classification algorithms for cancer identification after performing a dimensionality reduction of the DNA methylation data [26–28, 30]. However, to the best of the author's knowledge, no previous studies have been focused on the development of both tasks simultaneously. Novel deep learning algorithms have emerged optimizing the dimensionality reduction with unsupervised classification at the same time. These methods, called deep clustering algorithms, have outperformed the state-of-the-art results for different tasks as image classification [64–66], image segmentation [67], speech separation [68, 69] or RNA sequencing [70]. Therefore, our hypothesis is that since these algorithms perform well with high-dimensional data, they are likely to perform well for methylation data.

For all of the above, in this work, we proposed a deep-embedded refined clustering to distinguish cancer through DNA methylation data. In concrete, this work is developed using two public databases containing DNAm data from breast tissues with and without cancer. The proposed method is composed of an autoencoder to carry out the dimensionality reduction followed by a soft-assignment algorithm to perform an unsupervised classification. This algorithm is end-to-end trained to accomplish the data classification while optimizing the dimensionality reduction. As the main novelty, the method is optimized through a weighted loss function. This loss function is composed of two terms: (1) a reconstruction term in charge of optimizing the latent features provided by the autoencoder algorithm and (2) a clustering term used to improve the classification based on the latent features of the autoencoder. To the best of the authors' knowledge, no previous studies have addressed the distinction of cancer based on DNAm using an end-to-end trained dimensionality reduction and classification method. The proposed method is widely validated and compared to the use of autoencoder and variational autoencoder for dimensionality reduction with a subsequent unsupervised classification.

The rest of the paper is organized as follows: in Section 2.2, we introduce the databases used in this work, DNAm sets containing the methylation level (between 0 and 1) of different CpG regions related to cancer. In Section 3, we describe the methodology. In particular, Section 2.3.1 describes the statistical analysis performed on the DNAm data, Section 2.3.2 presents the dimensionality reduction algorithms used in this work, conventional and variational autoencoder, and Section 2.3.3 describes the details of the proposed deep clustering method. In Section 2.4, we describe the performed experiments in order to validate our method and in Section 2.5 we discuss the results obtained. Finally, Section 2.6 summarises the conclusions extracted with the carried out experiments.

2.2 Material

For this study, we used two methylation datasets obtained from Gene Expression Omnibus (GEO) website [71]. GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Specifically, we used the GSE32393 [72] and the GSE50220 [73] series to evaluate the proposed method. Note that the methodology proposed in this work was applied to the first series. The second series was used as an external database to perform an additional test and demonstrate the robustness of the proposed methodology. The GSE32393 series is composed of breast tissue samples from 114 breast cancers and 23 non-neoplastic breast tissues. The breast cancer tissue samples come from women from the United Kingdom (mean age 59.4) who were diagnosed with breast cancer. Among the cancers, 33 were at stage 1 and 81 at stage 2/3/4. All 23 non-neoplastic samples are from healthy women (mean age 47.6). The GSE50220 series is composed of breast tissue samples from 39 breast cancer and 9 normal control acquired at the Norwegian Radium Hospital (Norway). Among the breast cancer, 20 were non-irradiated breast cancer and the rest were irradiated tumors. In all cases, to obtain the methylation data, the Illumina Infinium 27k Human DNA methylation Beadchip v1.2 was used at approximately 27,000 CpGs from women with and without breast cancer. For each sample, 27,578 DNA methylation profiles were obtained. The methylation status of each CpG site varies from 0 to 1. Under ideal conditions, a value of 0 means the CpG site is completely unmethylated and the value of 1 indicates the site is fully methylated.

2.3 Methods

2.3.1 Statistical analysis

To conduct the prescreening procedure and obtain the methylation sites with the most differential methylation expression, a previous statistical analysis of the CpG methylation data was carried out. First, a hypothesis contrast to analyze the level of independence between pairs of variables was performed. For this purpose, the correlation coefficient ρ and the *p-value* of the correlation matrix were calculated to remove those variables that meet both *p-value* $\leq \alpha$ and $|\rho| \leq 0.90$, being α the level of significance with a value of 0.05 for this application. After that, we performed different contrasting hypotheses to analyze the discriminatory ability of each variable regarding the class. Depending on if the variables fit a normal distribution or not, the hypothesis test performed was the t-student or the Wilcoxon Rank-Sum, respectively. After the statistical analysis, we reduced the 27,578 DNA methylation features of the GSE32393 series to 10,153. These features were the input for the following stage.

2.3.2 Dimensionality reduction

In order to explore the well-known non-supervised algorithms to reduce the data dimensionality based on deep learning techniques, the conventional and the variational autoencoder were tested. In this section, we detail the characteristics of both algorithms as well as their main differences.

- *Conventional autoencoder*

Autoencoder (AE) is one of the most significant algorithms in unsupervised data representation. The objective of this method is to train a mapping function to ensure the minimum reconstruction error between input and output [74]. As it can be observed in Figure 2.2, the conventional autoencoder architecture is composed mainly of two stages: the encoder and the decoder stages. The encoder step is in charge of transforming the input data $\mathbf{X}$ into a latent representation $\mathbf{Z}$ through a non-linear mapping function, $\mathbf{Z} = f_\phi(\mathbf{X})$, where ϕ are the learnable parameters of the encoder architecture. The dimensionality of the latent space $\mathbf{Z}$ is much smaller than the corresponding input data to avoid the curse of dimensionality [65]. Since the latent space is a non-linear combination of the input data with smaller dimensionality, it can represent the most salient features of the data. The decoder stage produces the reconstruction of the data based on the features embedded in the latent space,

$\mathbf{R} = g_\theta(\mathbf{Z})$. The reconstructed representation $\mathbf{R}$ is required to be as similar to $\mathbf{X}$ as possible. Therefore, given a set of data samples $\mathbf{X} = \{x_i, ..., x_n\}$, being n the number of available samples, the autoencoder model is optimized with the following formula:

$$\min_{\theta,\phi} L_{rec} = \min \frac{1}{n} \sum_{i=1}^{n} ||x_i - g_\theta(f_\phi(x_i))||^2 \tag{2.1}$$

where θ and ϕ denote the parameters of encoder and decoder, respectively.

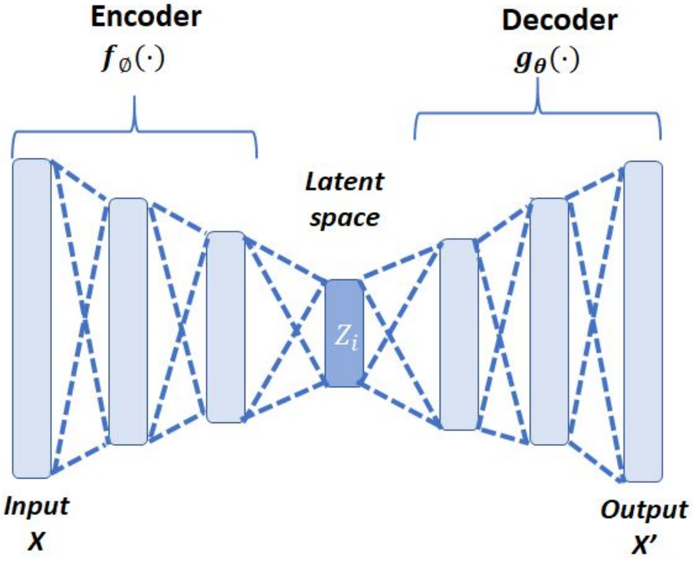


Figure 2.2: Architecture of the proposed conventional autoencoder used for the non-supervised dimensionality reduction.

The autoencoder architecture can vary between a simple multilayer perceptron (MLP), a long short-term memory (LSTM) network or a convolutional neural network (CNN), depending on the use case. In case the input data is 1-D and unrelated in time, both the encoder and decoder are usually constructed by a multilayer perceptron.

- *Variational autoencoder*

Variational autoencoder (VAE) is an unsupervised approach composed also of an encoder-decoder architecture like the conventional autoencoder aforementioned [29]. However, the main difference between a conventional and a variational autoencoder lies in the fact that the VAE introduces a regularisation into the latent space to improve its properties. With a VAE, the input data is coded as a normal multivariate distribution $p(z|x)$ around a point in the latent space. In this way, the encoder part is optimized to obtain the mean and covariance matrix of a normal multivariate distribution, see Figure 2.3.

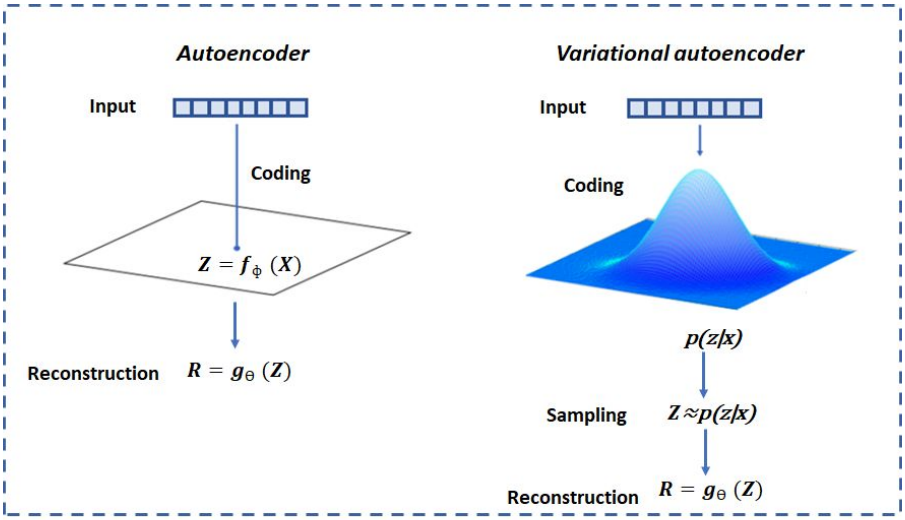


Figure 2.3: Main differences between a conventional and a variational autoencoder. Instead of just learning a function representing the data (a compressed representation) like conventional autoencoders, variational autoencoders learn the parameters of a probability distribution representing the input data.

The VAE algorithm assumes that there is no correlation between any latent space dimensions and, therefore, the covariance matrix is diagonal. In this way, the encoder only needs to assign each input sample to a mean and a variance vectors. In addition, the logarithm of the variance is assigned, as this can take any real number in the range $(-\infty, \infty)$, matching the natural output range from a neural network, whereas that variance values are always positive, see Figure 2.4.

In order to provide continuity and completeness to the latent space, it is necessary to regularize both the logarithm of the variance and the mean of the distributions returned by the encoder. This regularisation is achieved by

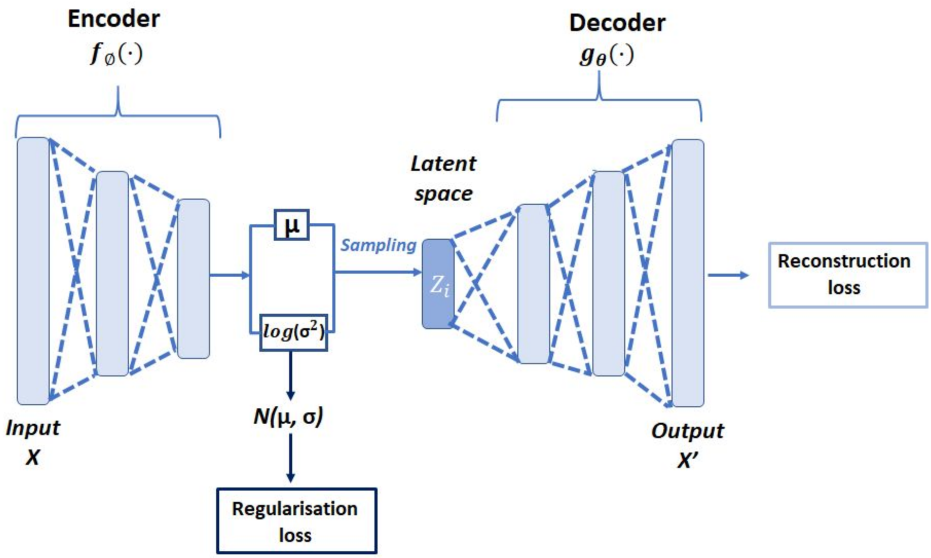


Figure 2.4: Architecture of a variational autoencoder. The proposed algorithm is optimized by minimizing two loss functions. One of them corresponding to the latent space regularisation and the other one corresponding to the input data reconstruction.

matching the encoder output distribution to the standard normal distribution ($\mu = 0$ and $\sigma = 1$).

After obtaining and optimizing the parameters of the mean and variance of the latent distributions, it is necessary to take samples of the learned representations to reconstruct the original input data. Samples of the encoder output distribution are obtained as follows:

$$Z \approx p(z|x) = \mu + \sigma \cdot \epsilon \tag{2.2}$$

where ϵ is randomly sampled from a standard normal distribution and $\sigma = \exp(\frac{\log(\sigma^2)}{2})$.

The minimized loss function in a variational autoencoder is composed of two terms: (1) a reconstruction term that compares the reconstructed data to the original input in order to get as effective encoding-decoding as possible and (2) a regularisation term in charge of regularizing the latent space organization, Figure 2.4. The regularisation term is expressed as the *Kulback-Leibler*

(KL) divergence that measures the difference between the predicted latent probability distribution of the data and the standard normal distribution in terms of mean and variance of the two distributions [75]:

$$D_{KL}[N(\mu,\sigma)||N(0,1)] = \frac{1}{2}\sum(1 + log(\sigma^2) - \mu^2 - \sigma^2) \tag{2.3}$$

The *Kulback-Leibler* function is minimised to 0 if $\mu = 0$ and $log(\sigma^2) = 0$ for all dimensions. As these two terms begin to differ from 0, the variational autoencoder loss increases. The compensation between the reconstruction error and the KL divergence is a hyper-parameter to be adjusted in this type of architecture.

2.3.3 Proposed method: Deep embedded refined clustering

Once the data dimensionality is reduced, we classify the samples in cancerous and non-cancerous. Reducing the data dimensionality without information about the different subjacent data distributions weakens the representativeness of the embedded features concerning the class and thereby, the performance of the subsequent classification worsens. For this reason, we consider that dimensionality reduction and classification should be optimized at the same time. In this context, we propose a deep refined embedded clustering (DERC) approach for classifying the DNA methylation data, see Figure 2.5. It is composed of an autoencoder in charge of the dimensionality reduction and a cluster assignment corresponding to the unsupervised classification stage (clustering layer in Figure 2.5). This approach is trained end-to-end optimizing the dimensionality reduction and the unsupervised classification in the same step and not in two different steps as all the algorithms proposed for DNA methylation analysis in the literature.

During the training process, the encoder and decoder weights of the autoencoder, W and W' respectively, are updated in each iteration in order to refine the latent features of the encoder output $\mathbf{Z}$. The proposed clustering layer (linked to the encoder output) obtains the soft-assignment probabilities $q_{i,j}$ between the embedded points z_i and the cluster centroids $\{\mu_j\}_{j=1}^{k}$ every T iterations, being k the number of cluster centroids. The soft-assignment probabilities ($q_{i,j}$) are obtained with the Student's t-distribution proposed in [65]. Using $q_{i,j}$, the target probabilities $p_{i,j}$ are updated, see Algorithm 1. These target probabilities allow the refinement of the cluster centroids by learning from the current high-confidence assignments. To take into account the refinement of the latent space carried out by the autoencoder while the

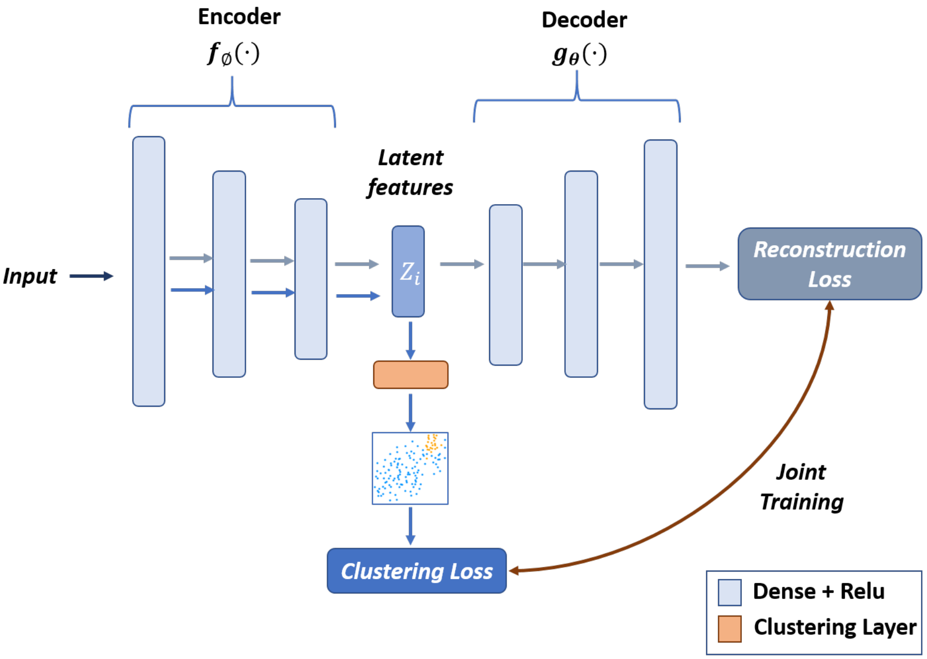


Figure 2.5: Architecture of the proposed method (DERC) to detect breast cancer using DNA methylation data. The proposed algorithm is trained to minimize both, clustering and reconstruction loss.

samples are classified in one of the two clusters (cancer and non-cancer), the proposed model is trained end-to-end minimizing both reconstruction L_{rec} and clustering loss $L_{cluster}$ terms:

$$L = L_{cluster} + \beta L_{rec} \tag{2.4}$$

where β balances the importance of the losses due to the reconstruction of the data. The term L_{rec} was defined in Equation (2.1) and it is minimized to obtain the maximum similarity between the input and the output data improving the representation of the latent space. $L_{cluster}$ is defined by the Kullback–Leibler (KL) divergence loss between the soft assignments and the target probabilities, $q_{i,j}$ and $p_{i,j}$ respectively:

$$L_{cluster} = \sum_i \sum_j p_{i,j} log \frac{p_{i,j}}{q_{i,j}} \tag{2.5}$$

The clustering term is minimized to achieve the soft-assignments $q_{i,j}$ and the target $p_{i,j}$ probabilities to be as similar as possible. In this way, the centroids are refined and the latent space obtained by the autoencoder is regularized to achieve a correct distinction between breast cancer and non-breast cancer samples. As discussed above, the hyper-parameter β balances the importance of losses due to the data reconstruction. If β is high, the data reconstruction term will predominate and the classification between cancerous and non-cancerous samples will worsen. Otherwise, if this term is too low, the reconstruction losses will be marginal and the features of the latent space will not be optimized correctly. Consequently, the latent features will be very different from the input data, decreasing the accuracy of the unsupervised classification. Therefore, β is a hyper-parameter that needs to be properly adjusted. In Step 2 of Algorithm 1, the methodology used to optimize the proposed DERC algorithm is detailed.

Note that to train the proposed method, a previous initialization of the centroids with latent characteristics is necessary (Step 1 of the Algorithm 1). In the experimental section, we present an experiment (Section 2.4.1.1) aimed at determining which of the dimensionality reduction models is optimal for this initialization.

2.4 Experimental Results

As we mentioned in Section 2.2, the DNA methylation databases used were obtained from the Gene Expression Omnibus (GEO) website. In this section, we used the dataset GSE32393 to evaluate the dimensionality reduction and the unsupervised deep clustering performance. The dataset GSE50220 was used as an external validation to demonstrate that the proposed method can generalise to other breast methylation databases. It should be noticed that all experiments were performed on an Intel i7 @ 3.10 GHz of 16 GB of RAM with a Titan V GPU of 12 GB of RAM. The proposed methods were executed in Python 3.5 using TensorFlow 2.0.

Algorithm 1 Proposed methodology for the DERC approach.

```
Input: Methylation data $\mathbf{X}$; number of clusters $k$; update interval $T$; batch-size $bs$; learning rate $lr$; number of samples $n$.
Output: Cluster assignment $\{c_i\}_{i=1}^{n}$ of each methylation sample $\{x_i\}_{i=1}^{n}$.
Step 1: Previous data dimensionality reduction
(1) Pre-train the proposed autoencoder algorithm.
(2) Obtain the centroid initialisation $\{\mu_j\}_{j=1}^{k}$ by running K-means on the latent space $\mathbf{Z}$ of the pre-trained autoencoder.
Step 2: Clustering using the proposed DERC method
End-to-end DERC optimization:
for $ite \leftarrow 1$ to MAXiter do
    %Choose a batch of samples $X_{bs} \subset \boldsymbol{X}$
    if $ite\%T == 0$ then
        $z_i \leftarrow f_\phi(x_i)$, $\forall x_i \in \mathbf{X}$ update $q_{i,j} \leftarrow \frac{(1+||z_i-\mu_j||^2)^{-1}}{\sum_{j'}(1+||z_i-\mu_j||^2)}$, $j \neq j'$ update $p_{i,j} \leftarrow \frac{q_{i,j}^2/f_j}{\sum_{j'} q_{i,j'}^2/f_{j'}}$, $f_j = \sum_i q_{i,j}$
    % Update encoder weights $W$, decoder weights $W'$ and centroids $\{\mu_j\}_{j=1}^{k}$:
    update $W \leftarrow W - \frac{lr}{bs}\sum_{i=1}^{bs}\left[\beta\frac{\partial L_{rec}}{W'} + \frac{\partial L_{cluster}}{\partial W}\right]$;
    update $W' \leftarrow W' - \frac{lr}{bs}\sum_{i=1}^{bs}\frac{\partial L_{rec}}{W'}$;
    update $\mu_j \leftarrow \mu_j - \frac{lr}{bs}\sum_{i=1}^{bs}\frac{\partial L_{cluster}}{\partial \mu_j}$;
Final prediction stage:
for $i \leftarrow 1$ to n do
    $c_i = argmax_j(q_{i,j})$
```

2.4.1 GSE32393 Series: Performance evaluation

2.4.1.1 Dimensionality reduction and unsupervised classification separately

As mentioned above, an initial latent space with a lower dimensionality than the input data for the cluster centroid initialisation is necessary. In this section, we detail a comparison between the latent space obtained using the conventional and the variational autoencoder and the unsupervised classification results after applying the K-means algorithm on each latent space. In this way, it will be demonstrated which algorithm is the most suitable for dimensionality reduction in the end-to-end proposed method.

Ablation experiment. The 10,153 CpG sites obtained after statistical analysis of the raw methylation data were the input of the proposed dimensionality reduction algorithms, conventional and variational autoencoders. In both cases, the dimensionality reduction was carried out using an architecture composed of 4 stacks. The number of neurons (input, output) of the 3 top layers were set to $\{(10153, 2000), (2000, 500), (500, 70)\}$, respectively (see Figure 2.6). These layers were composed of a dense layer with ReLU as activation function except for the last decoder layer that was constituted of the sigmoid function in order to obtain an output value between 0 and 1, range of the methylation data values. The kernel weights were initialised with random numbers drawn from a uniform distribution within $[-l, l]$, where $l = \sqrt{3 \cdot s / n_{input}}$, being $s = \frac{1}{3}$ and n_{input} the number of input units. The top layer output (latent space dimension) was set to $\{10,\ 20,\ 30\}$. Note that, these settings were obtained from empirical evaluations with a wide range of settings and we use only the best parameters here. After intense experiments, the optimal dimension of the latent space for both algorithms turned out to be 10 neurons.

To show the performance of AE and VAE and to demonstrate that they are not over-adjusted to the data, a first experiment was performed using 10% of the GSE32393 database as a validation set and the rest 90% as a training set. Subsequently, both algorithms were trained using the whole database (Entire prediction). The optimal hyper-parameters combination was achieved by training both algorithms during 300 epochs, using the Stochastic Gradient Descent (SGD) optimizer with a learning rate of 1 and a batch size of 8. Regarding the loss function, in the case of the conventional autoencoder, the mean square error (MSE) was used. However, the variational autoencoder loss function was composed of two terms: MSE weighted by 0.8 and Kullback–Leibler (KL) divergence. After training the dimensional reduction algorithms with the entire GSE32393 series and obtaining the features in the embedding space (encoder output), the classification results were obtained using K-means. To achieve this unsupervised classification, we ran K-means with 80 restarts and selected the best solution.

Qualitative and quantitative results. After training the proposed dimensionality reduction algorithms, the results in terms of reconstruction error for both autoencoders are shown in Table 2.1.

Table 2.1: Reconstruction error of the proposed dimensionality reduction algorithms. Conventional autoencoder (AE) and variational autoencoder (VAE).

Method	**Reconstruction Loss**		
	Training	Validation	Entire prediction
AE	**0.0062**	**0.0054**	**0.0057**
VAE	0.0082	0.0074	0.0082

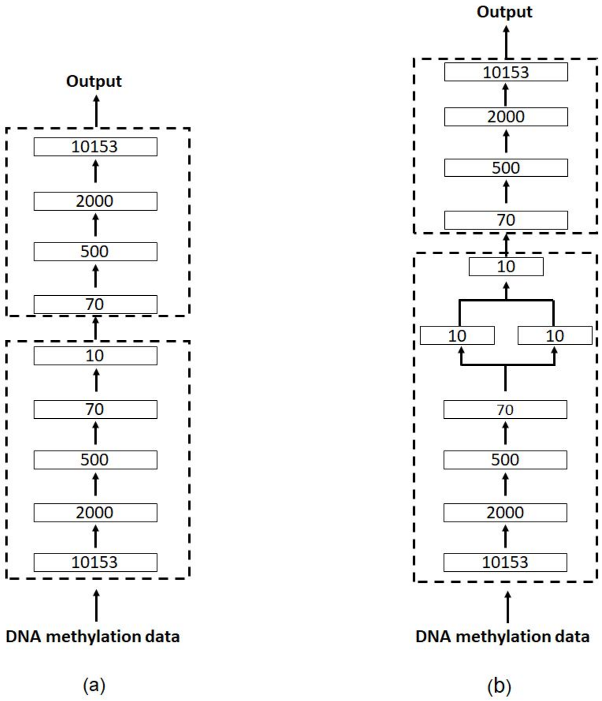
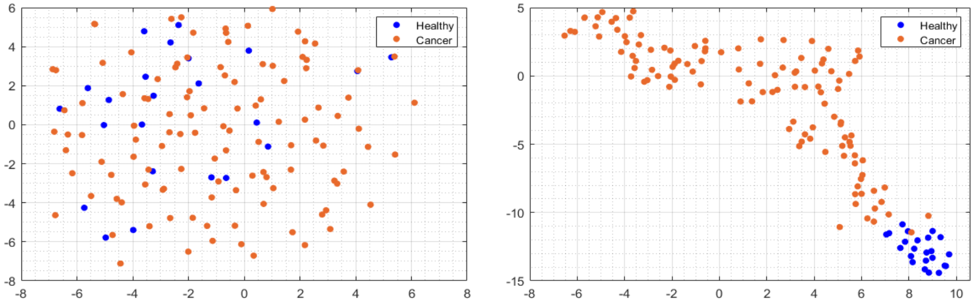


Figure 2.6: Final dimensionality reduction architectures. (a) Conventional autoencoder architecture composed of 4 stacks. (b) Variational autoencoder architecture composed of 4 stacks. Note that with the variational autoencoder algorithm, the latent space is obtained in two stages, two dense layers of 10 neurons representing the mean and the logarithm of the variance of the latent distribution and a sampling layer to obtain the points of the latent space.

In order to visualize in a qualitative way the effect of the tested dimensionality reduction methods (AE and VAE) over the data distribution, we used the t-distributed stochastic neighbor embedding (t-SNE) method to represent the latent space into a two-dimensional space. T-SNE is a nonlinear dimensionality reduction technique that embeds high-dimensional data into a space of two or three dimensions, which can then be visualized by a scatter plot [76]. In Figure 2.7, we show the representation of the data (latent space of the pre-trained variational autoencoder (a) and latent space of the pre-trained conventional autoencoder (b)) in a two-dimensional space.

To quantitatively evaluate the performance of the clustering assignments, several metrics were computed: the unsupervised clustering accuracy (ACC)



(a) (b)

Figure 2.7: Latent space of the dimensionality reduction algorithms. (a) Visualization of 10-dimensional features extracted by the latent space of the pre-trained variational autoencoder. (b) Visualization of 10-dimensional features extracted by the latent space of the pre-trained conventional autoencoder.

[74], the error rate (ER), the false positive (FP) and the false negative (FN) ratios, the adjusted rand index (ARI) [77] and the normalized mutual information (NMI) [78]. The ACC metric is defined as follows:

$$ACC = max_m \left(\frac{\sum_{i=1}^{n} 1\{y_i = m(c_i)\}}{n} \right) \tag{2.6}$$

where y_i is the ground-truth label, c_i is the cluster assignment generated by the algorithm, m is a mapping function that ranges over all possible one-to-one mappings between assignments and labels and n is the total number of samples. The error rate (%) is calculated according to the following formula:

$$ER\,(\%) = (1 - ACC) \cdot 100 \tag{2.7}$$

The adjusted rand index is defined as follows:

$$ARI = \frac{RI - E[RI]}{max(RI) - E[RI]} \tag{2.8}$$

where $RI = \frac{TP+TN}{TP+FN+TN+FP}$, TP and TN are true positive and true negative ratios and E[RI] is the expected index. The ARI can yield negative values if the index (RI) is less than expected index E[RI].

The normalized mutual information is defined by the following formula:

$$NMI = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{i,j} log \frac{n_{i,j}}{n_i \cdot \hat{n_j}}}{\sqrt{(\sum_{i=1}^{k} n_i log \frac{n_i}{n})(\sum_{j=1}^{k} \hat{n}_j log \frac{\hat{n_j}}{n})}} \tag{2.9}$$

where $n_{i,j}$ denotes the number of data points which are in the intersection between cluster c_i and class y_i, n_i is the number of data points in cluster c_i and $\hat{n}_j$ is the number of data points in class y_j.

In Table 2.2, the above-mentioned metrics were calculated for the input data + K-means clustering, the latent space of the pre-trained autoencoder(AE)+K-means and the latent space of the pre-trained variational autoencoder (VAE)+ K-means. Note that the input data is referred to the 10,153 features extracted after the statistical analysis.

Table 2.2: Comparison of the K-means clustering effect based on different feature extraction.

Method	***ACC***	***ER(%)***	***FN***	***FP***	***ARI***	***NMI***
Input Data + K-means	0.6715	32.85	45	0	0.1140	0.2355
AE + K-means	0.9343	6.57	9	0	0.7184	0.6300
VAE + K-means	0.5693	43.07	50	9	0.0133	0.0142
Proposed method	0.9927	0.73	1	0	0.9643	0.9212

2.4.1.2 Dimensionality reduction and unsupervised classification jointly (DERC)

After initializing the centroids using the algorithm with the lowest losses and the better prediction when K-means was used, in this case, the conventional autoencoder, the deep embedded refined clustering algorithm was trained. As we explained in Section 2.3.3, the β value, which weights the terms that composed the loss function of the DERC algorithm, is an important parameter to adjust. For this reason, we develop in this section a comparison between different β values exposing their influence on the clustering assignment.

Ablation experiment. After pre-training the conventional autoencoder model (with the parameters detailed in Section 2.4.1.1), we added the clustering layer to the output of the autoencoder latent space, see Table 2.3 for the layer layout in the final architecture.

Table 2.3: Architecture of the proposed deep embedded refined clustering model.

Layer name	**Output shape**	**Connected to**
Input_layer	10153	N/A
Encoder_0	2000	Input_layer
Encoder_1	500	Encoder_0
Encoder_2	70	Encoder_1
Encoder_3	10	Encoder_2
Decoder_3	70	Encoder_3
Decoder_2	500	Decoder_3
Decoder_1	2000	Decoder_2
Clustering_layer	2	Encoder_3
Decoder_0	10153	Decoder_1

In order to evaluate the β value on the performance of the clustering algorithm, we kept the rest of the hyper-parameters constant during the different experiments. In particular, the entire deep embedding clustering method was optimized by stochastic gradient descent (SGD) with a learning rate of 0.01 and a momentum of 0.9. The proposed method was trained during 50 epochs using a batch size of 8 samples and the target distribution of the clustering layer was updated every 10 iterations. Note that these hyper-parameters were obtained from empirical evaluations with a wide range of settings. β was a variable parameter and various experiments were conducted by setting its value with $\{0.95, 0.85, 0.75, 0.65\}$.

Quantitative results. In this case, we show the unsupervised classification results provided by the proposed deep embedded refined clustering (DERC) method depending on the β value (see Table 2.4).

Table 2.4: Comparison of the clustering effect of the proposed DERC based on different β values.

Method	***ACC***	***ER(%)***	***FN***	***FP***	***ARI***	***NMI***
DERC ($\beta = 0.95$)	0.9708	2.92	4	0	0.8643	0.7796
DERC ($\beta = 0.85$)	0.9781	2.19	3	0	0.8965	0.8198
DERC ($\beta = 0.75$)	**0.9927**	**0.73**	**1**	**0**	**0.9643**	**0.9212**
DERC ($\beta = 0.65$)	0.9854	1.4600	2	0	0.9298	0.8659

2.4.2 GSE50220 Series: Generalization ability of the DERC algorithm

In this section, we expose the results for the prediction of an external test set, see Table 2.5. The goal of this section is to demonstrate that the proposed DERC method could be valid to perform feature extraction and unsupervised classification from methylation data. Therefore, we made use of the GSE50220 series as an external test set to check the behavior of the proposed methods with new breast cancer samples.

Table 2.5: Results obtained over the external dataset. Note that in this case, the input data corresponds to the GSE50220 series after selecting the CpG sites extracted with the statistical analysis in the GSE32393 series.

Method	***ACC***	***ER(%)***	***FN***	***FP***	***ARI***	***NMI***
Input Data + K-means	0.6042	39.58	18	0	0.0445	0.2133
AE + K-means	0.8542	14.57	7	0	0.4727	0.4541
DERC ($\beta = 0.75$)	**0.9375**	**6.25**	**3**	**0**	**0.7374**	**0.6554**

2.4.3 Comparison with the state of the art

In order to provide the superiority of the proposed method for DNAm analysis, we compared our approach with well-known methods for high-dimensional clustering. In concrete, we use the divisive hierarchical clustering methods based on isometric mapping using the maximum distance between consecutive one-dimensional embeddings and the global minimum of the corresponding density estimator (i-DivClu-M and i-DivClu-D, respectively) [60] and the subspace methods based on local affine and convex hull [62] (LSC-aff. hull and LSC-conv. hull, respectively), see Table 2.6.

Table 2.6: Results reached by the state-of-the-art approaches in comparison with the proposed method when predicting the DNAm databases. Note that S1 refers to the primary set (GSE32393 series) and S2 to the external database (GSE50220 series).

	ACC		**ER (%)**		**FN**		**FP**		**ARI**		**NMI**	
	S1	**S2**	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**
i-DivClu-D [60]	0.7600	0.6700	24.00	33.00	33	16	0	0	0.2546	0.1000	0.3113	0.2422
i-DivClu-M [60]	0.8832	0.7917	11.68	20.83	11	10	5	0	0.5149	0.3203	0.3391	0.3618
LSC-aff. hull [62]	0.8248	0.5208	17.52	47.92	24	23	0	0	0.3934	-0.0560	0.3921	0.1545
LSC-conv. hull [62]	0.8248	0.5625	17.52	43.75	24	21	0	0	0.3934	-0.0210	0.3921	0.1700
Proposed	**0.9927**	**0.9375**	**0.73**	**6.25**	**1**	**3**	**0**	**0**	**0.9643**	**0.7374**	**0.9212**	**0.6554**

2.5 Discussion

In this work, we present a deep embedded refined clustering approach to automatically detect patients suffering for cancer using DNA methylation data. In concrete, the proposed algorithm was evaluated using two breast methylation datasets.

As it can be observed in Table 2.2, an optimal data dimensionality reduction is essential to improve the classification results when working with high-dimensional data. Using the K-means algorithm as non-supervised classifier, the dimensionality reduction carried out by the pre-trained AE on the input data (DNAm profiles obtained after statistical analysis) improves the ACC results from 0.6715 to 0.9343. However, with the VAE algorithm, the ACC results do not improve, obtaining a value of 0.5693. As it can be seen in Figure 2.7, the latent space of the VAE is centered around 0 due to the regularisation effect. This fact makes it impossible to distinguish between the different classes. Additionally, the reconstruction losses obtained by the VAE are higher than those reached by the AE (see Table 2.1). Therefore, it can be concluded that the conventional autoencoder is the most suitable algorithm to reduce the DNAm dimensionality.

Moreover, regarding the comparison between classifying separately and jointly to the dimensionality reduction, ACC results show an improvement from 0.9343 to 0.9927 when the dimensionality reduction and the unsupervised classification are optimized all at once. In Table 2.4, it can be observed the effect of the β value on the unsupervised classification results. In this way, it can be demonstrated that when the contribution of the reconstruction losses is too low (low β) or too high (high β), the ACC results are worse compared to a more balanced contribution. However, all the accuracy results shown in Table 2.4, joint optimization, are higher than those obtained when applying K-means in the autoencoder latent space, separate optimization. Therefore, it is proven that when the classification is carried out at the same time as the dimensionality reduction is optimized, the best results are obtained.

This fact can also be demonstrated when the results of the proposed method are compared to those obtained in the literature [26, 27]. As discussed in Section 2.1, in [27], the authors proposed a dimensionality reduction algorithm followed by several unsupervised classification algorithms. They applied their methods to the same breast cancer database used in this paper (GSE32393 series). Note that they obtained an error rate of 2.94 using a deep neural network (DNN) following a self-organizing feature map (SOM) compared to 0.73 obtained by the proposed method. Furthermore, their algorithm predicted 4 cancer examples as healthy, while our method only misclassifies one cancer sample. Therefore, it confirms that the proposed deep embedded refined clustering algorithm improves the results when it is applied to DNA methylation data. Additionally, the algorithm proposed in [26] used the same breast cancer database (GSE32393 series). In this case, they used a Non-negative matrix factorisation (NMF) for dimensionality reduction following by supervised algorithms for classification. Their main limitation is that, according to the authors, the NMF algorithm cannot directly reduce the number of features (27,578) to a lower dimension than the number of samples (137). Therefore, they used a method called column-splitting in which they separated the original data into different matrices. They could not reduce the original data to a single latent space because they had to reduce each data matrix independently. In this way, the overall information of all original features is not taken into account. They used a K-fold for the algorithm validation and obtaining a 100 % of accuracy when they used 900 and 2700 CpG sites. However, both resulting models were overfitted as it is demonstrated when they reduced the number of features to 540 and the accuracy dropped to 97.85 %, lower than achieved with the proposed method. Additionally, the authors of [26] claimed that it is important to reduce the number of features to a smaller space than the total number of examples. However, they were only able to reduce the features to 540 (due to NMF restrictions) which is about 5 times the number of examples they used to train their models.

The results obtained by our proposed method on the external database (GSE50220 series) reported closely similar values to those reached in the primary set (GSE32393 series). This fact indicates that the proposed deep clustering model is perfectly applicable to other breast tissue databases (see Table 2.5).

Furthermore, to objectively contrast the proposed method with other state-of-the-art high-dimensional clustering approaches, we replicated the experiments performed by [60, 62] with the two DNAm databases proposed in this paper (see Table 2.6). The results obtained show a clear outperformance of the proposed method with respect to the rest of the state-of-the-art models for all metrics. The methods based on convex hull clustering do not achieve satisfactory results on the DNAm databases due to their strong dependence on initialization. Both methodologies, isometric mapping for Divisive Clustering (i-Div) and Local Subspace Clustering (LSC) show a decrease in the performance when tested on the external database (S2), demonstrating that they are not scalable for the classification of methylation data, especially when the number of samples is limited.

2.6 Conclusion

In this paper, a deep embedded refined clustering based on breast cancer classification using DNA methylation data has been presented. To the best of the authors' knowledge, no previous studies using DNA methylation are based on algorithms that can optimise the dimensionality reduction and the classification of the data at the same time. As demonstrated throughout the manuscript, the method proposed in this paper improves the results of algorithms using dimensionality reduction and subsequent classification.

The proposed method allows the breast cancer classification using a latent space of only 10 features, which means a reduction in the dimensionality of 99.9637 %. The technology used in this study for data acquisition is the Illumina Infinium 27k Human DNA methylation Beadchip v1.2 which uses probes on the 27k array target regions of the human genome to measure methylation levels at 27,578 CpG dinucleotides in 14,495 genes. As verified through this work, many of the CpG sites obtained in the DNA methylation analysis are not relevant in the breast cancer classification. After ensuring model viability with a larger breast cancer database, the CpG sites from which the level of methylation is obtained could be reduced decreasing the cost and time of methylation analysis. Therefore, this work could contribute to a faster and more effective diagnosis of breast cancer, improving cancer care and advancing the future of breast cancer research technologies.

From a technical perspective, future lines of work will focus on adapting and applying the proposed method to identify and appropriately classify other challenging disorders, such as melanocytic tumours. In this way, the general applicability of the model for the detection of different types of cancer could be demostrated.

Chapter 3

An Attention-based Weakly Supervised framework for Spitzoid Melanocytic Lesion Diagnosis in Whole Slide Images

The content of this chapter corresponds to the author version of the following published paper: del Amor, R., Launet, L., Colomer, A., Moscardó, A., Mosquera, A., Monteagudo C. & Naranjo, V. An Attention-based Weakly Supervised framework for Spitzoid Melanocytic Lesion Diagnosis in Whole Slide Images. Artificial intelligence in medicine, 121, 102197 (2021).

Contents

3.1 Introduction **43**

3.2 Related work **47**

3.3 Material **49**

3.4 Methods **51**

3.4.1 Source model: ROI selection 51

3.4.2 Target model: WSI prediction 54

3.5 Ablation Experiments **55**

3.5.1 Database partitioning 55

3.5.2 Source model selection 56

3.5.3 Target model selection 58

3.6 Prediction Results **60**

3.7 Discussion **62**

3.7.1 Source model: ROI selection 62

3.7.2 Target model: WSI prediction 65

3.8 Conclusion **66**

An Attention-based Weakly Supervised framework for Spitzoid Melanocytic Lesion Diagnosis in Whole Slide Images

Rocío del Amor[1], Laëtitia Launet[1], Adrián Colomer[1], Anaïs Moscardó[2],
Andrés Mosquera-Zamudio[2], Carlos Monteagudo[2] and Valery Naranjo[1]

[1] Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano, Universitat Politècnica de València, 46022, Valencia, Spain.

[2]Pathology Department. Hospital Clínico Universitario de Valencia, Universidad de Valencia, Valencia, Spain.

Abstract

Melanoma is an aggressive neoplasm responsible for the majority of deaths from skin cancer. Specifically, spitzoid melanocytic tumors are one of the most challenging melanocytic lesions due to their ambiguous morphological features. The gold standard for its diagnosis and prognosis is the analysis of skin biopsies. In this process, dermatopathologists visualize skin histology slides under a microscope, in a highly time-consuming and subjective task. In the last years, computer-aided diagnosis (CAD) systems have emerged as a promising tool that could support pathologists in daily clinical practice. Nevertheless, no automatic CAD systems have yet been proposed for the analysis of spitzoid lesions. Regarding common melanoma, no system allows both the selection of the tumor region and the prediction of the benign or malignant form in the diagnosis. Motivated by this, we propose a novel end-to-end weakly supervised deep learning model, based on inductive transfer learning with an improved convolutional neural network (CNN) to refine the embedding features of the latent space. The framework is composed of a source model in charge of finding the tumor patch-level patterns, and a target model focuses on the specific diagnosis of a biopsy. The latter retrains the backbone of the source model through a multiple instance learning workflow to obtain the biopsy-level scoring. To evaluate the performance of the proposed methods, we performed extensive experiments on a private skin database with spitzoid lesions. Test results achieved an accuracy of 0.9231 and 0.80 for the source and the target

models, respectively. In addition, the heat map findings are directly in line with the clinicians' medical decision and even highlight, in some cases, patterns of interest that were overlooked by the pathologist.

3.1 Introduction

According to the World Health Organization, nearly one in three diagnosed cancers is a skin cancer [79]. The most dangerous skin cancer is melanoma which is responsible for 80 percent of skin cancer-related deaths [80]. Melanoma is an aggressive melanocytic neoplasm with numerous resistance mechanisms against therapeutic agents. In most melanocytic tumors, a precise pathological distinction between benign (nevus) and malignant (melanoma) is possible. However, there are still uncommon melanocytic lesions that represent a diagnostic challenge for pathologists. Among these, one of the most challenging lesions to diagnose is the so-called 'spitzoid melanocytic tumors' (SMTs), composed of spindled and/or epithelioid melanocytes with a large nucleus [81].

The final diagnosis of SMTs is confirmed by skin biopsies. The skin tumor is excised, laminated, stained with Hematoxylin and Eosin (H&E) and finally stored in crystal slides. Then, dermatopathologists analyze the sample under the microscope [81]. During the analysis of spitzoid lesions, different histopathological characteristics can be observed depending on the malignancy degree, see Figure 3.1. The regions with benign spitzoid lesions generally have a confluence of melanocytes in well-defined and organized nests. Figure 3.1 (a)-(b) shows sub-regions of a benign spitzoid melanocytic lesion. These regions show cellular and architectural maturation (both melanocytes and nests decrease in size towards the base of the lesion) throughout the dermis. In this case, this type of benign lesion is known as compound Spitz nevus. If the lesion only occurs in the epidermis and does not show extension into the dermis it would be called junctional nevus. In the case of spitzoid malignant lesions, cellular disorder is a frequent pattern, the melanocytic nests are ill-defined and are usually devoid of maturation, see Figure 3.1 (c). Additional features associated with malignancy of spitzoid melanocytic lesions include marked nuclear pleomorphism, pagetoid spread (individual cells or small aggregates of melanocytic cells grow and invade the upper epidermis from below) and a poor circumscription of lesions at their peripheries [82]. Figure 3.1 (d) shows an example of the pagetoid pattern. In addition to the cellular disorder, there are other local-level features associated with malignancy. Among these patterns, typical (bipolar and symmetrical) and atypical (aberrant mitotic figures, usually asymmetrical and/or multipolar) mitoses stand out. Note that benign melanocytic lesions can also have occasional typical mitoses, particularly in the most superficial areas. Therefore, if we find a typical mitosis in a spitzoid lesion, we should take into account additional factors such as the number of mitoses and their location within the lesion (deep typical mitoses are more suspicious of malignancy than the superficial ones) to determine if the neoplasm is malignant. Typical mitoses are only a sign of cellular proliferation and their mere presence cannot establish that a neoplasm is malignant. However, if numerous typical mitoses ($\succ 6/mm^2$) are found without evidence of a traumatic event, the probability of malignancy is high. Similarly, the presence of atypical mitoses in a spitzoid tumor favors malignancy. An example of typical and atypical mitoses on a malignant lesion are shown in Figure 3.1 (e)-(f), respectively. Table 3.1 summarizes the main features distinguishing normal tissue, tissue with benign and malignant spitzoid lesion. The manual diagnosis process is highly time-consuming and commonly leads to discordance between histopathologists due to the ambiguity of these neoplasms [83]. This is why these lesions represent a formidable diagnostic challenge.

Table 3.1: Main histological features of normal melanocytes and spitzoid lesions.

Histological features	Normal tissue	Benign spitzoid lesion	Malignant spitzoid lesion
Basal and periodically distributed isolated melanocytes	Yes	No	No
Melanocytic nests	No	Well defined	Ill defined
Pagetoid patterns	No	Rare	Yes/No
Typical mitoses	No	No/Few	Common (usually numerous)
Atypical mitoses	No	No	Yes/No
Necrosis	No	No	Yes/No
Ulceration	No	Very rare	Yes/No
Marked nuclear pleomorphism	No	No	Common

The computer-aided diagnosis systems (CADs) aim to support pathologists in the daily analysis of skin biopsies, reducing both the workload and the inconsistency generated. With the emergence of digital pathology, the digitization of histological crystals into whole-slide images (WSIs) has been standardized [84], leading the way to the application of computer vision methods. The development of CADs based on WSI analysis presents important hardware limitations because of their large size. For this reason, the typical approach generally involves extracting small patches from larger WSIs, resulting in thousands of patches per image. The convolutional neural networks (CNN)-based approaches have been extensively tested for the detection of breast cancer [85–87], prostate cancer [87–89] or lung cancer [90, 91]. However,

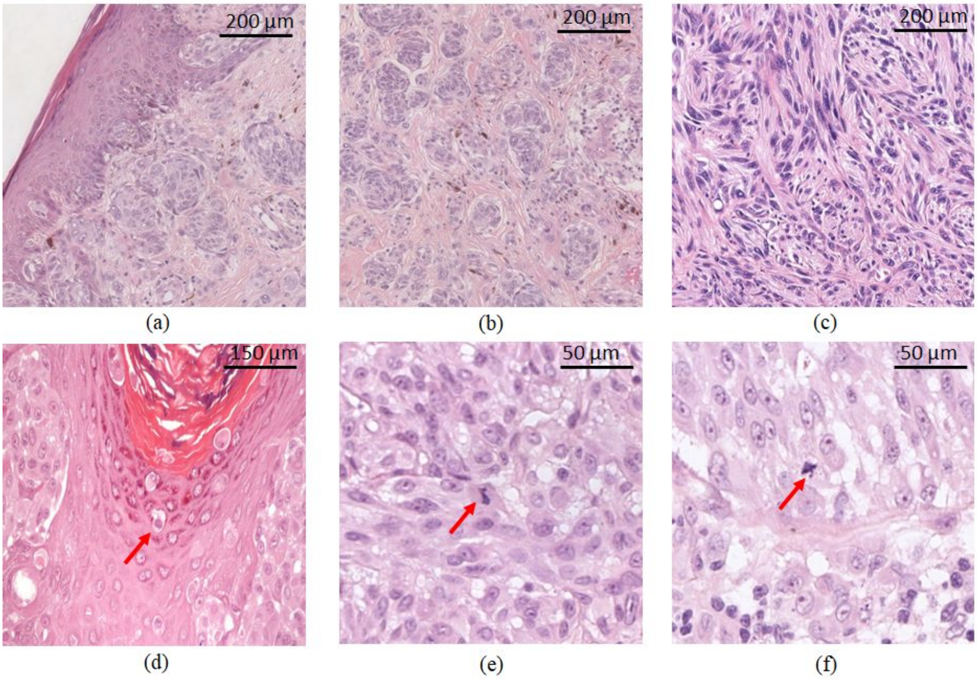


Figure 3.1: Representative patches extracted from WSIs presenting different spitzoid melanocytic lesions; (a)-(b): Benign spitzoid nevus containing well-defined melanocytic nests in an organized fashion; (c): Malignant lesion representative of the cellular disorder with ill-defined large tumor nest; (d): Malignant lesion with pagetoid spread, very common in this type of lesions; (e): Typical mitosis; (f): Atypical mitosis.

regarding skin cancer diagnosis, specifically for melanoma detection, most research was based on the analysis of dermoscopic images [92–100] and few studies have focused on the analysis of WSIs [101–104]. Hekler et al. [101] used transfer learning on a pre-trained ResNet50 CNN to differentiate between two classes, benign and melanoma tissues. The main limitation of this work is that they are not able to analyze entire WSIs but only a characteristic tumor sub-region. In De Logu et al. [102], a pre-trained Inception-ResNet-v2 network was then used to distinguish cutaneous melanoma areas from healthy tissues. However, this work didn't discriminate melanoma from nevi WSIs. In [103], the authors developed a deep learning system to automatically detect malignant melanoma in the eyelid from histopathological sections. The main limitation of this work is that the input of the algorithm is the tumor region and not the entire WSI image.

To the best of the authors' knowledge, no previous studies have focused on the SMTs distinction based on data-driven approaches. There is only one method based on hand-crafted feature extraction for SMTs identification [104]. In [104], the authors used a machine learning algorithm to assist in the diagnosis of SMT. In this study, a random forest classifier was used on numerical morphological characteristics extracted by the pathologists from histological images [104]. Therefore, the method does not extract features directly from the histological images. As SMTs are uncommon skin lesions, the available data is generally scarce. This is why this study used data from 54 patients.

Inspired by the main limitations of the studies focused on melanoma detection and more specifically on SMTs diagnosis, in this work, we put forward a novel semi-supervised inductive transfer learning strategy to conduct both the local automatic detection of tumor regions and the global prediction of an entire biopsy. In summary, the main contributions of this work are:

- Spitzoid histological images are used for the first time to develop an automatic feature extractor.
- A new attention-based backbone is proposed to extract more accurate features.
- A novel framework based on inductive transfer learning to solve at the same time ROI selection and malignancy detection is developed.
- Multiple instance learning-based solutions are formulated in a novel framework for spitzoid lesion detection using biopsy-level labels.
- A wide clinical interpretability of the results achieved with the proposed methods is provided.

The rest of the paper is structured as follows. Section 2 details the related work regarding inductive transfer learning and multiple instance learning strategies, then the underlying methodologies of the present work, and finally highlights the improvement introduced in medical research. In Section 3, we present the data used in this work, CLARIFYv1, a private database comprised of skin WSIs from patients with spitzoid tumors. In Section 4, we describe the proposed methodology, mainly composed of two stages: i) the development of a source model in charge of performing a patch-level classification to select tumor regions and ii) a target model based on a multiple instance learning approach to predict the malignancy degree at the biopsy level. Sections 5, 6 and 7 provide information on the performance outcomes related to the different classification tasks. Finally, in Section 8 we present our conclusions along with the future work.

3.2 Related work

Inductive transfer learning

Given a source domain D_S with a corresponding source task T_S, and a target domain D_T with a corresponding task T_T, transfer learning (TL) is the process of improving the target predictive function $f_T(\cdot)$ by using the related information from D_S and T_S, where $D_S \neq D_T$ or $T_S \neq T_T$ [105]. In the context of this work, we refer to inductive transfer learning (ITL) as the ability of the learning mechanism to enhance the performance on the target task (with a reduced number of labels) after having learned a different but related concept or skill on a previous task in the same domain [106]. The intuition behind this idea is that learning a new task from related tasks should be easier, faster and with better solutions or using less amount of labeled data than learning the target task in isolation. When the source and the target domain labels are available, the inductive transfer learning approach is known as multi-task learning.

Interest in this technique has grown in recent years in applications related to medical issues due to the promising results obtained. In this context, Caruana et al. suggested using multi-task learning in artificial neural networks and proposed an inductive transfer learning approach for pneumonia risk prediction [107]. Silver et al. introduced a task rehearsal method (TRM) as an approach to life-long learning that used the representation of previously learned tasks as a source of inductive bias. This inductive bias enabled TRM to generate more accurate hypotheses for new tasks that have small sets of training examples [108]. Zhang et al. used a technique based on inductive transfer learning to solve two-step classification problems: classification of malignant-nodule and non-nodule, and to classify the Serious-Malignant and the Mild-Malignant in malignant-nodule [109]. Tokuoka et al. provided an inductive transfer learning approach to adopt the annotation label of the source domain datasets to tasks of the target domain using Cycle-GAN based on unsupervised domain adaptation (UDA) [110]. Zhou et al. used an inductive transfer learning method to improve the performance of ocular multi-disease identification. In this case, the source and the target domain data were fundus images, but the source and target domain tasks were diabetic retinopathy lesion segmentation and multi-disease classification, respectively [111]. De Bois et al. used an inductive transfer learning approach to build a better glucose predictive model using a CNN-based architecture. A first model was trained on source patients that may come from different datasets and then, the model was fine-tuned to the target patients. Adding a gradient reversal layer, the patient classifier module made the feature extractor learn a feature representation that was general across the source patients [112].

In that context, we adopt an inductive transfer strategy to accurately classify instances from WSIs. The source model is trained to predict tumor regions by a patch-based CNN using inaccurate annotations with a large number of labels. After that, the backbone of the source model is retrained to classify nevus and malignant biopsies using a target model where the number of labels is reduced as this model is retrained at the biopsy level.

Multiple instance learning

Multiple instance learning (MIL), a particular form of weakly supervised learning, aims at training a model using a set of weakly labeled data [113]. In MIL tasks, the training dataset is composed of bags, where each bag contains a set of instances. A positive label is assigned to a bag if it contains at least one positive instance. The goal of MIL is to teach a model to predict the bag label. MIL approach has been successfully applied to computational histopathology for tasks such as tumor detection based on WSIs, reducing the time required to perform precise annotations [114–117]. In this vein, [114, 115] assigned the global label (cancerous against non-cancerous) to all patches of a slide. Campanella et al. [114] proposed a MIL-based deep learning system to accomplish the identification of three different cancers: prostate cancer, basal cell carcinoma and breast cancer metastases. In this case, they used an instance-level paradigm to obtain a tile-level feature representation through a CNN. These representations were then used in a recurrent neural network to integrate the information across the whole slide and report the final classification result to obtain a final slide-level diagnosis. Das et al. [115] used an embedded-space paradigm based on multiple instance learning to predict breast cancer. Specifically, they used a deep CNN architecture based on the pre-trained VGG19 network to extract the features of each bag. Then, the bag level representation is achieved by the aggregation of the features through the batch global max pooling (BGMP) layer at the feature embedding dimension. Silva et al. [117] used a novel weakly supervised deep learning model, based on self-learning CNNs, that leveraged only the global Gleason score of gigapixel whole slide images during training to accurately perform both, grading of patch-level patterns and biopsy-level scoring. Other works like [116] treated the tumor areas manually annotated by pathologists as a bag. In this case, the authors proposed a MIL method based on a deep graph convolutional network and feature selection for the prediction of lymph node metastasis using histopathological images of colorectal cancer. To the best of the authors' knowledge, no previous works have taken advantage of the promising MIL-based approaches for the diagnosis of melanocytic tumors yet. Our starting premise is that since there is at least one identifying patch of malignancy in a melanoma lesion, the MIL-based approach could assist in diagnosing a spitzoid lesion based on its whole context lessening the ambiguity between malignant and benign lesions. Additionally, in contrast to the works cited above, as in this study each bag contains the tumor region pseudo-labeled by the source model, the number of noisy labels is reduced, which will facilitate the model training-loop since the number of available samples is particularly limited.

3.3 Material

To evaluate the proposed learning methodology, we resort to a private database, CLARIFYv1, with histopathological skin images from different body areas that contain spitzoid melanocytic lesions. The database is composed of 53 biopsies from 51 different patients who signed the pertinent informed consent. The number of patients used in this study is relatively limited because these lesions are uncommon among the population. The tissue samples were sliced, stained and digitized using the Ventana iScan Coreo scanner at 40x magnification obtaining WSIs. The slides were analyzed by an expert dermatopathologist at the University Clinic Hospital of Valencia (CM). Specifically, 21 of the 51 patients under study were diagnosed as malignant melanocytic lesions (melanoma) and the rest as benign melanocytic lesions (nevus).

The global tumor regions, areas with spitzoid lesions, were annotated by the pathologists (AM, AM-Z and CM) using an in-house software based on the OpenSeadragon libraries [118]. With these annotations, WSIs were divided into regions of interest or ROI (tumor region) and non-interest regions (the rest of the WSI). Note that the tumor region denotes the part of the biopsy where the spitzoid lesion is found. After defining the tumor regions, the pathologist classified them as benign or malignant. Figure 3.2 shows the annotation of benign and malignant regions. To streamline the annotation task, these annotations were performed in a coarse way, so in some sub-regions there are

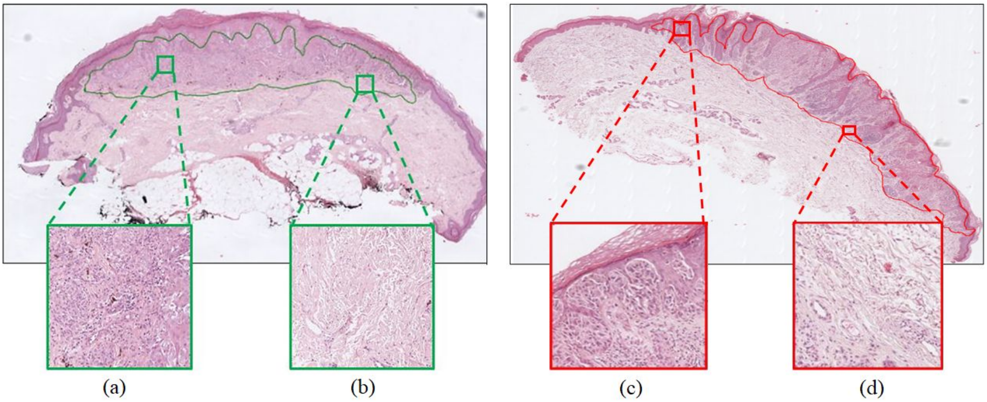


Figure 3.2: Annotation of a benign and a malignant spitzoid lesion. Patches (a) and (c) show characteristic patterns of the tumor region, benign and malignant respectively. Although patches (b) and (d) are inside the interest region annotated by the pathologists, these patches correspond to reactive stroma and do not contain tumor cells.

tumor discontinuities not considered. This fact is shown in Figure 3.2 (b) and (d), where these patches have not patterned related to the tumor lesion.

In order to process the large WSIs, these were downsampled to $10x$ resolution, divided into patches of size 512x512x3 with a 50% overlap among them. Aiming at pre-processing the biopsies and reduce the noisy patches, a mask indicating the presence of tissue in the patches was obtained by applying the Otsu threshold method over the magenta channel. Subsequently, the patches with less than 20% of tissue were excluded from the database. A summary of the database description is presented in Table 3.2. Note that, due to the irregular morphology of these lesions, the tumor shape is very different among patients, with the number of patches per patient varying considerably.

Table 3.2: CLARIFYv1 database description. Amount of whole slide images with their respective biopsy label (first row), number of patches of each tumor region (second row) and number of non-interest region (third row).

	Benign	**Malignant**
# **WSI**	30	21
# **Tumor patches**	3652	4726
# **Non tumor patches**	5842	8139

3.4 Methods

The methodological core of the proposed approach is a semi-supervised CNN classifier able to detect the tumor region in a WSI and classify it into either benign or malignant spitzoid lesions. The proposed workflow is composed of a source and a target model, (θ^s) and (θ^t) respectively. The first model (θ^s) allows to automatically obtain the patches with significant features of spitzoid neoplasms, Figure 3.3. Tumor patches selected by the first model are then transferred to a second model (θ^t), Figure 3.4. This second model discerns malignant and benign biopsies using a MIL paradigm.

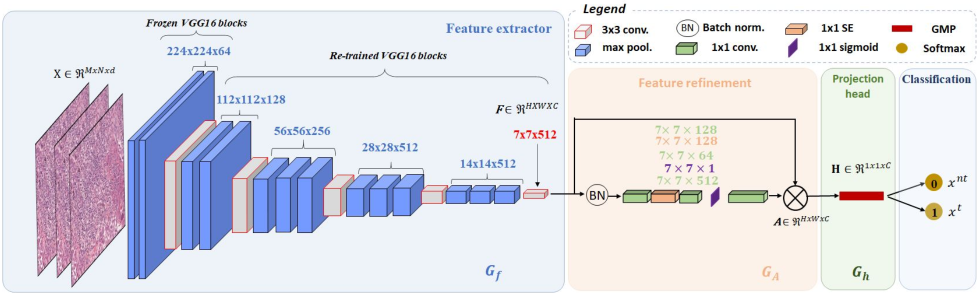


Figure 3.3: Overview of the proposed source model to conduct the tumor region detection. Blue and orange frames correspond to the base encoder network consisting of feature extraction and refinement. Note that VGG16 has been used as the feature extractor. After that, a projection head (green frame) maps the embedded representations in a lower-dimensional space to maximize the agreement in the classification stage (cyan frame).

3.4.1 *Source model: ROI selection*

The objective of this stage is to build a 2D-CNN architecture able to extract discriminatory features from WSI patches to distinguish tumor regions.

A. *Backbone*

(1) Feature extractor. The patch-level feature extractor $G_f : x \rightarrow F$ is a CNN which maps an image x into an F feature volume. Since the deep learning models trained from scratch report worse performance in comparison to fine-tuned models when the amount of available data is limited, we fine-tuned several well-known architectures: VGG16 [119], ResNet50 [120], InceptionV3 [121] and MobileNetV2 [122]. All architectures were pre-trained with around 14 million natural images corresponding to the ImageNet dataset. For the

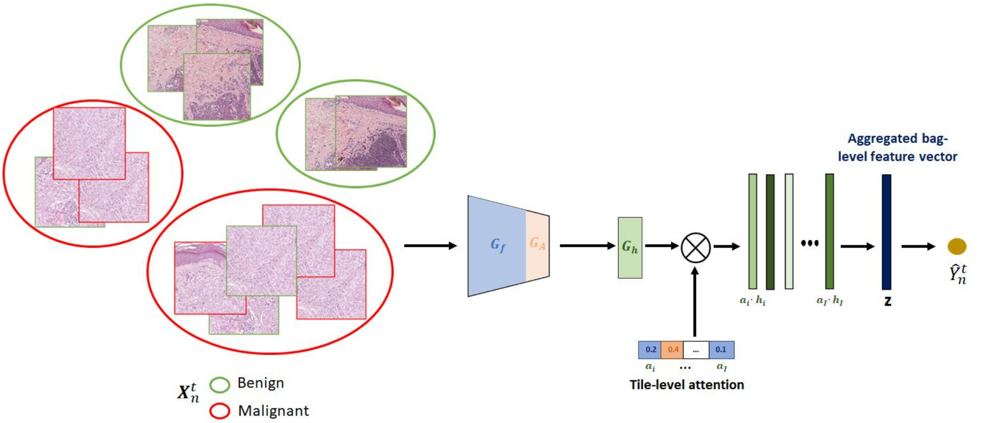


Figure 3.4: Pipeline showing the embedded-level approach for spitzoid melanocytic lesion classification. The weights of the pre-trained feature extractor and feature refinement of the source model (σ^s and δ^s) are used to initialize this approach. After that, we use the output of the projection head and tile-level attention to weight the patches in the prediction of a whole biopsy. Using an aggregated bag-level feature vector we classify the entire biopsy.

feature extraction stage, the base model is extracted from those pre-trained models and partially retrained. Since the patterns of the ImageNet dataset are very different from the histological ones (the value of the Frechet Inception Distance metric is around 68), it is optimal to keep the low-level features only (contours, combination of basic colors, general shapes, etc.). To this end, the weights of the first convolutional blocks from the pre-trained model are frozen, while the rest are re-trained to adapt the model to the specific application. The layer from which the freezing strategy is applied is empirically optimized for each architecture and it is specified in the experimental part of the paper, Section 5. Therefore, given a histological image $x \in \mathbb{R}^{M \times N \times d}$, where $M \times N \times d = 224 \times 224 \times 3$, a feature-embedded map $F \in \mathbb{R}^{H \times W \times C}$ is provided by the feature extractor. It is denoted as $F = G_f(x; \sigma^s)$ where σ^s is the set of trainable parameters of this source model.

(2) Feature refinement (SeaNet). Medical images always contain some irrelevant information that can disrupt the decision-making. For this reason, to solve ambiguous classification problems, it is essential to refine the features extracted by the CNN model. To this end, an attention module $G_A(F; \delta^s)$ was proposed to mimic the clinical behavior by focusing on the key features for the prediction, $G_A : F \to A$. In this case, the input of the attention module corresponds to the output feature map generated by the feature extractor,

$F \in \mathbb{R}^{H\times W\times C}$. The proposed attention module works as a kind of autoencoder composed of 1×1 convolutions in which the filters are decreased and increased, respectively. Therefore, the feature maps obtained at the output of each of these convolution layers will have the same spatial dimension as the previous feature map, with the difference that the number of channels will have been changed to accomplish a combination of the features. In order to explore the dependencies existing among the different feature channels as well as the contextual information, the blocks called 'Squeeze-and-Excitation' (SE) [123] were implemented between the different convolutional reduction layers of the attention module, see Figure 3.5.

The input to the SE block, $G \in \mathbb{R}^{H\times W\times R}$, is embedded into a $s \in \mathbb{R}^{1\times 1\times R}$ vector by a global average pooling (GAP) layer, which provides a global distribution of responses by channels. Note that the number of filters R, corresponds to the number of channels at the output of the convolutional layers of the attention module. In the following step, s is transformed into $\hat{s} = \phi(W_2(\partial(W_1 s)))$ where ϕ is the sigmoid activation function, $W_1 \in \mathbb{R}^{\frac{R}{r}\times R}$ and $W_2 \in \mathbb{R}^{R\times\frac{R}{r}}$ are the weights of two completely fully-connected layers (FC) and ∂ is the Relu activation function. The parameter r is the reduction ratio for dimensionality reduction, in this case $r = 4$, indicating the bottleneck. After the sigmoid activation, the activations of $\hat{s}$ are ranged to [0,1] and it is used to recalibrate the input $\boldsymbol{G} = [g_1, g_2, ..., g_c]$ where $g_i \in \mathbb{R}^{HxW}$. The output feature map of this block is $\boldsymbol{G}_{se} = [\hat{s_1} g_1, \hat{s_2} g_2, ..., \hat{s}_c g_c]$.

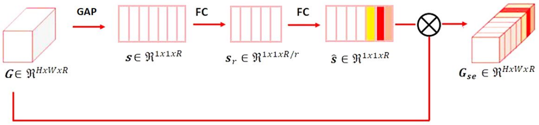


Figure 3.5: Architecture of the Squeeze-and-Excitation blocks used to exploit the dependencies between feature channels.

The last reduction layer of the attention module has the sigmoid as activation function to recalibrate the inputs and force the network to learn useful properties from the input representations. After increasing the number of filters to the same number as the input layer to this module, the output of the attention module is pondered with the output of the feature extractor obtaining a refined feature map $A \in \mathbb{R}^{H\times W\times C}$.

B. *Projection head module*

In this paper, we instantiate a projection head network, $G_h : A \rightarrow Z$, that maps the representations A to an embedding vector Z where the classification stage is addressed in a lower-dimensional space. In this case, different configurations already applied in the literature were tested in Section 5. In contrast to other widely used approaches such as the flattening of the activation volume resulting from the final convolutional block and the class prediction through consecutive fully-connected layers, the global max pooling (GMP) and the global average pooling (GAP) layers reduce the number of parameters decreasing the complexity of the model. At the end of the convolutional network, a softmax-activated dense layer is applied to address the tumor region identification.

3.4.2 Target model: WSI prediction

The target model aims to classify spitzoid lesions under an embedded-space paradigm using the biopsy-level labels for learning. To that end, our main goal is to find a compact embedding for the instances of a bag/WSI and combine these instance embeddings to a single embedding that represents the entire bag, see Figure 3.4.

Specifically, we denote each individual bag as $X_n^t = \{x_{n,1}^t, ..., x_{n,i}^t, x_{n,I_n}^t\}$, where $x_{n,i}^t$ is the i-th predicted tumor instance by the source model and I_n denotes the total number of predicted tumor region patches in a slide. Note that I_n can vary across bags. Hence, the objective of the target model becomes to obtain the label of a slide ($\hat{Y}_n^t$) from the tumor instances predicted by the source model ($x_{n,i}^t$), which can be defined as follows:

$$\hat{Y}_n^t = f(\{x_{n,1}^t, ..., x_{n,i}^t, ..., x_{n,I_n}^t\}, \omega^t) \tag{3.1}$$

where ω^t denotes the target model weights.

In order to find an embedding representation of each bag, we use the pre-trained backbone and the projection head module of the source model. In this manner, following an inductive learning strategy, the backbone already has prior knowledge concerning basic features of the histological database. After embedding each bag, $\mathbf{h}_n = G_h(G_A(G_f(X_n^t)))$, we obtain a C-dimensional feature vector for each instance. The bag label predictor $G_y : \{\mathbf{h}_i\}_{i \in I_n} \rightarrow \hat{Y}_n^t$ aggregates the C-dimensional feature vectors $\{\mathbf{h}_i\}_{i \in I_n}$ into a feature vector $Z_n \in \mathbb{R}^{1 \times C}$ representative of the bag. In the literature, there exist different aggregation functions such as batch global max pooling (BGMP) or batch global average pooling (BGAP). However, such functions are not flexible since they do not have trainable parameters. For this reason, in this work we use a trainable aggregation function [124]. In this case, $G_y(\cdot; \omega^t)$ is characterized by a set of trainable parameters $\mathbf{V} \in \mathbb{R}^{L \times C}$ and $\mathbf{w} \in \mathbb{R}^{L \times 1}$. The embedded feature vector per bag is obtained as $Z_n = \sum_{i \in I_n} a_i \cdot \mathbf{h}_i$, where a_i is defined as:

$$a_i = \frac{exp(\mathbf{w}^T tanh(\mathbf{V}\mathbf{h}_i))}{\sum_{j \in I_n} exp(\mathbf{w}^T tanh(\mathbf{V}\mathbf{h}_j))} \tag{3.2}$$

The attention-based aggregation function is differential and can be trained in a end-to-end manner using gradient descent. Additionally, the attention module not only provides a more flexible way to incorporate information from instances, but also enables us to localize informative tiles. The superiority of this aggregation function for spitzoid prediction will be shown in Section 5. Finally, the Z_n vector attaches to the dense layer with a sigmoid function-activated neuron to obtain the prediction at the biopsy level.

3.5 Ablation Experiments

In this section, we present the results of the different experiments carried out to show the performance of the proposed approach for the different classification tasks: patch-level classification (source model) and WSI prediction (target model). Note that a comparison with the current state-of-the-art methods was not possible as there are no algorithms focused on histological images of spitzoid tumors. Additionally, no public databases of histological images with melanocytic neoplasms have been found to apply our algorithms.

3.5.1 Database partitioning

Making use of the spitzoid database (CLARIFYv1), we carried out a patient-level data partitioning procedure to separate training and testing sets, aiming at avoiding overestimating the performance of the system and ensuring its ability to generalize. Specifically, 30% of patients were used to test the models, whereas the remainder of the database was employed to train the algorithm. To train the proposed models and optimize the hyperparameters involved in this process, the training set was divided following a 4-fold cross-validation strategy. We used four validation cohorts to optimize both the source and the target models. To encourage the source model to select the most relevant tiles, we used an instance dropout over the non-tumor region, since these represent the majority class. Specifically, instances were randomly dropped during the training, while all instances were used during the model evaluation.

3.5.2 Source model selection

A. *Backbone optimization*

According to the literature for histopathological image analysis, we compared as feature extractors the well-known ResNet50 and VGG architectures since they have reported the best performance [101, 103]. Additionally, we applied the proposed feature refinement SeaNet, Squeeze and Excitation Attention Network, on each of these feature extractors in order to evaluate the enhancement introduced. To address an objective comparison of the proposed backbones, we kept the projection head module constant using a GAP layer. In Table 3.3, we contrast the validation results achieved by the different backbones trained in a binary-class scenario. The comparison was handled by means of different figures of merit, such as sensitivity (SN), specificity (SPC), positive predictive value (PPV), false positive rate (FPR) negative predictive value (NPV), F1-score (F1S), accuracy (ACC) and area under the ROC Curve (AUC). Note that the figures of merit listed above report the results for the average of the validation cohorts in the cross-validation process.

Table 3.3: Classification results reached during the validation stage with the proposed fine-tuned architectures. SeaNet: Squeeze-and-Excitation network.

	VGG16	**SeaNet (with VGG16)**	**RESNET50**	**SeaNet (with RESNET50)**
SN	0.8057 ± 0.1247	**0.8310 ± 0.1061**	0.8200 ± 0.1223	0.7494 ± 0.1736
SPC	0.9070 ± 0.0343	**0.9298 ± 0.0185**	0.8850 ± 0.0243	0.9290 ± 0.0422
PPV	0.8448 ± 0.0856	**0.8814 ± 0.0495**	0.8061 ± 0.1005	0.8800 ± 0.0316
FPR	0.0930 ± 0.0343	**0.0702 ± 0.0185**	0.1150 ± 0.0243	0.0828 ± 0.0235
NPV	0.8894 ± 0.0649	**0.9100 ± 0.0232**	0.8830 ± 0.0761	0.8693 ± 0.0516
F1S	0.8183 ± 0.0865	**0.8654 ± 0.0805**	0.8022 ± 0.1126	0.8100 ± 0.0927
ACC	0.8752 ± 0.0357	**0.9031 ± 0.0262**	0.8611 ± 0.0558	0.8770 ± 0.0329
AUC	0.8600 ± 0.0584	**0.8810 ± 0.0566**	0.8400 ± 0.0813	0.8500 ± 0.0737

Additionally, class activation maps (CAMs) were computed to highlight the regions of interest at patch-level in which the proposed source model paid attention to predict the samples, see Figure 3.6 and Figure 3.7. The backbone reporting the best performance during the validation stage was selected as the base encoder network to address the head projection optimization.

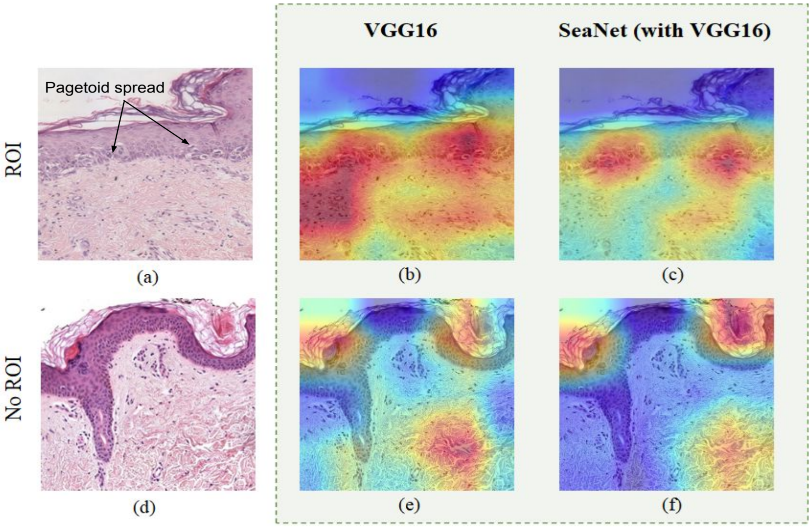


Figure 3.6: Class activation maps (CAMs) for images correctly classified as tumor region or ROI (first row) and non-tumor regions (second row). First column: original images; Second column: CAMs obtained using the VGG16 model. Third column: CAMs using Squeeze and excitation network (SeaNet) with VGG16 as the backbone. SeaNet model focuses on the most distinctive features and, in this case, pays attention to the pagetoid spread to define a patch as tumorous and to the healthy stromal region for the non-tumoral region.

Training details. All the contrasting approaches were implemented using Tensorflow 2.3.1 with Python 3.6. Experiments were conducted on the NVIDIA DGX A100 system. NVIDIA DGX A100 is the universal system for all artificial intelligence (AI) workloads, offering unprecedented compute density, performance, and flexibility in a 5 petaFLOPS AI system. After intense experiments, the optimal hyperparameters combination was achieved by training the models for 120 epochs using a learning rate of 0.001 with a batch size of 64. A stochastic gradient descent (SGD) optimizer was applied to minimize the binary cross-entropy (BCE) loss function at each epoch. The base model of the fine-tuned feature extractor was also optimized, selected to freeze the first convolutional block for VGG16 and setting all layers as trainable for ResNet50.

B. *Head projection optimization*

In this section, we report the validation performance using different projection head modules. Specifically, we compare a small multi-layer perceptron (MLP) with one hidden layer of 128 neurons non-linearly activated by the ReLU

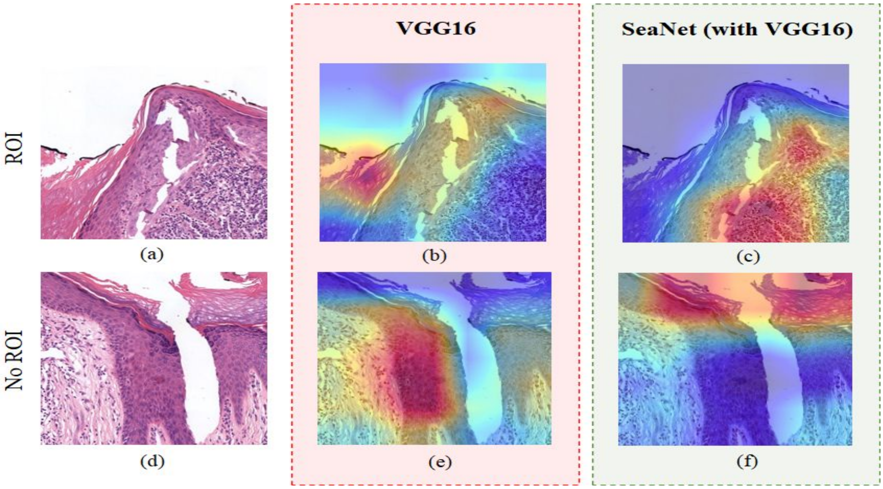


Figure 3.7: Original images (first column) and Class Activation Maps (CAMs) obtained with the VGG16 model (second column) and the Squeeze and excitation network (SeaNet) with VGG16 (third column). (b) and (e): Patches misclassified by the VGG16 model predicted as no ROI and ROI, respectively; (c) and (f): Patches well classified, ROI and No ROI respectively.

function, a global max-pooling (GMP) layer and a global average-pooling (GAP) layer, see Table 3.4. It is important to note that the comparison was conducted using the proposed SeaNet (with VGG16) backbone for all the scenarios.

Training details. The same hardware and software systems as for the backbone section were used to optimize the head projection. Additionally, we use the same learning rate, batch size, loss function and number of epochs as in the previous section. In this case, we only changed the head projection.

3.5.3 Target model selection

A. *WSI label predictor optimization*

As mentioned throughout the manuscript, the backbone and the projection head module of the target model were optimized during the ROI selection, via the source model. After obtaining an embedded feature vector of each tile in a bag, it is necessary to implement an aggregation function. In this section, we compare the results, when three different aggregation functions were used:

Table 3.4: Classification results reached during the validation stage using different projection head modules. SeaNet: Squeeze-and-Excitation network (with VGG16 as backbone), MLP: multi-layer perceptron, GMP: global max-pooling, GAP: global average-pooling.

	SeaNet+MLP	**SeaNet+GMP**	**SeaNet+GAP**
SN	0.8716 ± 0.3000	**0.8729 ± 0.0371**	0.8310 ± 0.1061
SPC	0.9076 ± 0.0478	0.9143 ± 0.0131	**0.9298 ± 0.0185**
PPV	0.8460 ± 0.1018	0.8589 ± 0.0710	**0.8814 ± 0.0495**
FPR	0.0927 ± 0.0340	0.0857 ± 0.0131	**0.0702 ± 0.0185**
NPV	0.9100 ± 0.0348	**0.9140 ± 0.0283**	0.9100 ± 0.0232
F1S	0.8606 ± 0.0655	**0.8708 ± 0.0541**	0.8654 ± 0.0805
ACC	0.8940 ± 0.0320	0.9020 ± 0.0164	**0.9031 ± 0.0262**
AUC	0.8800 ± 0.0391	**0.8935 ± 0.2490**	0.8810 ± 0.0566

batch global max pooling (BGMP), batch global average pooling (BGAP) and batch global attention summary (BGAS), Table 3.5.

Table 3.5: Classification results reached during the validation stage using different aggregation functions. BGMP: batch global max-pooling; BGAP: batch global average-pooling; BGAS: batch global attention summary.

	BGMP	**BGAP**	**BGAS**
SN	0.5000 ± 0.3953	0.5833 ± 0.3062	**0.7500 ± 0.2764**
SPC	**0.9000 ± 0.3953**	0.8500 ± 0.1658	0.8500 ± 0.2764
PPV	0.6250 ± 0.4330	0.8375 ± 0.1709	**0.8667 ± 0.1414**
FPR	**0.1000 ± 0.2909**	0.1500 ± 0.3062	0.1500 ± 0.2764
NPV	0.7625 ± 0.1546	0.7848 ± 0.1388	**0.8869 ± 0.1207**
F1S	0.5018 ± 0.3873	0.6000 ± 0.1541	**0.7472 ± 0.1473**
ACC	0.7361 ± 0.0977	0.7361 ± 0.0417	**0.8229 ± 0.0262**
AUC	0.7000 ± 0.1744	0.7167 ± 0.0841	**0.8000 ± 0.0963**

Training details. In order to generate bags and train the algorithms, a maximum of 300 image patches were randomly extracted from the source model prediction. In this case, the optimal results were obtained re-training the whole models during 100 epochs using a learning rate of 0.001 and a batch size of 1, in other words, one slide per batch. To minimize the BCE loss function at every epoch, the SGD optimizer was used.

3.6 Prediction Results

In this section, we show the quantitative and qualitative results achieved by the proposed strategies during the prediction of the test set. For both methods developed in this work, ROI selection and WSI classification, predictions were performed using the architectures with the best performance during the validation stage.

Quantitative results. Table 3.6 shows the results reached in the test prediction for the proposed source and target models.

Table 3.6: Classification results reached during the prediction stage. SM: source model; TM: target model. The proposed source model (SM) was composed of the SeaNet (with VGG16) + global max-pooling (GMP). The proposed target model (TM) used the batch global attention summary (BGAS) layer as an aggregation function.

	SM	**TM**
SN	0.9285	0.6700
SPC	0.9202	0.8900
PPV	0.8622	0.8000
FPR	0.0798	0.1111
NPV	0.9599	0.8000
F1S	0.8942	0.7300
ACC	0.9231	0.8000
AUC	0.9244	0.7800

Qualitative results. To qualitatively show the performance of the ROI selection model, we obtained probability heatmaps of representative samples indicating the presence of tumor region in the WSIs, Figure 3.8.

In the probability maps, for each pixel, the predicted probabilities for the ROI are estimated by bilinearly interpolating the predicted probabilities of the closest patches in terms of euclidean distance to the center of the patches. In addition, using these heatmaps, we visualize the distribution of attention weights, which were calculated for cases correctly classified into benign and malignant neoplasms, see Figure 3.9.

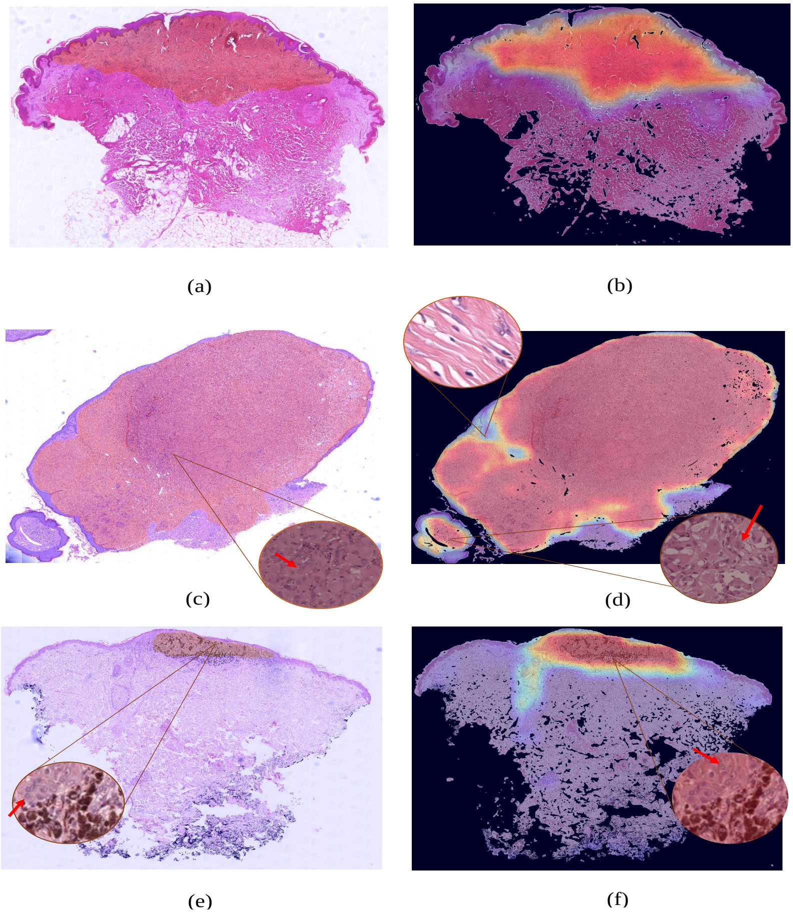


Figure 3.8: Whole slide image-level prediction for the source model (ROI estimation). (a) Manual annotation by experts; (b) System prediction completely in line with the annotation of (a); (c) Manual annotation by experts with expansion of areas with melanocytic nests characteristic of the lesion; (d) System prediction with certain areas annotated by the pathologists predicted as non-tumor regions. The expansion of the areas where there are no activations demonstrate that there are no melanocytic nests characteristic of the lesion; (e) Manual annotation by experts with expansion of area where melanocytic cells with melanosomes are found; (d) System prediction with expansion in the regions not annotated by the pathologist to demonstrate the presence of tumor cells.

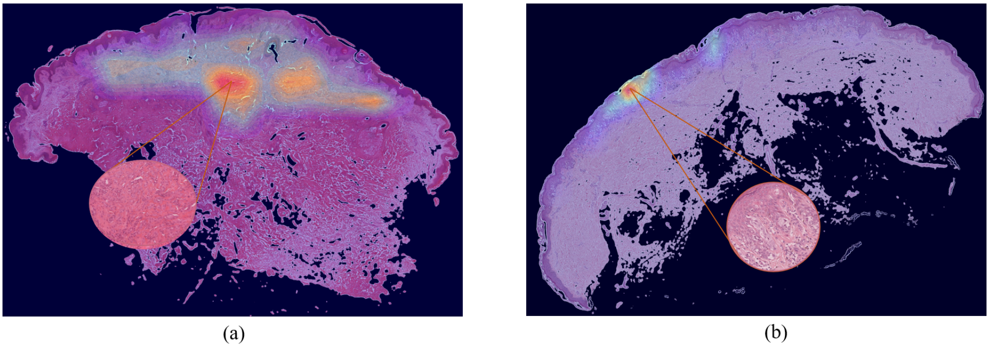


Figure 3.9: Visualization of the attention weights of the bag aggregation function in heat maps. (a) Benign sample; (b) Malignant sample.

3.7 Discussion

In this section, we make reference to the main contributions detailed throughout the paper and review the results obtained.

In contrast to the state-of-the-art studies for histological images classification, in which the input of the prediction model is the tumor region annotated by the pathologist, in this paper, we propose a framework able to first automatically select neoplastic regions of interest and then predict the malignancy or benignity of spitzoid neoplasms. Note that no previous studies seem to have proposed any automated method for the detection of these challenging neoplasms. Due to the absence of public spitzoid databases, the developed algorithms could not be validated with external databases, which can lead to biased results, according to the database used.

3.7.1 Source model: ROI selection

A. *About the ablation experiment*

Backbone selection. As a first stage, we carried out an optimization of the feature extractor for the selection of the tumor regions. Considering the limited amount of available samples, we decided to use the fine-tuning technique on the VGG16 and RESNET architectures. Particularly, from Table 3.3 we can observe that the use of sequential approaches (VGG16) provided slightly better results than architectures with residual blocks (RESNET). This fact is evidenced in several works in the literature for histopathological analysis where the sequential models used outperform residual ones [88]. Additionally, the proposed SeaNet module, characterized by the refinement of the features via convolutional attention blocks, reported a significant outperforming. Specifically, the SeaNet module via fine-tuning VGG16 architecture, achieved the best results. The use of the attention module provides more distinctive feature maps and allows a considerable reduction in the incidence of false positive and false negative samples, leading to improve global metrics. Aiming at qualitatively observing the enhancements introduced by the refinement module, the CAMs of the best models (SeaNet with VGG16 and VGG16 alone) were obtained for correctly classified images (see Figure 3.6) and for images misclassified by the VGG16 model (see Figure 3.7). In Figure 3.6, we can see that both for the prediction of patches belonging to the tumor region (a) and for non-tumor ones (b), the SeaNet activations are focused on smaller regions. For the ROI prediction, the SeaNet (with VGG16) model is mainly focused on the pagetoid pattern present within the epidermis, defining the region as tumor. However, the VGG16 model extends its activations to lymphocytes found within the dermis. In this case, the lymphocytes do not necessarily determine that the region is tumorous, since this small amount of lymphocytes can also be found in healthy regions. Therefore, the VGG16 model without the attention module introduces certain noise in the prediction. Regarding the prediction of non-tumor regions, both models are focused on the epidermis and stromal region of the dermis. Regarding the cases where VGG16 misclassifies tumor regions, Figure 3.7 (b), the activations are focused on the epidermis region. In this case, the epidermis region has no patterns indicative of a melanocytic lesion, but for a correct classification, the activations would have to be focused on the melanocyte aggregate found in the upper region, as in the case of the SeaNet model, see Figure 3.7 (c). In this region, we find a large number of melanocytic cells with a high concentration of lymphocytes indicating an inflammatory reaction to a tumor region. For the case of the non-tumor region shown in Figure 3.7 (d), the VGG16 model erroneously predicts it by focusing on the melanocytic cells found in the epidermis, see Figure 3.7 (e). Normally, in healthy skin, the dermo-epidermal junction is composed of isolated melanocytic cells with a certain spacing between them. It is representative of a tumor when these cells ascend to the upper layers of the epidermis forming what is known as a pagetoid pattern or infiltrate the dermis forming nests. Furthermore, in this case, the epidermis has no patterns that would be representative of a melanocytic lesion. Unlike the VGG model, the SeaNet (with VGG16) model reports its activations in the epidermal region and based on it establishes the correct prediction, classifying this patch as non-characteristic of a spitzoid lesion, see Figure 3.7 (f).

In any case, the inclusion of the proposed attention module outperforms the popular pre-trained architectures of the state of the art and reduces the number of noisy patches used as input to the target model.

Projection head module selection. After optimizing the backbone, we proceeded to select the projection head module that provided the best results. For this purpose, we tested three projection head modules: multilayer perceptron (MLP), global average pooling (GAP) and global max pooling (GMP). Table 3.4 shows that the modules based on GAP and GMP provide very similar and significantly better results than those reported by the MLP. The outperforming of GMP and GAP compared to the fully-connected configuration could be explained by the reduction in the number of weights to be optimized, making the model simpler and more capable of generalizing to new images. Comparing the results provided by GAP and GMP, we can conclude that they are very similar. The main difference between these techniques lies in the method of squeezing the spatial dimension. While GMP considers only the maximum value for the feature map, in the GAP layer the whole spatial region contributes to its output. This explains why the GMP layer enhances SN results and the GAP layer improves SPC results. With the GMP layer, it is more likely to correctly classify a patch belonging to the tumor region, even if it contains a minimal tumor region. However, GAP takes into account the whole context so that regions with small tumor areas are likely to be discarded. Although both show a very similar result, global metrics such as F1S and AUC exhibit a slight improvement with the GMP layer. Therefore, the GMP layer will be preferred as the optimal head projection module.

B. *About the prediction results*

Table 3.6 shows the results reached by the proposed ROI selection model. All the metrics reported here outperform those obtained in the validation phase. Figure 3.8 shows the probability maps for the lesion region of three test samples. The majority of the lesion regions predicted by the algorithm are depicted in Figure 3.8 (b), in which the prediction is completely in line with the annotation performed by the pathologists, Figure 3.8 (a). Some activation maps, such as those shown in Figure 3.8 (d), predict certain areas annotated by the pathologists as non-tumor regions. However, if we visualize the expansion of the areas where there are no activations, we can see that there are no melanocytic nests characteristic of the lesion, and therefore, we may be facing a discontinuity of the lesion as explained in Section 3.3. In contrast, in the lower part of Figure 3.8 (d), there are activations of tumor regions that have not been annotated by expert pathologists, see Figure 3.8 (c). However, if these regions are enlarged, it can be concluded that tumor cells are present. At times, due to the large amount of material in a lesion, pathologists can overlook some tumor areas. In the case of Figure 3.8 (e) and (f), there is also some discrepancy between the annotations performed by the pathologists and the activations predicted by the model. In these figures, we find melanocytic cells with melanosomes that give them their characteristic brown color. It is difficult to differentiate these tumor cells from melanophages (cells with brown staining and all of the same size) that are not tumor cells. In this case, if we zoom the activations of the algorithm (Figure 3.8 (f)) in those regions not annotated by the pathologist, we can see that there are also tumor cells. Therefore, the developed algorithm could help the decision-making in cases where there is ambiguity for the pathologists. In this context, the developed method enhances the detection of tumor areas.

3.7.2 Target model: WSI prediction

A. *About the ablation experiment*

WSI label predictor optimization. As discussed throughout the document, the backbone used by the target model was optimized during the selection of the source model. Therefore, in this case, it was only necessary to optimize the aggregation function required to perform a prediction using a MIL approach. From Table 3.5, we can observe that the use of the feature average of all patches containing a bag to obtain the embedded representation provides the best results (BGAP and BGAS aggregation functions). Additionally, the BGAS aggregation function improves the results provided by BGAP thanks to the introduction of optimized attention weights by updating the bag-level predictor weights (ω^t), achieving a validation accuracy of 0.8229. Therefore, we can conclude that the introduction of the attention module allows focusing on more relevant patterns, thus improving the final classification.

B. *About the prediction results*

Table 3.6 shows the results reached by the proposed target model in the test set. The results are in line with those obtained in the validation phase. Although the results are promising, there are some biopsies that are misclassified by the algorithm. This is because these types of lesions occasionally do not have universally accepted guidelines that can guarantee their specific diagnosis. Figure 3.9 shows the attention weights of the BGAS aggregation function for benign (Figure 3.9 (a)) and malignant (Figure 3.9 (b)) samples. The attention weights were normalized between 0 to 1 in each bag. The red regions in the attention weight maps represent the highest contribution for classification in each bag. Therefore, the bag class label is predicted by only using instances for which the attention values are large. In the case of a benign sample (Figure 3.9 (a)), the regions contributing to the class establishment are distributed over a wide area of the lesion, these areas being aggregates of melanocytes. However, the large attention weights for a malignant lesion are focused on small region characteristics of malignancy (in this case pagetoid pattern) as shown in Figure 3.9 (b).

3.8 Conclusion

In this work, we propose an inductive transfer learning framework able to perform both ROI selection and malignant prediction in spitzoid melanocytic lesions using WSIs. Our proposed framework is composed of a source model in charge of selecting the patches with characteristic lesion patterns. The source model introduces an attention module able to refine the features of the latent space to maximize the classification agreement. Using the backbone of the source model as a patch-level feature extractor and under a multiple instance learning approach, the target model predicts the malignancy degree by taking as input the tumor patches predicted by the first model. This innovative approach carried out in an end-to-end manner reported promising results for both ROI selection and WSI classification, achieving a testing accuracy of 0.9231 and 0.8000 for the source and the target models, despite the limited number of samples. Thus, our framework bridges the gap with respect to the development of automatic diagnostic systems for spitzoid melanocytic lesions. In future research lines, efforts should focus on improving the discrimination of malignancy and benignity with the acquisition of new samples and enhancements to the implemented attention module in the multiple instance learning approach.

Chapter 4

Constrained Multiple Instance Learning for Ulcerative Colitis prediction using Histological Images

The content of this chapter corresponds to the author version of the following published paper: del Amor, R., Meseguer, P., Parigini T., Villanacci v., Colomer A., Launet L., PICASSO team, & Naranjo, V. Constrained Multiple Instance Learning for Ulcerative Colitis prediction using Histological Images. Computer Methods and Programs in Biomedicine, 224, 107012. (2022).

Contents

4.1 Introduction **69**
4.2 Related work **71**
4.2.1 Multiple instance learning 71
4.2.2 Constrained CNNs 72
4.3 Methodology **73**
4.3.1 Problem formulation 73
4.3.2 MIL backbone with location constraints 74
4.3.3 MIL attention-embedding weights 76
4.4 Experiments and Results **76**
4.4.1 Implementation 76
4.4.2 Ablation experiments 78
4.4.3 Comparison to the literature 81
4.5 Conclusion **82**

Constrained Multiple Instance Learning for Ulcerative Colitis prediction using Histological Images

Rocío del Amor [1], Pablo Meseguer [1], Tommaso Lorenzo Parigini [1], Vincenzo Villanacci [3], Adrián Colomer [1], Laëtitia Launet [1]. PICASSO team and Valery Naranjo[1]

[1]Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano, Universitat Politècnica de València, 46022, Valencia, Spain

[2] Department of Biomedical Sciences, Humanitas University, Milan, Italy

[3] Institute of Pathology, ASST Spedali Civili, University of Brescia, Brescia, Italy

Abstract

Deep learning-based models applied to digital pathology require large, curated datasets with high-quality (HQ) annotations to perform correctly. In many cases, recruiting expert pathologists to annotate large databases is not feasible, and it is necessary to collect additional labeled data with varying label qualities, e.g., pathologists-in-training (henceforth, non-expert annotators). Learning from datasets with noisy labels is more challenging in medical applications since medical imaging datasets tend to have instance-dependent noise and suffer from high inter/intra-observer variability. In this paper, we design an uncertainty-driven labeling strategy with which we generate soft labels from 10 non-expert annotators for multi-class skin cancer classification. Based on this soft annotation, we propose an uncertainty estimation-based framework to handle these noisy labels. This framework is based on a novel formulation using a dual-branch min-max entropy calibration to penalize inexact labels during the training. Comprehensive experiments demonstrate the promising performance of our labeling strategy. Results show a consistent improvement by using soft labels with standard cross-entropy loss during training ($\sim 4.0\%$ F1-score) and increases when calibrating the model with the proposed min-max entropy calibration ($\sim 6.6\%$ F1-score). These improvements are produced at negligible cost, both in terms of annotation and calculation.

4.1 Introduction

Ulcerative colitis (UC) is a chronic inflammatory bowel disease (IBD) affecting the colon and the rectum with a propensity to arise in adolescents and young adults. The incidence of UC has been increasing globally [125] and currently ranges from 4 to 20 per 100,000 in North America and Europe [126].

The treatment of UC aims to extinguish bowel inflammation and prevent complications. Histological assessment plays a critical role in determining inflammatory activity. In this vein, histologic remission (HR) (also referred to as histologic healing, HH) is emerging as the most rigorous target of treatment and is associated with favorable clinical outcomes [127–130]. However, incorporating histology into clinical practice remains challenging. This is due to: (1) the lack of a universal definition of HR that varies depending on the histological score/index applied, (2) the complexity of most scores and (3) the high inter-observer variability between pathologists [128, 131–133].

Over the past decades, more than 30 histological scores have been developed, although their adoption in clinical practice remains modest [134, 135]. Similarly, different definitions and criteria of HR have been proposed, ranging from 'elimination of mucosal ulceration/erosion' to 'complete histological normalization'. Almost all investigators now agree that the absence of neutrophilic infiltration ('neutrophil-free' mucosa) is the key to define HR [135–138]. Indeed, this has been endorsed by two independent expert panels [138, 139]. Recently, our medical team developed a simplified histological score, PICASSO Histological Remission Index or PHRI, see Table 4.1 [41].

The primary aim of PHRI was to create a simple 'neutrophil only' histologic evaluation that predicted specified clinical outcomes. The structures of the biopsy where to evaluate the presence or absence of neutrophils and predict histological remission are: (a) lamina propia, (b) surface epithelium, (c) cryptal epithelium and (d) cryptal lumen, see Figure 4.1.

The computer-aided diagnosis systems (CADs) based on artificial intelligence (AI) aim to support pathologists in the daily analysis of histological biopsies, reducing both the workload and the inconsistency generated. Their final goal is to produce a reliable and reproducible real-time assessment of disease activity. With the emergence of digital pathology, the digitization of histological tissue sections into whole-slide images (WSIs) has been standardized, leading to the application of computer vision methods. Additionally, previous research showed the applicability of computer vision methods based on deep-learning approaches using WSIs for cancer detection, inflammatory prediction, etc.

Table 4.1: PICaSSO Histologic Remission Index (PHRI) to predict histological remmision.

Histologic finding	**Score**
Neutrophil infiltration in lamina propria	
Absent (No)	0
Present (Yes)	1
Neutrophil infiltration in epithelium	
Absent (No)	0
Present (Yes)	
- Surface epithelium	1
- Cryptal epithelium	1
- Crypt abscess	1
Total Score = sum of all above (maximum 4)	

Regarding the detection of UC activity based on deep learning techniques, available research has focused on the analysis of endoscopic images [140–144], but so far, only one study has approached the analysis of WSIs [145]. In [145], the authors used a deep learning algorithm to quantify the density of eosinophils in sigmoid colon biopsies from consecutive UC patients with histologically active disease. The algorithm was applied to sigmoid and colon biopsies from a cross-sectional cohort of 88 UC patients with histologically active disease as measured by the Geboes score and Robarts histopathology index (RHI). However, this study does not differentiate between remission and active WSI.

To the best of our knowledge, no previous study based on deep learning has been carried out to identify UC activity based on neutrophils detection using WSI, which has proven to be an accurate indicator of disease activity. In this work, we present a novel deep learning strategy to distinguish histological remission from activity based on the detection of neutrophils following the PHRI index. In summary, the main contributions of this work are:

- A deep learning framework used for the first time to accurately predict ulcerative colitis activity based on neutrophil detection.

- A novel constrained formulation that leverages prior knowledge in terms of relative tissue location (i.e. neutrophil location in the WSI) by imposing constraints on the feature extractor at bag (WSI)-level.

- A new attention weight for embedding-level MIL, which enlarges the relevance of the positive instances.

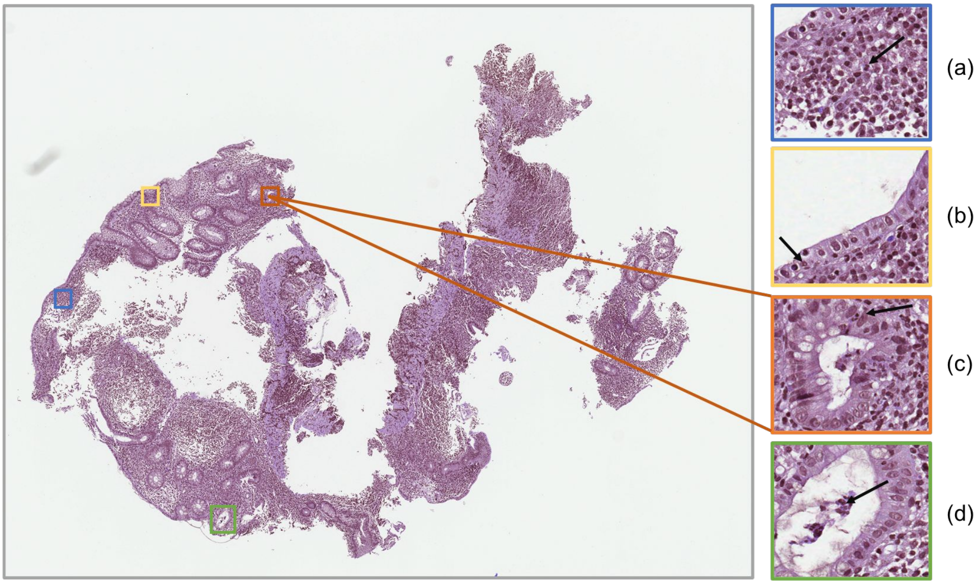


Figure 4.1: The larger image corresponds to a Whole-Slide Image (WSI) of a patient suffering from ulcerative colitis. The patches marked with colours denote different interest structures. Specifically: (a) lamina propia, (b) surface epithelium, (c) cryptal epithelium and (d) cryptal lumen. The black mark indicates the presence of a neutrophil.

- We benchmark the proposed model against relevant body of literature on PICASSO-MIL, a large cohort of biopsies collected and digitalized in 7 centers in the UK, Germany, Belgium, Italy, Canada and USA.
- Comprehensive experiments demonstrate the superior performance of our model. By simply incorporating information about neutrophil location during the training, we found improvements of nearly 10% for bag-level classification compared to prior MIL methods.

4.2 Related work

4.2.1 Multiple instance learning

Multiple instance learning (MIL), a particular form of weakly-supervised learning, aims at training a model using a set of weakly labeled data [113]. In MIL tasks, the training dataset is composed of bags, where each one contains a set of instances and its goal is to teach a model to predict the bag label. A positive label is assigned to a bag if it contains at least one positive instance. MIL approaches have been successfully applied to computational histopathology for tasks such as tumor detection based on WSIs, reducing the time required to perform accurate annotations [114–117, 146, 147]. Some of these works use convolutional neural networks (CNNs) for the feature extraction process in each instance independently and then combine the instance-level information into one bag-level output. Methods that combine instance-level features are known as embedding-based, which require a later classification layer. In the case of [115], the bag level representation is achieved by the aggregation of the features through a simple batch global max-pooling (BGMP). Recent methods have proposed weighted-average embeddings, using instance-specific attention weights learned via a multi-layered perceptron projection or recurrent neural networks. In contrast, instance-based architectures combine instance-level predictions directly into the bag classification. In this vein, [114] obtained a tile-level feature representation through a CNN. These representations were then used in a recurrent neural network to integrate the information across the whole slide and report the final classification result to obtain a final slide-level diagnosis.

In most MIL-based papers, the WSIs employed have broad features that determine that a bag is positive. However, in this case, small cells (neutrophils) with features very similar to others in the tissue differentiate whether a bag is positive. Therefore, the typical MIL approach is not useful as the extracted activations are degrade and do not allow satisfactory classification.

4.2.2 Constrained CNNs

Constrained classification aims to guide the training of a CNN towards a solution that satisfies a given condition, which takes advantage of additional knowledge to the global labels. This learning paradigm has gained popularity on weakly-supervised scenarios (e.g. weakly supervised segmentation or MIL) since it allows to incorporate local information for improving the final task. Several works have tackled the problem of weakly-supervised segmentation by imposing constraints on deep CNNs [43–46]. In [43], the authors proposed a latent distribution and KL-divergence to constrain the output of a segmentation network. It is used in a semi-supervised setting to impose size constraints and image-level tags (i.e., force the presence or absence of given labels) on the regions of unlabeled images. Moreover, an L2 penalty term was proposed in [44] to impose equality constraints on the size of the target regions in the context of histopathology image segmentation which considerably improved the results. More recently, the authors showed in [45] that imposing inequality constraints on size directly in gradient-based optimization, also via an L2 penalty term, provided better accuracy and stability when few pixels of an image are labeled. Similarly, Zhou et al. embedded prior knowledge on the target size in the loss function by matching the probabilities of the empirical and predicted output distributions via the KL divergence. As directly minimizing this term by standard SGD is difficult, they proposed to optimize it by using stochastic primal-dual gradient [46]. While these works have helped to improve segmentation in a weakly-supervised setting, few studies focused on classification frameworks. In this work, by means of location constraints, we force the activations of the feature extractor to focus on those regions where neutrophils are localized. In this way, a reduced number of annotations can significantly improve the classification results.

4.3 Methodology

Here, we build an end-to-end MIL method as our baseline to perform image-to-image learning and prediction. The MIL formulation, based on CNNs, enables to detect neutrophils in WSIs and classify them into either histological remission or adverse outcome (UC activity). In Figure 4.2, the proposed framework is shown. In the following, we describe the problem formulation and each of the proposed components.

4.3.1 Problem formulation

In MIL tasks, the training dataset is composed of bags, where each bag contains a set of instances (patches). A positive label is assigned to a bag if it has at least one positive instance. The goal of MIL is to teach a model to predict the bag label.

We denote our training dataset by $\mathcal{S} = (X_k, Y_k)$ with $k = \{1, 2, 3, \ldots, N\}$, where X_k denotes the k-th input bag (WSI) and $Y_k \in 0, 1$ refers to the global label (ground truth label) assigned to the k-th input WSI. Here, $Y_k = 0$ refers to a WSI with remission and $Y_k = 1$ refers to ulcerative colitis activity. Note that we denote each individual bag or WSI as: $X_k = \{x_{k,1}, ..., x_{k,t}, x_{k,I_n}\}$, where $x_{k,t}$ is the t-th instance of the bag and I_n denotes the total number of patches or instances in a slide. The number of instances varies considerably between slides.

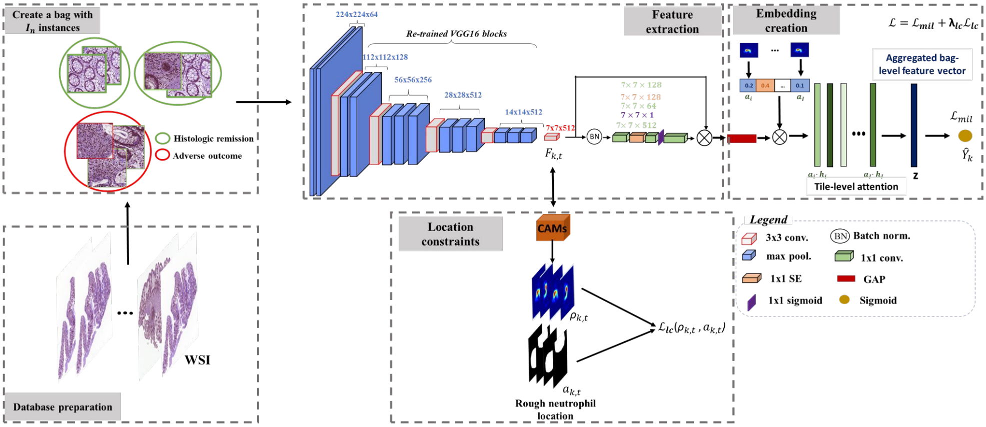


Figure 4.2: Pipeline showing the embedded-level approach for ulcerative colitis detection. By incorporating the proposed location constraints, we force the backbone to extract more significant features from each patch belonging to a given bag. After that, we classify the entire biopsy using an aggregated bag-level feature vector weighted by the proposed attention-embedding weights.

The loss function used to optimize the end-to-end MIL approach is the cross-entropy cost function:

$$\mathcal{L}_{mil} = \sum_{k}(I(Y_k = 1)log\hat{Y}_k + I(Y_k = 0)log(1 - \hat{Y}_k) \tag{4.1}$$

where $I(.)$ is an indicator function.

4.3.2 MIL backbone with location constraints

As will be shown in the experiment section, our baseline MIL formulation produces a decent result for the proposed task but still with room for improvement. One problem is that the positive instances predicted by the algorithm tend to outgrow the true regions with inflammation (UC activity) progressively. We propose using a neutrophil area constraint term to restrict the expansion of positive instances during training. We refer to our algorithm as location constrained MIL, abbreviated as LCMIL.

We denote our training set as $\mathcal{S} = (X_k, Y_k, A_k)$ with $k = \{1, 2, 3, \ldots, N\}$, where X_k denotes the k-th bag, $Y_k \in \{0, 1\}$ refers to the global label (ground truth label) assigned to the k-th input WSI and A_k specifies a rough estimation of the relative area in which the neutrophils are located within the image X_k. Being $a(i,j)_{k,t}$ the pixel (i,j) in the t-th patch from the bag k-th, $a(i,j)_{k,t} = 1$ if it corresponds to a pixel that is located around a neutrophil, whereas $a(i,j)_{k,t} = 0$, otherwise. Note that the rough annotations of neutrophil areas only are used for optimizing the parameters of the networ (θ) and not for the prediction phase.

A Global-aggregation layer is implemented to obtain an activation map representing the distribution of the features extracted from each of the instances belonging to a given bag. This layer summarizes the information from all spatial locations in the feature-embedded map $F_{k,t} \in \mathbb{R}^{H\times W\times C}$ (corresponding to the last volume of features extracted by the backbone) to one representative map $\rho \in \mathbb{R}^{H\times W}$. Note that $H \times W$ are the dimensions of the instances and C is the number of filters. Therefore, $\rho \in \mathbb{R}^{H\times W}$ is defined as follows:

$$\rho(i,j)_{k,t} = \frac{1}{C}\sum_{c\in C} F_{k,t}(i,j,c) \tag{4.2}$$

In this way, we have a representation of how the backbone attention is distributed over the instance surface. In order to have the same dimension as the input instances (224^2), a bilinear interpolation is performed to the activation map ρ. In the following step, ρ is transformed into $\rho_s = \phi(\rho)$, where ϕ is the sigmoid activation function. The aim of the sigmoid activation function is to range the map activation function into [0-1]. Then, we define an area constraint as the L_2 penalty:

$$\mathcal{L}_{lc} = \sum_{k,t} I(Y_k = 1 \textit{ and } a(ij)_{k,t} > 0)\ ((a_{k,t} - \phi(\rho_{k,t}))^2) \tag{4.3}$$

Naturally, the global loss function can be updated from Equation (4.1) to:

$$\mathcal{L} = \mathcal{L}_{mil} + \lambda_{lc}\mathcal{L}_{lc} \tag{4.4}$$

where $\lambda_{lc} \in \mathbb{R}^{+}$ weights the importance of the constraint during training.

4.3.3 MIL attention-embedding weights

After the feature extraction of each instance, we obtain a C-dimensional feature vector. The bag label predictor is in charge of aggregating the C-dimensional feature vectors $\{\mathbf{h}_t\}_{t \in I_n}$ into an embedding vector $Z_k \in \mathbb{R}^{1 \times C}$ representative of each bag. In the literature, there exist different simple aggregation functions such as batch global max-pooling (BGMP) or batch global average pooling (BGAP). However, these operators have a clear disadvantage. They are predefined and non-trainable. Other works use trainable aggregation functions [124]. However, in some situations, these attention weights have the same value for all instances in the bag, which is not suitable to determine a positive bag. This could be due to the complexity of the instance in some bags and the over-fitting tendency of neural networks. To solve this problem, we propose to use a weighted average of instances where weights are obtained from the representative maps $\rho_{k,t}$. Note that the weights of these maps are updated each epoch using the $\mathcal{L}_{lc}$ term. Additionally, the weights must sum to 1 to be invariant to the size of a bag.

Therefore, the embedded feature vector per bag is obtained as $Z_k = \sum_{t \in I_n} a_t \cdot \mathbf{h}_t$, where a_t is defined as:

$$a_t = \frac{exp\{\sum \rho(i,j)/S\}}{\sum_{I_n} exp\{\sum \rho(i,j)/S\}} \tag{4.5}$$

where $S = H \cdot W$.

This attention vector promotes variability between instances of a positive bag. If there is no activation corresponding to neutrophils in the map ($\rho_{k,t}$), the value of a_t will be low and therefore, the embedding features $\mathbf{h}_t$ will have smaller weight in the final prediction. In the case of a negative bag, the attention values will be very similar and all instances will contribute equally. The superiority of this aggregation function for neutrophil identification and HR prediction will be shown in Section 4.

4.4 Experiments and Results

4.4.1 Implementation

All the tested approaches were implemented using Tensorflow 2.3.1 with Python. Experiments were conducted on the NVIDIA DGXA100 system.

Table 4.2: Database description. Amount of whole-slide images (first row), number of patches (second row) and percentage of slides with PHRI>0, ulcerative colitis (third row).

	Training	**Validation**	**Test**
Number of WSI	84 (64,6%)	46 (35,4%)	100
patches	61.1 ± 54.2	58.2 ± 36.4	481.2 ± 292.1
PHRI score>0	51,1 %	39,15%	48%

1) **Dataset (PICASSO-MIL)**:We analyzed 230 colorectal biopsies from UC patients enrolled in a prospective international multicenter study to evaluate the proposed deep-learning methodology. Note that the slides belong to 7 different hospitals [148]. To process the large WSIs, these were downsampled to 20x resolution, divided into patches of size 512x512x3 with a 50% overlap among them. Aiming at pre-processing the biopsies and reducing the noisy patches, a mask indicating the presence of tissue in the patches was obtained by applying the Otsu threshold method over the magenta channel. Subsequently, the patches with less than 20% of tissue were excluded from the database. Using this database, we carried out a patient-level data partitioning procedure to separate training and validation sets, aiming to avoid overestimating the system's performance and ensuring its ability to generalize. Additionally, 100 non-annotated images at pixel-level were used to test the framework, see Table 4.2. During training, the human pathologists (with more than 35-year clinical experience) make two image-level annotations for each WSI, indicating each image as HR or UC activity depending on PHRI, and roughly estimating which areas of the image show neutrophils and inflammation. Only the bag label is necessary to evaluate the proposed method.

2) **Model parameters**: The MIL loss is known to be hard to train and special care is required for choosing training hyperparameters. To reduce fluctuations in optimizing the MIL loss, all training data are used in each iteration (the minibatch size is equal to the size of the training set). The network is trained with stochastic gradient descent (SGD) optimizer and a fixed learning rate of 0.01. The number of epochs was adapted in function of the experiment performed.

3) **Backbone network**: We choose the SeaNet (with VGG16) proposed in [50] as the CNN architecture of our framework since it demonstrated the improvement over standard methods in histological imaging. This framework is composed of VGG16 as a feature extractor and a squeeze and excitation attention network. In addition, we performed fine-tuning of this model, as it

had previously been trained with histological images, in a different task, the detection of skin tumors.

4) **Evaluation**: The quantitative comparison of the different methodologies was handled by means of different figures of merit, such as sensitivity (SN), specificity (SPC), positive predictive value (PPV), false-positive rate (FPR) negative predictive value (NPV), F1-score (F1S), accuracy (ACC) and area under the ROC Curve (AUC).

4.4.2 Ablation experiments

In the following, we provide comprehensive ablation experiments to validate several elements of our model (LCMIL), and motivate the choice of the values employed in our formulation, as well as our experimental setting.

1) **Weight of location constraint loss**: The weight of the constraint loss is crucial for LCMIL since it directly decides the strength of constraints. Strong constraints may make the network unable to converge, while weak constraints have little help with learning. Therefore, we optimized the proposed formulation with the location constraint term in Eq. 4. Using the training setting previously described, we cross-validated different values of λ_{ac} = $\{0.1, 0.1, 1, 1, 5\}$. Additionally, we tried two loss functions, $\mathcal{L}_1$ and $\mathcal{L}_2$, to check for differences. We obtained bag-level ACC from the validation subset using the ACC on validation subset as early stopping criteria. Results are presented in Figure 4.3.

These results show that the inclusion of the $\mathcal{L}_{lc}$ term improves the performance at bag level. Nevertheless, using a too large slope once the performance is satisfied can lead to a worsening of the results. Thus, we selected $\lambda_{lc} = 1$, which led to the best results at bag level in the validation cohort.

Additionally, we want to get a more intuitive view of how the proposed methodology location constraint term influences the extraction of discriminative features. For that purpose, we depict the feature representation of the embedding space produced by the encoder networks of MIL without $\mathcal{L}_{lc}$ and the proposed encoder on the instance-level labeled validation. Concretely, we obtained the class activation maps for regions of a bag where neutrophils are found (cryptal lumen, cryptal epithelium, lamina propia and surface epithelium). In Figure 4.4, the annotations made by the pathologists, the activation maps obtained by a MIL module without $\mathcal{L}_{lc}$ and the proposed method are compared.

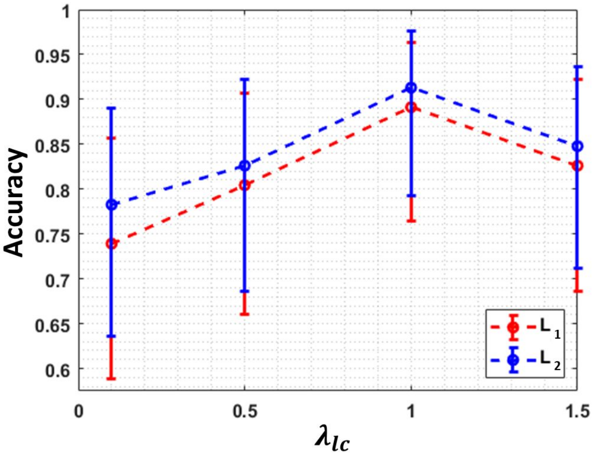


Figure 4.3: Ablation studies on MIL formulation. Hyperparameters study for λ_{ac} are performed for bag-level accuracy on validation set. Confidence intervals are shown at 95%.

The MIL without location constraint module does not focus its attention on the areas where neutrophils are located by the pathologist but on other cells found in the tissue. Note that neutrophils are very similar to other cells found in the tissue, such as eosinophils, macrophages, etc., but in this case, they do not determine that a patient has active ulcerative colitis. This is why the specificity of this model is very low. In contrast, the inclusion of the location constraints module forces the network to focus its attention on the real determining cells, the neutrophils. In this way, we can therefore obtain precise instance-level maps for unannotated images that allow us to detect the neutrophils.

2) **Attention weights for bag classification**: Using the best configuration reached for the λ_{lc} term, we optimized the embedded feature vector per bag, see Table 4.3. This Table compares the best-known methodologies for constructing the embedded vector (BGAP, BGMP and MIL-Attention) versus the proposed method. Since the features that discriminate a positive bag are relatively small compared to the dimension of the different instances, in this case, the BGMP layer improves the results of the BGAP and MIL-Attention layers. However, the proposed aggregation method outperforms all previous methods.

To compare the distribution of the attention weights of [124] with those proposed here, we show the histogram of these values in a positive bag, see Figure 4.5. In this case, the bag comprises 80 instances, of which only 15% are positive, i.e., contain neutrophil structures. In Figure 4.5

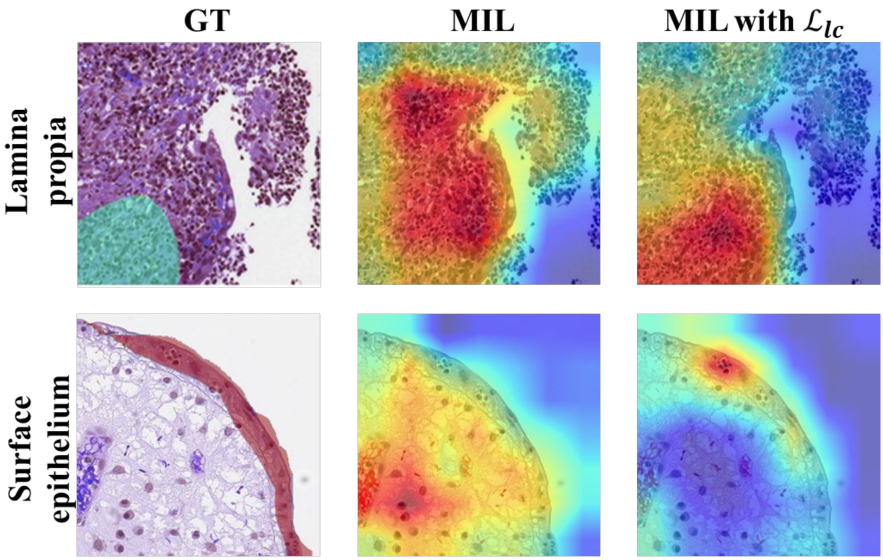


Figure 4.4: Class activation maps (CAMs) of some regions where neuthophils are found. First column: original images with pathologist annotation (green and red annotations); Second column: CAMs obtained using the normal MIL model. Third column: CAMs using the proposed location constraints.

(b), attention proposed in [124], the different values of weights have similar probabilities. Therefore, no discriminatory weighting is performed to separate negative and positive instances. However, with the proposed method, most instances (around 60) have a low weight, which would belong to the instances without neutrophils. The remaining weights are spread across instances with neutrophils, with higher weights assigned to those with more significant features. Therefore, the proposed attention-based MIL allows to assign more discriminate weights to instances within a bag and hence the final representation of the bag is highly informative for the bag-level classifier.

Table 4.3: Comparison of the different attention embedding weights on the validation set. BGAP: batch global average pooling, BGMP: batch global max-pooling, LCMIL: neutrophil constrained weak supervision (proposed). Note that in all cases the location constraint proposed is integrated into the backbone.

	BGAP	**BGMP**	**Attention [124]**	**LCMIL**
SN	0.9643	0.9643	0.8889	**0.9643**
SPC	0.6667	0.7778	0.7778	**0.8333**
PPV	0.8182	0.8710	0.8571	**0.9000**
NPV	0.9231	0.9333	0.8235	**0.9375**
F1S	0.8852	0.9153	0.8727	**0.9310**
ACC	0.8478	0.8913	0.8444	**0.9130**
AUC	0.8155	0.8710	0.8333	**0.8988**

4.4.3 Comparison to the literature

To compare the proposed method with the MIL baselines, a comparative analysis of the test cohort is performed in this section, see Table 4.4. For this purpose, we included the current state-of-the-art deep MIL models, the attention based pooling operator (ABMIL) [124], non-local attention based pooling operator (DSMIL) [146], single-attention-branch (CLAM-SB) [147] and recurrent neural network (RNN) based aggregation (MIL-RNN) [114].

The figures of merit are obtained at the biopsy label because only these labels are available in the test set. In general, the specificity of the MIL baseline models drops considerably. The best state-of-the-art model (CLAM-SB) achieves a specificity of 0.8033 compared to 0.9615 obtained by the proposed model (LCMIL). State-of-the-art models are not able to discriminate between neutrophils and other tissue cells and therefore are not optimal for predicting diseases such as ulcerative colitis, which are caused by very precise histological patterns. Under our proposed formulation (LCMIL), the model can detect neutrophils at the instance level and, therefore, predicts ulcerative colitis with a good performance. Obviously, there is a high consistency between the fine annotation area and CAMs obtained in Figure 4.4, illustrating great interpretability and attention visualization of the proposed framework. Therefore, with a small volume of training annotations, the model can improve the accuracy of the best baseline MIL approach by almost 10%.

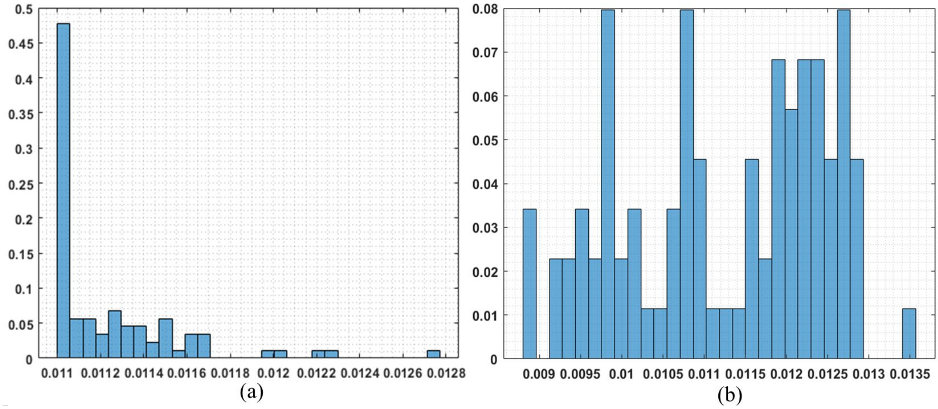


Figure 4.5: Distribution of embedding weights across the instances that comprise a WSI. (a) Proposed attention embeddings. (b) Attention weights proposed in [124].

Table 4.4: Comparison of the different baseline frameworks in the test cohort. Note that for the test cohort only the global bag label are available.

	ABMIL	DSMIL	CLAM-SB	MIL-RNN	LCMIL
SN	0.9583	0.8293	0.9302	0.8667	**0.9583**
SPC	0.6923	0.7288	0.8033	0.7797	**0.9615**
PPV	0.7419	0.6800	0.7692	0.7500	**0.9583**
NPV	0.9473	0.8600	0.9423	0.8846	**0.9615**
F1S	0.8393	0.7473	0.8421	0.8041	**0.9583**
ACC	0.8200	0.7700	0.8558	0.8173	**0.9600**
AUC	0.8253	0.7546	0.8321	0.8009	**0.9599**

4.5 Conclusion

Whole-slide images (WSI) have shown applicability to developing computer vision models, but few studies have approached the use of deep learning models to detect ulcerative colitis (UC). In this work, we propose an location constraint framework able to perform histological remission prediction using WSIs of patients with UC. Our framework comprises a feature extraction backbone with an attention module to refine the patch-level features and a MIL approach to predict the UC activity in each bag. We introduce a location constraint module that forces the feature extractor to focus on the most significant patterns in the patches that form a bag. The biopsy classification comes from the bag-level feature vector that the attention embedding has ponderated. This approach reaches a test accuracy of 0.9600 in a more significant subset than the training set, which shows that the extra pixel-level annotation gives crucial information to the algorithm.

Future research lines need to focus on detecting neutrophils in the different biopsy regions and grading PHRI accordingly, not being limited to the histological activity or remission prediction. The location constraint approach also promises applicability to other pathologists in which histological analysis is based on identifying single cells.

Chapter 5

Labeling confidence for uncertainty-aware histology image classification

The content of this chapter corresponds to the author version of the following published paper: Del Amor, R., Silva-Rodríguez, J. & Naranjo, V. Labeling confidence for uncertainty-aware histology image classification. Computerized Medical Imaging and Graphics, 102231 (2023).

Contents

5.1 Introduction **87**
5.2 Related work **88**
5.2.1 Skin WSIs 88
5.2.2 Uncertainty estimation 89
5.3 Methods **90**
5.3.1 Labeling uncertainty 92
5.3.2 Dual-branch uncertainty calibration 93
5.4 Experimental setting **95**
5.4.1 Dataset 95
5.4.2 ROI extraction 96
5.4.3 Implementation details 97
5.4.4 Evaluation metrics 97
5.5 Results **97**
5.5.1 Comparison to the literature 97
5.5.2 Ablation studies 101
5.6 Conclusion **104**

Labeling confidence for uncertainty-aware histology image classification

Rocío del Amor[1], Julio Silva-Rodríguez[2] and Valery Naranjo[1]

[1]Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano, Universitat Politècnica de València, 46022, Valencia, Spain

[2]ÉTS Montréal,
Montréal, Québec, Canada

Abstract

Deep learning-based models applied to digital pathology require large, curated datasets with high-quality (HQ) annotations to perform correctly. In many cases, recruiting expert pathologists to annotate large databases is not feasible, and it is necessary to collect additional labeled data with varying label qualities, e.g., pathologists-in-training (henceforth, non-expert annotators). Learning from datasets with noisy labels is more challenging in medical applications since medical imaging datasets tend to have instance-dependent noise and suffer from high inter/intra-observer variability. In this paper, we design an uncertainty-driven labeling strategy with which we generate soft labels from 10 non-expert annotators for multi-class skin cancer classification. Based on this soft annotation, we propose an uncertainty estimation-based framework to handle these noisy labels. This framework is based on a novel formulation using a dual-branch min-max entropy calibration to penalize inexact labels during the training. Comprehensive experiments demonstrate the promising performance of our labeling strategy. Results show a consistent improvement by using soft labels with standard cross-entropy loss during training ($\sim$ 4.0% F1-score) and increases when calibrating the model with the proposed min-max entropy calibration ($\sim$ 6.6% F1-score). These improvements are produced at negligible cost, both in terms of annotation and calculation.

5.1 Introduction

Digital pathology research has experienced significant growth in recent years thanks to the advent of novel computer vision techniques based on deep learning [34]. The deployment of convolutional neural networks (CNNs) has allowed the automatic identification of new biomarkers and innovative features in the whole slide images (WSIs) that support the diagnostic process. In particular, these techniques have shown promising results for computer-aided diagnosis on different applications such as prostate [149], breast [150] and skin cancer detection [50], tissue segmentation [151], or mitosis detection [152], among others. Nevertheless, deep learning models require large and curated datasets with high-quality (HQ) annotations to perform properly. In the case of digital pathology, a popular choice is the use of weakly supervised strategies with WSI-level annotations. In the multi-class scenario, an expert pathologist assigns a unique label to the whole biopsy based on diagnostic or prognostic features. Then, deep learning models are trained using multiple instance learning (MIL) to automatically solve the task at hand. However, this pipeline does not consider real-world limitations and noise sources inherent to the annotation process, which may hinder the performance of the model. These limitations are accentuated in some applications requiring a high level of expertise, such as several skin neoplasm diagnosis (i.e., cutaneous spindle cell neoplasms, one of the most challenging skin neoplasms not studied in previous studies [153]). In many cases, recruiting expert pathologists to annotate large databases is not feasible. Unfortunately, without sufficient labels, the data-hungry learning-based methods often struggle with overfitting, leading to inferior performance [47]. To alleviate this issue, collecting additional labeled data with varying label qualities, e.g., pathologists-in-training (henceforth, non-expert annotators) or using machine-generated labels is a common practice. However, directly introducing data with low-quality (LQ) noisy labels may confuse the network training, which easily leads to performance degradation [48, 49]. Therefore, how to effectively and robustly exploit the additional information in plentiful LQ noisy labeled data is crucial to the medical image analysis community.

Learning from noisy labels is a widely recognized challenge in classical image recognition. Several efforts have been made to mitigate the negative impact of LQ labels in medical image analysis [49, 154–156]. However, this is still an under-explored area, as existing literature on learning with noisy labels lacks a clear distinction of applicable scenarios, leading to ambiguous benchmarks. Some approaches [155, 156] assumed mixed data from multiple sources, i.e., set-HQ and set-LQ labels are indiscriminate. In contrast, other techniques [49, 154] were developed for a scenario where experts label a small data set, making LQ and high-quality (HQ) labels separated. A main body of literature exploits multiple annotators in a crowdsourcing scenario, to extract the underlying noise-free label distribution. Nevertheless, gathering multiple annotators in the medical context may be unrealistic. The high level of expertise required, as well as the time-consuming nature of such annotation, is a barrier to the implementation of these methods in real-world applications. These findings highlight the need for developing uncertainty-aware pipelines to address the inherent uncertainty in the annotation process, which may not require from multiple label sources.

Based on these observations, we propose a novel uncertainty-driven labeling strategy for histology skin cancer classification. The key contributions of our work can be summarized as follows:

- A single-annotator uncertainty-aware labeling strategy with which we generate soft labels from 10 non-expert annotators for multi-class skin cancer classification that quantify uncertainty in the annotations.
- Based on these annotations, we present an extensive study for the use of soft label model calibration compared to the ground truth, labeled by an expert pathologist.
- In addition, we propose a novel formulation based on dual-branch entropy calibration (DBEC) to calibrate both, overconfident outputs and uncertain soft labels, during training.
- Comprehensive experiments demonstrate the promising performance of our labeling strategy. By incorporating uncertainty during labeling we found average improvements of nearly $\sim 4.0\%$ in averaged F1-score using the baseline methods, which increases up to $\sim 6.6\%$ using the proposed dual-branch calibration.

5.2 Related work

5.2.1 Skin WSIs

According to the World Health Organization, nearly one in three diagnosed cancers worldwide is a skin cancer [79]. Different techniques, such as dermatoscopy, wood lamp, CT scan and histopathology, are utilized for the diagnosis of skin diseases. However, the gold standard for skin cancer detection is histological image analysis. Traditionally, histological slides would be viewed with a light microscope. However, digitization has created opportunities for automated analysis using WSI. Applying deep-learning models to computer vision problems shows excellent potential in skin cancer detection. Most research was based on the analysis of dermoscopic images [96–100, 157, 158] and few studies have focused on the analysis of WSI [50, 101, 102, 104, 150, 159]. In this vein, MIL approaches have been successfully applied to Basal carcinoma (BCC) [150] or melanoma [50], reducing the time required to perform precise annotations. However, many types of skin cancer have not yet been explored. These include cutaneous spindle cell neoplasms (CSC), predominantly composed of spindle-shaped neoplastic cells arranged in sheets and fascicles [160]. These lesions are relatively common. For example, cutaneous squamous cell carcinoma is the second most common epidermal cancer representing 20 % to 50% of skin cancers [161] and spindle cell melanoma contributes 3% to 14% of all melanoma cases [162]. CSC neoplasms are challenging to diagnose due to the considerable morphological overlap between the different tumor types that make up this group [153], which poses a particular problem for less experienced pathologists. This hampers an accurate diagnosis and the application of effective clinical treatment [163] in neoplasms in which early detection and appropriate treatment are essential for a good prognosis in malignant cases. Despite the complexity of these neoplasms, they had not been previously studied in the literature. Therefore, the main objective of this paper is to classify, under a MIL-based approach, the seven types of fusocellular skin neoplasms identified by expert pathologists as the most challenging: leiomyomas (lm), leiomyosarcomas (lms), dermatofibromas (df), dermatofibrosarcomas (dfs), spindle cell melanomas (mfc), fibroxanthomas (fxa) and squamous cell carcinoma (cef).

5.2.2 Uncertainty estimation

Uncertainty estimation methods are expected to improve the understanding and quality of deep learning models to enhance their generalization during inference. These methods have an outstanding interest in medical applications due to the high expertise required to obtain quality labels, the variability in acquisition systems and noise present in many databases [164], and the known inter-annotator variability in different medical applications [165, 166]. For these reasons, training uncertainty-aware models is key to the success of diagnostic support systems in medical applications. An uncertainty-aware deep learning model training usually covers two steps: uncertainty quantification and model calibration. Uncertainty quantification aims to assess the prior probability of error for certain samples during training. From the perspective of noisy labels, a main core of previous literature use multiple annotators in crowd-sourcing scenarios to quantify inter-observer agreement for each sample [167–170]. Thus, crowd-sourcing methods aim to predict the underlying noise-free label distribution by simultaneously training annotator-specific projections over the feature space [167–171]. Other solutions focus on prior task-specific knowledge such as avoiding overconfident outputs on neural networks [172] or leveraging high confidence on non-informative regions [173]. Other uncertainty quantification approaches focus on sample noise estimation, which may raise from image quality, feature extraction, or out-of-distribution domains. Previous literature in this regard use a trained student model to study the confidence of the model via Monte Carlo dropout with image augmentations [159, 170, 174], curriculum learning [175], or co-teaching [176, 177]. After uncertainty estimation, deep learning models are calibrated to overcome the limitations detected in the training samples. Some approaches include sample weighting based on divergence observed by the Student-based methods [170], or calibrating the output of the network based on label smoothing [178] and entropy regularization [172, 179, 180].

In this paper, we focus on label-noise calibration, and we study the feasibility of estimating uncertainty from single annotator labels. Contrary to much of the previous literature, we study the case in which multiple annotators are not available. To this end, we define a soft label-based annotation protocol. Then, we propose a dual-branch criterion for calibrating the trained neural network based on entropy regularization. The underlying idea is two-fold: (i) penalizing overconfident predictions on high-certain samples, and (ii) forcing the network to produce confident outputs on uncertain cases, to overcome the limitations of the noisy labels based on the features of each sample. Note that although we trained 10 models, one for each non-expert to validate the proposed methodology, these models are independent since only the labels of a single annotator are used to train the algorithm each time.

5.3 Methods

An overview of our proposed method is depicted in Figure 5.1. In the following, we describe the problem formulation and each of the proposed components.

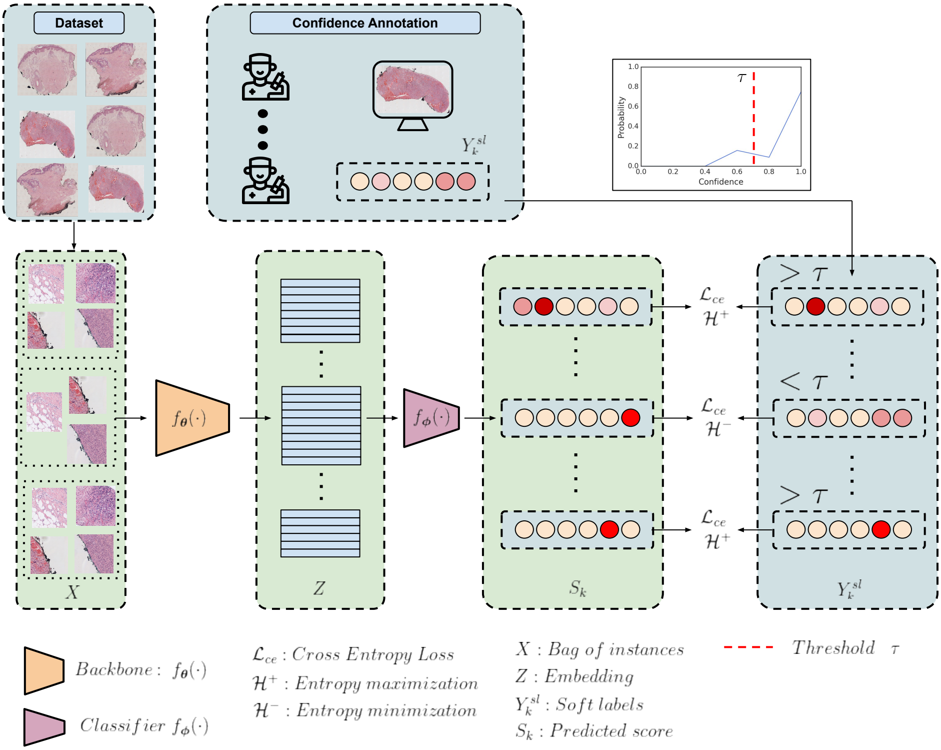


Figure 5.1: Method overview. In this work, we address weakly supervised histology image classification on skin WSIs by quantifying the uncertainty of the individual annotators during labeling. Concretely, we train an embedding-based Multiple Instance Learning (MIL) model to predict up to six different categories using standard cross-entropy loss. We propose to quantify annotator-specific uncertainty by following a soft labels annotations protocol, such that $Y_k^{sl} = [0, 1]$, and $\sum_k Y_k^{sl} = 1$. In this fashion, our model captures information regarding inter-category dependencies and avoids over-fitting to uncertain, noisy annotations. Then, we propose a dual-branch min-max uncertainty calibration (DBEC) based on the annotated soft labels. Based on uncertainty calibration using Shannon entropy regularization (see Eq. 5.3), we propose to (i) maximize the entropy on high-confidence labeled samples, by entropy maximization (H^+), and (ii) to minimize the entropy on samples labeled with low-confidence (H^-). Thus, entropy minimization encourages the network to produce confident outputs on uncertain cases, based on the features of the sample, and thus diminishing noise propagation. A threshold τ is empirically fixed to differentiate low and high certain labels, and the dual-branch min-max uncertainty is combined with cross-entropy loss (see Eq. 5.5). Circles in bag-level predictions and references indicate soft-max scores. The more intense the color, the higher the score.

Problem Formulation Under the paradigm of Multiple Instance Learning (MIL), instances are grouped in bags of instances $X = \{x_n\}_{n=1}^{N}$ that exhibit neither dependency nor ordering among them, and its number N is arbitrary for each bag. In the multi-class scenario, each bag is a member of one of K mutually exclusive classes, such that $Y_k \in \{0, 1\}$. Note that, in contrast to other MIL formulations, the individual instances do not have an associated label, but rather the label of the bag is determined by the combination of features of the different instances.

Embedding-based MIL In this work, we aim to train a model capable of predicting bag-level labels using a combination of features extracted at the instance level. This learning strategy falls under the embedding-based MIL paradigm[1]. Let us denote a neural network model, $f(\cdot) : \mathcal{X} \rightarrow \mathcal{Z}$, parameterized by , which projects instances $x \in \mathcal{X}$ to a lower dimensional manifold $\in \mathcal{Z} \subset \mathbb{R}^d$, with d the embedding dimension. Then, we define an aggregation, $f_a(\cdot)$, which is in charge of combining the instance-level projections into a global embedding, Z. In particular, we use a global-average pooling along instances, such that: $Z = \frac{1}{N} \sum_n \{f(x_n)\}_{n=1}^{N}$. Finally, a neural network classifier, $f(\cdot) : \mathcal{Z} \rightarrow \mathcal{S}$, is in charge of predicting softmax bag-level class scores, S_k, such that $S_k \in [0, 1]$. The optimization of the model parameters and is driven by the minimization of standard categorical cross-entropy loss between the reference labels and predicted scores such that:

$$\mathcal{L}_{ce} = -\frac{1}{K} \sum_{k=1}^{K} Y_k \cdot log(S_k) \tag{5.1}$$

5.3.1 *Labeling uncertainty*

Uncertainty estimation methods assume that different noise sources are present in the dataset, both in image noise and inter and intra- annotator variability. The objective is to calibrate the trained model to account for quantified uncertainties. Regarding inter-annotator variability, a large body of literature quantifies this uncertainty by obtaining labels from multiple annotators. However, obtaining multiple annotators may not be possible in specific scenarios requiring a high level of specialization or covering proprietary solutions, such as medical applications. To overcome this limitation, we propose an annotator-level uncertainty quantification by annotating the confidence associated with each sample in the form of soft labels. To this

[1] Based on the denomination proposed in [181]

end, we differentiate between the labeled samples using hard labels (HL), Y_k^{hl}, and soft labels (SL), Y_k^{sl}. As previously described, hard labels assign a discrete value for each label such that $Y_k^{hl} \in \{0, 1\}$, where $Y_k = 1$ indicates that the corresponding sample belongs to the class k. It is worth mentioning that, in the multi-class scenario, categories are considered mutually excluded, and only one tag is given to each sample. Nevertheless, this labeling strategy fails to capture the certainty of the annotator for each sample. To gather this information, we propose to use soft labels, such that $Y_k^{sl} \in [0, 1]$. Note that in this case, Y_k is a continuous value that corresponds to the probability that the annotator assigns to each class, such that: $\sum_k Y_k^{sl} = 1$. For instance, in a case with high uncertainty, the annotator might assign the following labels: $Y^{sl} = [0, 0, 0, 0, 0.9, 0.1, 0]$, whereas in a uncertain case, the total probability might be more distributed among categories: $Y^{sl} = [0.2, 0.2, 0, 0, 0.6, 0, 0]$. Then, the MIL classification model previously described is trained using standard cross-entropy loss in Eq. 5.1 using soft annotation labels. We believe that, in this fashion, the model might capture information regarding inter-category dependencies and avoid over-fitting to uncertain cases, as supported in the experimental stage of the present work.

5.3.2 Dual-branch uncertainty calibration

The aforementioned soft-labeling strategy can differentiate between high-certain and uncertain labels provided by the annotator. Still, using standard cross-entropy might produce ill-calibrated models. These limitations include reaching trivial solutions by producing overconfident outputs from high-certain samples or trivial, uniform outputs on low-certainty samples. In addition, we want to consider that samples labeled with low confidence might belong to a class other than the one most likely to be noted. To this end, we propose calibrating the model during training to deal differently with both types of samples in a dual-branch fashion.

Shannon entropy for confidence regularization One of the main approaches to calibrating neural networks is using an auxiliary term to regulate the output probabilities. Originally developed to reduce overconfident predictions, which are produced by training models using cross-entropy and hard labels, one of the main approaches lies in forcing the output distribution to approximate a uniform distribution [172, 178]. To this end, the neural network is trained to minimize the Kullback − Leibler (KL) distance, $D_{KL}(p||u) = H(p, u) - H(p)$ between an output distribution, p and an uniform distribution, u. Note that $H(p, u)$ indicates the cross-entropy between both distributions, and $H(p) = H(p,p)$ is the Shannon entropy or self-entropy, such that $H(p) = -\frac{1}{K}\sum_k p_k \cdot log(p_k)$. It is straightforward to see that, in the case of a target uniform distribution, minimizing the KL distance is equivalent to maximizing the Shannon entropy of the output distribution.

$$D_{KL}(p||q) = H(p,q) - H(p) =^c -H(p) \tag{5.2}$$

where $=^c$ indicates equality up to an additive constant.

Thus, standard model calibration using Shannon entropy includes a regularization term to the standard cross-entropy loss weighted by an hyper-parameter $\beta > 0$, such that:

$$\mathcal{L} = \mathcal{L}_{ce} - \beta H(p) \tag{5.3}$$

Dual-branch min-max entropy calibration Inspired by previous literature on model calibration, we propose to use the Shannon entropy regularization in a dual-branch fashion. First, we want the model to avoid overconfident outputs on high-certainty labeled samples, similarly to Eq. 5.3. Secondly, we aim to calibrate the model to assign a confident category to each sample, even though the annotator might have high uncertainty in the label. For the latter, we draw on Shannon entropy minimization, which encourages the output scores to differ from the uniform distribution (see Eq. 5.2). It is worth mentioning that, in the case of minimum entropy, the output scores tend to produce hard labels. Thus, we hypothesize that the model may be able to overcome the potential noise from the uncertain labels, and produce more accurate predictions based on the features of the sample. This formulation is inspired by the semi-supervised learning literature, in which entropy maximization is used as a proxy to learn from unlabeled samples [182]. From now on, and for simplicity in the context of loss functions, we refer to the entropy-maximization criteria $-H(p)$ as H^+, and the opposite minimization term as $H^- = H(p)$.

Thus, we propose a dual-branch optimization criterion to independently calibrate low and high-certainty labeled samples, using the bag-level predicted scores, S_k, such that:

$$\mathcal{L}_H = \begin{cases} H^+(S_k), & \text{if } \max_k Y_k^{sl} > \tau \\ H^-(S_k), & \text{otherwise} \end{cases} \tag{5.4}$$

where τ is an empirically-fixed threshold that divides the input samples based on its certainty, quantified by the confidence of the predominant category per sample, $\max_k Y_k^{sl}$.

Since using entropy calibration alone may yield trivial results [183], the MIL model is trained with annotated soft labels, Y_k^{sl}, and the dual-branch entropy calibration, using the overall following loss function:

$$\mathcal{L} = \alpha^{+/-}\mathcal{L}_{ce} + \beta^{+/-}\mathcal{L}_H \tag{5.5}$$

Note that $\mathcal{L}_H$ is the cross entropy loss at bag level in Eq. 5.1, and $\mathcal{L}_H$ refers to the dual-branch calibration presented in Eq. 5.4, and $\alpha^{+/-}$ and $\beta+/-$ are disentangled in two terms, one for high-certainty labeled samples (α^+, β^+), and other for the opposite case (α^-, β^-). It is worth mentioning that the values of threshold value τ in Eq. 5.4 as well as the relative weight of the min-max entropy duality, β^+ and $\beta-$, and cross-entropy loss, α^+ and α^-, are hyperparameters empirically optimized during the experimental stage. Hereafter, we refer to this dual-branch min-max entropy calibration term as DBEC.

5.4 Experimental setting

5.4.1 Dataset

To validate the proposed approach, we use the *AI4SKINV1* database. This database comprises two private databases (DSV and DSG) from the University Clinic Hospital of Valencia (Spain) and San Cecilio University Hospital in Granada (Spain). DSV and DSG are composed of histopathological skin images from different body areas that contain cutaneous spindle cell (CSC) neoplasms, i.e, leiomyomas (lm), leiomyosarcomas (lms), dermatofibromas (df), dermatofibrosarcomas (dfs), spindle cell melanomas (mfc), fibroxanthomas (fxa) and squamous cell carcinoma (cef). Each database (DSV and DSG) comprises 180 and 91 different patients who signed the pertinent informed consent. Two expert pathologists established the WSI-level label of the whole database, 271 images. A summary of the database description is presented in Table 5.1.

Regarding the non-experts labeling, an annotation protocol was designed to ensure that 106 WSIs were annotated by all non-expert annotators (dense set).

Table 5.1: Database distribution. DSV: database from Valencia; DSG: database from Granada. Lm:leiomyomas; lms: leiomyosarcomas; df:dermatofibromas; dfs: dermatofibrosarcomas; fxa: fibroxanthomas; spindle cell melanomas; cef: squamous cell carcinoma.

	lm	**lms**	**df**	**dfs**	**mfc**	**fxa**	**cef**	**Total**
DSV	28	19	52	11	32	28	10	180
DSG	27	9	16	7	6	26	-	91
Total	55	28	68	28	38	44	10	271

In contrast, the rest were only annotated by some non-expert pathologists (non-dense set). It is worth mentioning that the use of a dense set allows us to establish data-balanced comparisons between annotators, without requiring everyone to annotate the entire data set, with the burden that this process entails. Table 5.2 shows images used by each non-expert annotator for training, validation and testing of the models. To establish fair comparisons the validation and test images belonged to the dense set. Note that the images were annotated following the soft strategy proposed in Sec. 5.3.1 [2].

To process the large WSIs, these were downsampled to 10x resolution and divided into patches of size 512x512x3 with a 50% overlap. Aiming at preprocessing the biopsies and reducing the noisy patches, a mask indicating the presence of tissue in the patches was obtained by applying the Otsu threshold method over the magenta channel. Subsequently, the patches with less than 20% of tissue were excluded from the database.

Table 5.2: Number of images used for training, validation and testing the models of each non-expert annotator (ten in total). Note that for the validation and test set the same samples labeled by all non-experts were used.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Tain	148	142	151	143	154	145	155	149	152	150
Val	26									
Test	54									

5.4.2 ROI extraction

To select the instances with tumor from the WSI to train and validate the proposed approach, we extend the model proposed in [36] for the six neoplasms under study. This method was based on a teacher-model paradigm to increase the annotated database while avoiding manual annotations. In this vein, this approach enhances the detection of tumor regions in WSI using pseudolabels

[2] The soft labels will be available on request.

from non-labeled data. As the output of this section, we obtain the patches with tumor lesions used as input for the MIL-based model.

5.4.3 Implementation details

The proposed methods were trained using the different train subsets for each non-expert annotator (10 in total), see Table 5.2. The backbone $f(\cdot)$ used was a VGG16 [184] pre-trained on Imagenet [185], using patches resized to 224×224 images. Models were trained during 120 epochs with a batch size of 1 whole slide image, using a learning rate of $\eta = 1 \cdot 10^{-3}$ with SGD optimizer. The model performance was continuously monitored on the validation subset, and early stopping was applied to keep the model with the best accuracy on this subset. The proposed uncertainty calibration DBEC in Eq. 5.5 was trained similarly, but the learning rate was exponentially decreased in the last 20 epochs to ensure stability. In this case, early stopping was not applied since the calibration moved predictions away from the domain of the training labels. Hyperparameters were fixed empirically such that: $\alpha^{+} = 1$, $\beta^{+} = 0.1$, $\alpha^{-} = 0.1$, $\beta^{-} = 1$, and $\tau = 0.7$. For the motivation of these values, we refer the reader to the ablation experiments. All the validated experiments were implemented using Pytorch version 1.9.1 and Python 3.7. Experiments were conducted on the NVIDIA DGXA100 system. The code is publicly available on `https://github.com/cvblab/Labeling_Uncertainty`.

5.4.4 Evaluation metrics

In order o evaluate the performance of the proposed approaches regarding previous literature, we use standard metrics for multi-class classification. In particular, we obtain accuracy (ACC) and macro-averaged F1-score. It is worth mentioning that, although explicitly mentioned, metrics are obtained using as reference the ground truth, labeled by the expert pathologists, Y_k.

5.5 Results

5.5.1 Comparison to the literature

In this subsection, we study the obtained results by the proposed methods, concerning previous literature. We also carried out a detailed study of the success cases and limitations encountered, by means of a detailed study of the annotations made by the in-training pathologists.

Quantitative evaluation The quantitative results obtained training the model using expert labels, and non-expert labels using hard labels (HL), annotated soft labels (SL), and the proposed dual-branch entropy calibration (DBEC) on the respective test subset of each non-expert annotator are depicted in Table 5.3. Results obtained using annotated soft labels from non-expert pathologists reach an average F1-score of 0.364, which shows an improvement of $\sim$ 4.0% compared to hard labels by simply training the model using standard cross-entropy loss. This fact demonstrates that the annotation protocol developed in the paper is optimal for model training when expert labels are not available. Once our proposed dual-branch entropy calibration (DBEC, see Eq. 5.4) is incorporated during training, results achieve an average F1-score of 0.389. In addition, some noteworthy improvements can be observed for some non-expert annotators. For example, annotators 1, 2, and 8 show improvements of $\sim$ 13.1%, $\sim$ 21.5% and $\sim$ 13.2%, respectively. Although the results obtained are still far from those obtained using the ground truth from the expert pathologists, the models obtained bridge the gap, going from a difference of $\sim$ 25% to $\sim$ 18% regarding F1-score. Furthermore, this paper is the first study to address the multi-class problem of spindle cell neoplasms. While previous studies focus on binary problems to identify benignity or malignity of neoplasms [36], in this study we try to identify the distinct neoplasms that have considerable morphological overlap between them. Therefore, the results obtained in this paper establish a benchmark for the comparison of further models.

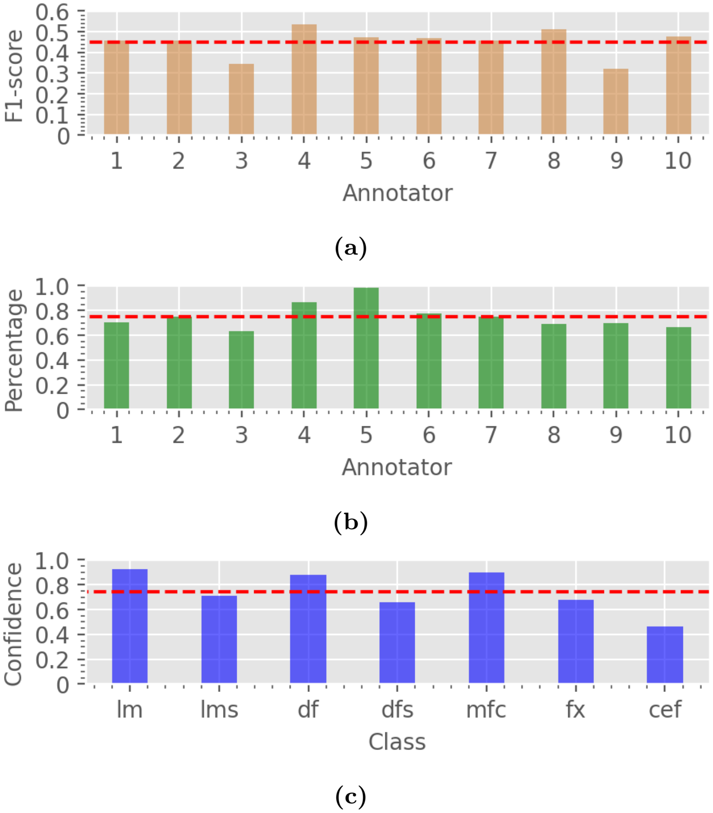
In-depth results analysis Although, as discussed above, the methodology based on confidence annotation offers promising results, the variability in the results observed among different annotators calls for an in-depth analysis of the annotated labels, their advantages, and limitations. To this end, we proceed to study the accuracy of the annotations made by non-expert pathologists in the training subset, the number of samples labeled with low confidence, and their distribution in relation to the classes, in Figure 5.2. Likewise, we display the confusion matrices obtained by the non-expert annotators concerning the expert annotations, as well as those obtained using the model trained with hard labels and the proposed dual-branch entropy calibration, in Figure 5.3.

Regarding the gap observed between models trained using the ground truth or non-expert labels, this is due to the quality of the latter labels, which shows an average F1-score of 0.4510 (see Figure 5.2 (a)), which sets an upper limit on the results that the model can extract using pathologist-in-training labels. As observed in the corresponding confusion matrix (see Figure 5.3 (a)), this problem accentuates in certain classes such as lms and cef, which

Table 5.3: Quantitative comparison to prior literature. The metrics presented are the accuracy and micro-averaged F1-score (ACC/F1-score). The model trained with expert labels (second column) is used as the upper bound of the non-expert-based models. Colored values indicate the relative improvement of each method concerning the baseline using hard labels from non-expert in terms of the F1-score. Green indicates improvement and red a worsening lack. HL: hard labels; SL: soft labels; H^+: entropy maximization.

Annotator	**Expert**	**Non-Expert**				
	HL+H$^+$	**HL**	**SL**		**DBEC**	
1	0.653/0.620	0.408/0.277	0.428/0.295	↑ 1.8%	0.530/0.408	↑ 13.1%
2	0.571/0.467	0.408/0.288	0.530/0.424	↑ 13.6%	0.571/0.503	↑ 21.5%
3	0.612/0.584	0.448/0.386	0.489/0.401	↑ 1.5%	0.428/0.330	↓ 5.6%
4	0.551/0.520	0.448/0.309	0.428/0.355	↑ 4.6%	0.489/0.364	↑ 5.5%
5	0.673/0.601	0.551/0.448	0.571/0.460	↑ 1.2%	0.530/0.442	↓ 0.0%
6	0.591/0.555	0.428/0.298	0.428/0.304	↑ 0.6%	0.428/0.315	↑ 1.7%
7	0.673/0.602	0.469/0.348	0.551/0.427	↑ 7.9%	0.530/0.444	↑ 9.6%
8	0.693/0.655	0.367/0.259	0.408/0.270	↑ 1.1%	0.469/0.391	↑ 13.2%
9	0.653/0.614	0.387/0.280	0.469/0.323	↑ 4.3%	0.387/0.299	↑ 1.9%
10	0.632/0.525	0.469/0.353	0.530/0.390	↑ 3.7%	0.530/0.398	↑ 4.5%
Avg.	0.630/0.574	0.438/0.324	0.473/0.364	↑ 4.0%	0.489/0.389	↑ 6.6%

show lower prevalence concerning other classes in the used dataset (see Table 5.1). In addition, it can be observed how non-expert pathologists show lower confidence when labeling a sample corresponding to those categories (see Figure 5.2 (b)). This make sense since, for example, in the case of lms the pathologists-in-training are often confused with lm as they have the same morphological features. These limitations produce the drop in results between both types of labels observed in the quantitative metrics, which can be observed in the corresponding confusion matrix (see Figure 5.3 (b)). Interestingly, once the proposed dual-branch calibration is used, obtained results for those low-confidence classes improve (see Figure 5.3 (c)). Concretely, promising improvements for the classes lms, dfs, and fx are observed, which coincide with those categories that pathologists show the least confidence (see Figure 5.2 (c)). This may be produced by the lower-confidence entropy minimization, which encourages the model to produce confident predictions in those cases in which confidence falls below the fixed threshold τ. In this fashion, predicted labels move away from the annotator bias, based on the inherent features of each sample, and show the best generalization compared to expert annotations. Although the proposed approach offers consistent improvements among most annotators, still some limitations can be observed. For instance, it shows the least effect when noise increases. Annotators 3 and 9, which show low accuracy on the training dataset (see Figure 5.2 (a)), also offer worse results regarding



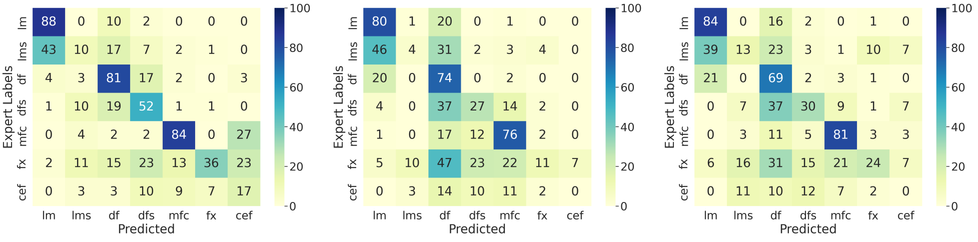
(a)

(b)

(c)

Figure 5.2: In-depth study of the soft labels annotated by in-training pathologists. (a) Quality of the labels, in terms of F1-score, in the training subset. Reference labels are the expert ground truth. (b) Percentage of samples with maximum confidence above the threshold $\tau = 0.7$. (c) Average confidence per each class, on positive samples. Dashed, red lines indicate average values.

the proposed approach. Also, if no use is made of soft labels (see 5.2 (b), annotator 5), the results remain the same as using hard labels.



(a) **(b)** **(c)**

Figure 5.3: Normalized confusion matrices, averaged among non-expert annotators, obtained using (a) raw hard labels, (b) the model trained using hard labels, and (c) the model trained using the dual-branch entropy calibration proposed in Eq. 5.5. Reference labels are the expert ground truth.

5.5.2 Ablation studies

The following experiments aim to demonstrate the convenience of the proposed approaches in an empirical fashion. First, we compare the benefits of labeling uncertainty instead of using a direct calibration of hard labels. Then, we motivate the choice of the components and hyper-parameters used for the proposed dual-branch uncertainty calibration setting in Eq. 5.5.

Artificial *vs.* annotated soft labels As previously discussed, we propose in this work to calibrate the model training to the inherent uncertainty of non-expert labeling by annotating the confidence for each independent class per sample. The benefit of calibrating CNNs to avoid overconfident predictions has already been demonstrated in previous literature [172]. We follow two main artificial methods used in this regard: label smoothing (LS) [178] and entropy regularization (H) [172]. Concretely, LS modifies the hard labels to assign a uniform distribution over non-positive categories such that: $Y_k^{LSR} = (1-\epsilon)Y_k + \frac{\epsilon}{K}$. Entropy calibration is based on Shannon entropy maximization (H^+), as described in the method section (see Eq. 5.3). In our experiments, we empirically optimized the hyper-parameters for both $\epsilon = 0.2$ and $\beta = 0.2$. We depict in Figure 5.4 the results using hard labels (HL), both artificial regularization approaches (LS and H^+), and the model trained using the proposed annotated soft labels (SL).

The obtained results show that regularizing neural network outputs improves the model performance. In particular, entropy-based regularization outperforms label smoothing, as indicated by previous literature [172, 180]. Concretely, average improvements of F1-score of $\sim 0.6\%$ and $\sim 2.4\%$ are obtained,

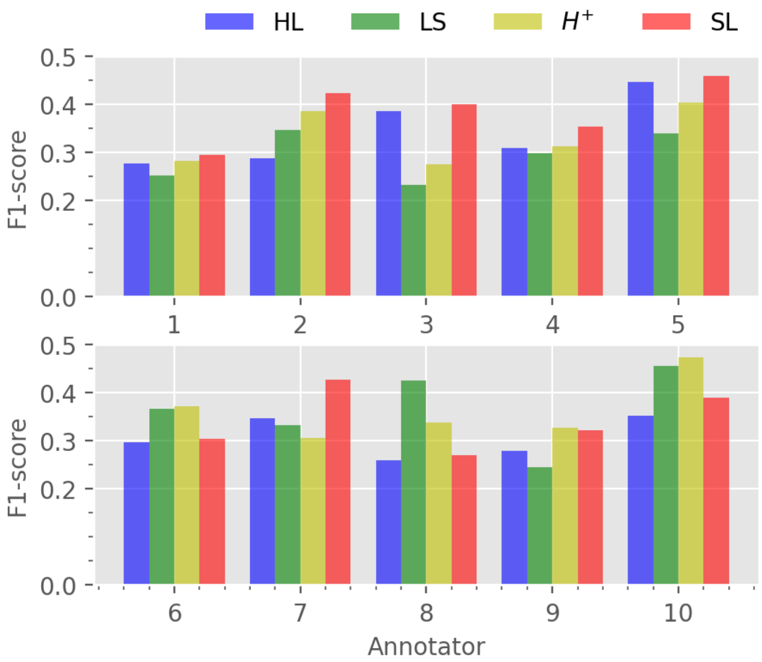


Figure 5.4: Ablation study on the use of artificial model calibration of hard labels (HL) or annotated soft labels (SL). For the first approach, label smoothing (LS) and entropy maximization (H^+) are used. F1-score is presented for each method and non-expert annotator.

respectively. The proposed labeling confidence approach outperforms the artificial entropy-based calibration across most annotators (see Figure 5.4 annotators 2, 3, 4, $8-10$). Concretely, an average improvement of $\sim 4\%$ is observed, as already depicted in Table 5.3. This indicates that labelling the confidence of the annotator for the different classes for each sample offers benefits beyond preventing the model from producing overconfident outputs. It is worth mentioning that this improvement is produced at a negligible cost, both in terms of annotation time and computational level. This may be because it introduces a sample-dependent distribution over labels, as opposed to these artificial methods.

Uncertainty calibration optimization The following experiments aim to demonstrate the convenience of the different components of the dual-branch entropy calibration (DBEC) for uncertainty assessment proposed in Eq. 5.4 when trained using soft labels (SL). Concretely, we fix the used threshold $\tau = 0.7$, then train and modify the relative weight of both branches to emulate the absence of each term. First, each term is trained individually, by using $\beta^{-} = 0$ and $\alpha^{-} = 0$, (DBEC (H^{+}) configuration), and $\beta^{+} = 0$ and $\alpha^{+} = 0$, (DBEC (H^{-}) configuration), respectively. Then, both terms are included as indicated in the implementation details. Average results among the 10 in-training pathologists are presented in Table5.4.

Table 5.4: Ablation experiment on the components of the proposed calibration formulation.

	Target Criteria			
	SL	DBEC (H^{+})	DBEC (H^{-})	DBEC ($H^{+/-}$)
ACC	0.438	0.461	0.386	0.489
F1-score	0.324	0.334	0.281	0.389

The results show that using only the positive entropy term, which calibrates the network by penalizing confident predictions, improves around $\sim$ 2% in terms of the F1-score. In contrast, using only low-confidence samples during training does not show good results. However, by incorporating this term into the general formulation, the figures of merit reach the improvements discussed earlier in the article. These results show the usefulness of including both terms in the proposed double-branch formulation.

In the following, we perform a study regarding the threshold used to compute the positive or negative entropy calibration, τ. Concretely, we sample homogeneously τ values between $[0, 1]$. The obtained results for representative annotators are depicted in Figure 5.5.

The performance of the DBEC proposed in relation to the τ value shows a characteristic shape. The non-expert annotators that show an improvement in the model performance using the proposed term first drop the obtained results when increasing τ. Then, an absolute maxima is reached around τ values of 0.7 and 0.8. Finally, increasing the hyper-parameter from this value worsen the performance, since entropy minimization is applied to all samples, even when high confidence is annotated. Based on these observations, we fixed $\tau = 0.7$ for the implementation of the dual-branch calibration.

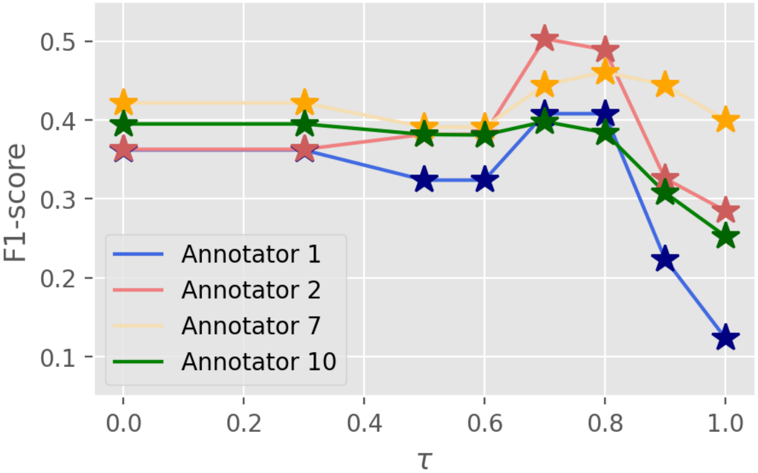


Figure 5.5: Ablation study on the effect of the confidence threshold τ on the proposed dual-branch entropy calibration (DBEC) based on annotated soft labels.

5.6 Conclusion

A relevant body of literature on uncertainty estimation requires multiple annotators to quantify individual sample noise and inter-annotator variability. Nevertheless, acquiring multiple rater views is a limiting factor in a wide range of applications, such as medical imagining. In particular, in the case of digital pathology imaging, a high level of expertise is required to perform image labeling, which may make it unfeasible to recruit multiple annotators. To address this limitation, in this work we have proposed to capture individual uncertainties by annotating soft labels instead of unique categories. In addition, and inspired by previous literature on model calibration using Shannon entropy, we have proposed a dual-branch min-max entropy calibration (DBEC) criteria that optimize the model training to (i) avoid overconfident outputs by entropy maximization, and (ii) produce confident outputs on samples labeled with high uncertainty by Shannon entropy minimization, which focuses on inherent features of each sample.

The proposed uncertainty estimation method is validated in the challenging context of skin whole slide image (WSI) multi-class image classification, under the multiple instance learning (MIL) paradigm. It is worth highlighting the scarce literature on this field since, to the best of our knowledge, this is the first work that aims to distinguish among 6 different relevant pathological categories. Over the AI4SKIN dataset, we have generated new uncertainty-driven soft labels from 10 in-training pathologists, so-called non-expert annotators. Uncertainty-aware MIL models have been trained using soft labels, and the novel dual-branch min-max entropy calibration, and they have been evaluated using a ground truth annotated by expert pathologists. Results show a consistent improvement by using soft labels with standard cross-entropy loss during training ($\sim 4.0\%$ F1-score), and increases when calibrating the model with the proposed min-max entropy calibration DBCE ($\sim 6.6\%$ F1-score). In addition, we have observed that improvements using the DBCE appear in categories that non-expert annotators presented high uncertainty, which supports our claim that the entropy minimization term in this case helps the model to move away from the annotator bias. These improvements are produced at a negligible cost, both at the level of annotation and calculation.

Still, during the experimental stage, we found some limitations in our study. First, the proposed formulations are still highly dependent on the quality of the produced labels. In the context of non-expert annotators, this may produce limitations when labels are too noisy. Likewise, the annotation of soft labels depends on the commitment of the experts recruited and does not bring improvements when performed in a very low proportion. We believe that the framework developed in this work opens the door to different interesting lines of further research. Learning how to combine certain expert labels with uncertain non-expert labels might be of great interest, such as crowd-sourcing methods able to obtain the underlying label distribution using the least number of annotators, among others.

Chapter 6

Final conclusions

This chapter relates the findings from each paper to the final aim of the PhD thesis. It summarises concluding remarks and suggests future research lines for each proposed learning framework.

Contents

6.1 Global remarks 109
6.2 Specific remarks 109
6.3 Future work 112

6.1 Global remarks

In this thesis, we have designed, developed and validated different weakly-supervised methods to solve real-world challenges in the medical sector when deep learning techniques are used. Concretely, our research has focused on comprehensively analyzing two widely used data modalities for detecting cancer and inflammatory diseases: genomic and histological data. For genomic data analysis, we have proposed unsupervised learning algorithms based on deep clustering. These algorithms effectively address the challenge of high-dimensional data by simultaneously clustering them without requiring any labels for training. As a result, this approach significantly enhances the accuracy of prognostic predictions. Then, this thesis has explored weakly supervised methods in WSI to improve the diagnosis of different disorders: ulcerative colitis, spitzoid and spindle cell cancer. A significant contribution of this thesis is the introduction of self-supervised learning algorithms for classifying gigapixel spitzoid histology images, utilizing inductive learning to overcome the limited number of available biopsies. Then, we have presented a novel formulation for applying weakly supervised methods to detect ulcerative colitis. Our research has also incorporated prior knowledge into constraint formulations, effectively leveraging valuable insights. Finally, we have created a new annotation protocol to measure the uncertainty of non-expert annotators. Leveraging these weak annotations, we have designed an uncertainty-aware pipeline to effectively handle the inherent uncertainty in the annotation process. The proposed methods in this thesis have undergone rigorous validation and, where applicable, have been compared with other state-of-the-art techniques. The results demonstrate their effectiveness and potential in advancing medical diagnostics and improving patient outcomes.

6.2 Specific remarks

In Chapter 2, we have presented the application of deep clustering methodologies for breast cancer detection using genomic data. In concrete, we have used the CpG island methylation levels related to the development of several types of cancer. The challenges associated with the high dimension of the methylation data and the reduced number of samples have prompted the exploration of cutting-edge techniques based on deep-embedded clustering. Our results demonstrate the promising performance of this end-to-end method, validated on two breast cancer databases, achieving accuracies of 0.9927 and 0.9375, respectively. The proposed method allows breast cancer classification using a latent space of only ten features, reducing the input data dimensionality by

99,9637%. Moreover, our proposed system outperforms other state-of-the-art methods based on classical clustering, surpassing their accuracy by more than 10%. This approach has also been validated in other disorders, specifically spitzoid melanocytic lesions. This study contributes to the advancement of DNA methylation research, potentially leading to personalized therapy due to the highly promising results of this technique.

The large size of digitalized histological images (WSIS) and the difficulty for pathologists to perform accurate annotations make it necessary to apply novel artificial intelligence-based methods to the automatic analysis of these gigapixel images. In Chapter 3, we have processed entire WSIs of spitzoid neoplasm in a weakly supervised manner under the multiple instance learning paradigm. In this line, we have proposed an inductive transfer learning framework based on a source CNN that allows for patch-level selection of tumor regions, which are then fed into a target model that focuses on the specific diagnosis of the entire biopsy. The inductive learning adopted allows the target model to start the biopsy-level training process from the backbone weights of the source model, facilitating its convergence. Different source architectures have been explored, highlighting the superiority of the new backbone proposed in this thesis (SeaNet). SeaNet is an improved convolutional neural network that introduces an attention module that can refine the latent space's features to maximize the classification agreement. Additionally, the use of attention weights on embedded-based MIL (in the target model) has shown the best performance since the embedding vector is more representative of the WSI. This innovative approach, carried out in an end-to-end manner, have reported promising results for both tumor selection and WSI classification, achieving a testing accuracy of 0.9231 and 0.800 for the source and the target models, despite the limited number of samples. The heat map findings are directly in line with the clinicians' medical decision and even highlight, in some cases, patterns of interest that were overlooked by the pathologist. Therefore, the proposed AI-based solution is valuable in addressing human eye fatigue and assisting inexperienced pathologists by suggesting inadvertent areas of interest to avoid biased diagnosis.

In most of the state-of-the-art works based on MIL, the WSIs employed have broad features that determine a positive bag. However, there are cases in which small cells differentiate whether a bag is positive. Therefore, the typical MIL approaches are not useful as the extracted activations are degraded and do not allow satisfactory classification. For that reason, we have presented in Chapter 4 a constraint optimization able to incorporate prior knowledge, in the form of relative localization of crucial elements, to the multiple instance learning formulation. Additionally, in the literature, there exist different weighted aggregation functions to obtain the biopsy embedding. However, in some situations, these attention weights have the same value for all instances in the bag, which is not suitable for determining a positive bag. To solve this problem, we have proposed to use a weighted average of instances where weights are obtained from the representative constrained activation maps optimized during the training. This new formulation has been applied to detect histological remission in ulcerative colitis using a new index developed by the PICASSO group. In the developed index, the leading biomarker for assessing histologic remission is the presence or absence of neutrophils. Therefore, the finding of this cell in specific colon structures indicates that the patient has ulcerative colitis activity. Using neutrophil localization during training, which pathologists promptly carry out, has greatly enhanced the results. We have demonstrated the robustness of the deep learning-based model in a multicentre study composed of a large cohort of biopsies collected and digitalized in 7 centers in the UK, Germany, Belgium, Italy, Canada and the USA. In general, the specificity of the MIL baseline models drops considerably. The best state-of-the-art model (CLAM-SB) has achieved a specificity of 0.8033 compared to 0.9615 obtained by the proposed model (LCMIL). Additionally, including the location constraints module forces the network to focus on the real determining cells, the neutrophils. In this way, we have obtained precise instance-level maps for unannotated images that allow us to detect neutrophils.

Finally, we have studied in Chapter 5 the condition that expert labels are not available since, in many cases, recruiting expert pathologists to annotate large databases is not feasible. In this line, we have designed an uncertainty-driven labeling strategy to generate soft labels from 10 non-expert annotators for multi-class skin cancer classification. The new protocol is based on categorical labeling and confidence percentage. Based on this soft annotation, we have proposed an uncertainty estimation-based framework to handle these noisy labels. This framework is based on a novel formulation using a dual-branch min-max entropy calibration to penalize inexact labels during the training. In contrast to the literature on uncertainty estimation that requires multiple annotators to quantify individual sample noise and inter-annotator variability, this model only needs individual annotations. Results have shown a consistent improvement by using soft labels with standard cross-entropy loss during training (4.0 % F1-score) and increase when calibrating the model with the proposed min-max entropy calibration DBCE (6.6% F1-score). In addition, we have observed that improvements using the DBCE appear in categories that non-expert annotators presented high uncertainty, which supports our claim that the entropy minimization term, in this case, helps the model to move away from the annotator bias. These improvements are produced at negligible cost in terms of annotation and calculation. The proposed system aims to assist less experienced pathologists in diagnosing the specific type of spindle cell neoplasm when they can recognize it as such but are uncertain about its classification or potential malignancy.

6.3 Future work

In this thesis, different weakly-supervised methods have been explored to mitigate the dependency of deep learning on expert-labeled data. This thesis has also aimed to incorporate explainability techniques to enhance the understanding of the developed models. Specifically, attention mechanisms and instance-level saliency maps have been employed, enabling the visualization of relevant elements and their contributions to predictions. However, there are many research possibilities to go further in the explainability domain. Future studies can explore cutting-edge methods such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), which provide instance-level explanations and highlight the most influential features. Moreover, concept-based explanations can also be sought to discover the high-level concepts learned by the models, utilizing techniques like Concept Activation Vectors (CAVs) or Generative Adversarial Networks (GANs). Using CAVs enables the identification of the attributes or features that contribute most significantly to specific concepts, thereby providing insights into how the model interprets and understands that concepts. Additionally, GANs can generate synthetic examples that maximize the activation of a particular concept, facilitating a deeper understanding of the model's learned representations. Integrating these advanced explainability methods will offer a more comprehensive understanding of model decisions, encouraging transparency, trustworthiness, and deeper insights into the developed models.

Finally, we would like to highlight some future possibilities regarding real-world applications. Within the computer vision field, most research has focused on developing static models trained when all training data is available at once. However, this context does not match real-world scenarios like the medical field, where new data sets are continually arising. To tackle this problem, further research should focus on incremental learning, also known as continual or lifelong learning. It attempts to mimic natural vision systems capable of integrating new information while retaining previous knowledge. Incremental learning pretends to address the stability-plasticity dilemma. Specifically, continual learning strategies aim to force the model to retain knowledge from the old data (stability) while acquiring the pertinent knowledge to perform correctly on the novel one (plasticity). In this sense, we also aim to use the potential of large-scale foundation models for the classification or segmentation of histological data. Foundation models are trained using various centers, acquisition systems, study types and tasks. These models tend to offer better transferability when updated on new tasks and domains.

Merits

Journal papers

Golfe, A., **del Amor, R.**, Colomer, A., Sales, M.A., Terradez, L., & Naranjo, V. ProGleason-GAN: Conditional Progressive Growing GAN for prostatic cancer Gleason Grade patch synthesis. Computer Methods and Programs in Biomedicine (2023).

Pulgarín-Ospina, CC., **del Amor, R.**, Colomer, A., Silva-Rodríguez, J., & Naranjo, V. HistoColAi: An Open-Source Web Platform for Collaborative Digital Histology Image Annotation with AI-Driven Predictive Integration. Computer Methods and Programs in Biomedicine, 2023 (under review).

del Amor, R., Pérez-Cano, J., López-Pérez, M., Terradez,L., Aneiros-Fernandez, J., Morales,S., Mateos, J., Molina,R., & Naranjo, V. Annotation Protocol and Crowdsourcing Multiple Instance Learning Classification of Skin Histological Images: the CR-AI4SkIN Dataset. Artificial Intelligence in Medicine 2023 (under review).

del Amor, R., Silva-Rodríguez, J., & Naranjo, V. Labeling confidence for uncertainty-aware histology image classification. Computerized Medical Imaging and Graphics, 102231, (2023).

Iacucci, M., Lorenzo Parigi, T., **del Amor, R.**, Meseguer, P., ... Naranjo,V., & Villanacci, V. Artificial Intelligence enabled histological prediction of remission or activity and clinical outcomes in ulcerative colitis. Gastroenterology, 164, (2023).

Villanacci, V., Lorenzo Parigi, T., **del Amor, R.**, Meseguer, P., ... Naranjo,V., & Iacucci, M. 277: A New Simplified Histology Artificial Intelligence

System For Accurate Assessment of Remission in Ulcerative Colitis, Gastroenterology. 162, (2023).

Xianyong, G., Bazarova, A., **del Amor, R.**, ... Naranjo,V., Ghosh, S., & Iacucci, M. PICaSSO Histologic Remission Index (PHRI) in ulcerative colitis: development of a novel simplified histological score for monitoring mucosal healing and predicting clinical outcomes and its applicability in an artificial intelligence system. GUT, 71, (2022).

Iacucci, M., Cannatelli, R., Parigi, T L., ... **del Amor, R.**, ... Ghosh, S., & Grisan, E. OP16 The first virtual chromoendoscopy artificial intelligence system to detect endoscopic and histologic remission in Ulcerative Colitis. Journal of Crohn's and Colitis, 16, (2022).

del Amor, R., Meseguer P., Lorenzo-Parigi, T., Villanacci, V., Colomer A., ... & Naranjo, V. Constrained multiple instance learning for ulcerative colitis prediction using histological images. Computer Methods and Programs in Biomedicine, 224, (2022).

Villanacci, V., Parigi, T L., **del Amor, R.**, ... Naranjo,V., & Iacucci, M. OP15 A new simplified histology artificial intelligence system for accurate assessment of remission in Ulcerative Colitis. Journal of Crohn s and Colitis, 16, (2022).

Iacucci, M., Cannatelli, R., Parigi, T L., ..., **del Amor, R.**, ... Naranjo,V., Ghosh S., & Grisan, E. A virtual chromoendoscopy artificial intelligence system to detect endoscopic and histologic activity/remission and predict clinical outcomes in ulcerative colitis. Endoscopy, 55, (2022).

Berenguer-Vidal, R., Verdú-Monedero, R., Morales-Sánchez, J., Sellés-Navarro, I., **del Amor, R.**, García, G., & Naranjo, V. Automatic segmentation of the retinal nerve fiber layer by means of mathematical morphology and deformable models in 2D optical coherence tomography imaging. Sensors, 23, 8027 (2021).

García, G., **del Amor, R.**, Colomer, A., Verdú-Monedero, R., Morales-Sánchez, J., & Naranjo, V. Circumpapillary OCT-focused hybrid learning for glaucoma grading using tailored prototypical neural networks. Artificial Intelligence in Medicine, 118, 102132 (2021).

del Amor, R., Colomer, A., Monteagudo, C., & Naranjo, V. A deep embedded refined clustering approach for breast cancer distinction based on DNA methylation. Neural Computing and Applications, 34, (2021).

del Amor, R., Launet, L., Colomer, A., Moscardó, A., Mosquera-Zamudio, A., Monteagudo, C., & Naranjo, V. An attention-based weakly supervised framework for spitzoid melanocytic lesion diagnosis in whole slide images. Artificial Intelligence in Medicine, 121, (2021).

Morales, S., Colomer, A., Mossi., JM., **del Amor, R.**, Woldbye, D., Klemp, k., Larsen, M., & Naranjo, V. Retinal layer segmentation in rodent OCT images: Local intensity profiles fully convolutional neural networks. Computer Methods and Programs in Biomedicine, 198, (2021).

del Amor, R., Morales, S., Colomer, A., Mogesen, M., Jensen, M., Israelsen, NM., Bang, O., & Naranjo, V. Automatic Segmentation of Epidermis and Hair Follicles in Optical Coherence Tomography Images of Normal Skin by Convolutional Neural Networks. Frontiers in Biomedicine, 198, (2020).

National conferences

del Amor, R., Colomer, A., & Naranjo, V. *El rol de la inteligencia artificial generativa en la educación: beneficios potenciales de ChatGPT para promover el aprendizaje en tareas de programación en Python* in *IX Congreso de Innovación Educativa y Docencia en Red (Inred)* (2023).

International conferences

del Amor, R., Colomer, A., Morales, S., Pulgarín-Ospina, C., Terradez, L., Aneiros-Fernandez, J., & Naranjo, V. *A Self-Contrastive Learning Framework for Skin Cancer Detection using Histological Images* in *2022 IEEE International Conference on Image Processing (ICIP)* (2022) 2291-2295.

Pastor-Naranjo, F., **del Amor, R.**, Silva-Rodríguez, J., Ferrer Contreras, M., Piñero, G., & Naranjo, V. *Conditional Generative Adversarial Networks for Acoustic Echo Cancellation* in *2022 European Signal Processing Conference (EUSIPCO)* (2022) 85-89.

Launet, L., **del Amor, R.**, Colomer, A., Mosquera-Zamudio, A., Moscardó, A., Monteagudo, C., Zhao, Z., & Naranjo, V. *Federating Unlabeled Samples: A Semi-supervised Collaborative Framework for Whole Slide Image Analysis* in *2022 Intelligent Data Engineering and Automated Learning (IDEAL)* (2022) 64-72.

del Amor, R., Curieses, F. J., Launet, L., Colomer, A., Moscardó, A., Mosquera-Zamudio, A., Monteagudo, C., & Naranjo, V. *Multi-Resolution Framework For Spitzoid Neoplasm Classification Using Histological Data* in *2022 IEEE Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)* (2022).

Ramírez, E., **del Amor, R.**, García, G., & Naranjo, V. *Glaucoma Grading Via Mean Defect Back Propagation From Oct Images* in *2022 30th European Signal Processing Conference (EUSIPCO)* (2022), 1278-1282.

del Amor, R., Colomer, A., Monteagudo, C., Garzón, MJ., García-Giménez, J., & Naranjo, V. *A Deep Embedded Framework for Spitzoid Neoplasm Classification Using DNA Methylation Data* in *2021 29th European Signal Processing Conference (EUSIPCO)* (2021), 1271-1275.

García, G., **del Amor, R.**, Colomer, A. & Naranjo, V. *Glaucoma Detection From Raw Circumpapillary OCT Images Using Fully Convolutional Neural Networks* in *2020 IEEE International Conference on Image Processing (ICIP)* (2020), 2526–2530.

del Amor, R., Morales, S., Colomer, A. Mossi, JM., Woldbye, D., Klemp, K., Larsen, M., & Naranjo, V. *Towards Automatic Glaucoma Assessment: An Encoder-decoder CNN for Retinal Layer Segmentation in Rodent OCT* *images* in *2019 27th European Signal Processing Conference (EUSIPCO)* (2019).

Research awards

Villanacci, V., Parigi, T L., **del Amor, R.**, ... Naranjo,V., & Iacucci, M. OP15 A new simplified histology artificial intelligence system for accurate assessment of remission in Ulcerative Colitis. Journal of Crohn s and Colitis, 16, (2022), *ECCO 2022 congress highlight (best 10 presentation).*

Xianyong, G., Bazarova, A., **del Amor, R.**, ... Naranjo,V., Ghosh, S., & Iacucci, M. PICaSSO Histologic Remission Index (PHRI) in ulcerative colitis: development of a novel simplified histological score for monitoring mucosal healing and predicting clinical outcomes and its applicability in an artificial intelligence system. GUT, 71, (2022), *GUTTopPAper 2022.*

Technology transfer activities

D4HEALTH: Digital diaries for diagnosis of psychological disorders for mental health tracking (S-097-2021). Naranjo Ornedo, Valeriana; Colomer Granero, Adrián; López Mir, Fernando; del Amor del Amor, Rocío. 13/07/2021.

Pixnormous - plataforma web para la visualización, anotación y análisis de imágenes gigapixel digitalizadas (S-172-2023). Naranjo Ornedo, Valeriana; Colomer Granero, Adrián; López Mir, Fernando; del Amor del Amor, Rocío. 20/07/2023.

Bibliography

1. McCarthy, J. What is artificial intelligence (2007).
2. Helm, J. M. *et al.* Machine learning and artificial intelligence: definitions, applications, and future directions. *Current reviews in musculoskeletal medicine* **13,** 69–76 (2020).
3. Chauhan, N. K. & Singh, K. *A review on conventional machine learning vs deep learning* in *2018 International conference on computing, power and communication technologies (GUCON)* (2018), 347–352.
4. Ojala, T., Pietikäinen, M. & Harwood, D. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition* **29,** 51–59 (1996).
5. Dalal, N. & Triggs, B. *Histograms of oriented gradients for human detection* in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* **1** (2005), 886–893.
6. Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* **60,** 91–110 (2004).
7. Mammone, A., Turchi, M. & Cristianini, N. Support vector machines. *Wiley Interdisciplinary Reviews: Computational Statistics* **1,** 283–289 (2009).
8. Cutler, A., Cutler, D. R. & Stevens, J. R. Random forests. *Ensemble machine learning: Methods and applications,* 157–175 (2012).
9. Ahmad, A. & Dey, L. A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering* **63,** 503–527 (2007).
10. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nature medicine* **25,** 24–29 (2019).
11. Lee, J.-G. *et al.* Deep learning in medical imaging: general overview. *Korean journal of radiology* **18,** 570–584 (2017).

12. Elyan, E. *et al.* Computer vision and machine learning for medical image analysis: recent advances, challenges, and way forward. *Artificial Intelligence Surgery* **2** (2022).
13. Schneider, L. *et al.* Integration of deep learning-based image analysis and genomic data in cancer pathology: a systematic review. *European journal of cancer* **160,** 80–91 (2022).
14. Mueller, B., Kinoshita, T., Peebles, A., Graber, M. A. & Lee, S. Artificial intelligence and machine learning in emergency medicine: a narrative review. *Acute medicine & surgery* **9,** e740 (2022).
15. Schwab, E., Gooßen, A., Deshpande, H. & Saalbach, A. *Localization of critical findings in chest X-ray without local annotations using multi-instance learning* in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* (2020), 1879–1882.
16. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **22,** 1345–1359 (2010).
17. Deng, J. *et al. Imagenet: A large-scale hierarchical image database* in *2009 IEEE conference on computer vision and pattern recognition* (2009), 248–255.
18. Krizhevsky, A., Hinton, G., *et al.* Learning multiple layers of features from tiny images (2009).
19. Lin, T.-Y. *et al. Microsoft coco: Common objects in context* in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (2014), 740–755.
20. Cheplygina, V., de Bruijne, M. & Pluim, J. P. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis* **54,** 280–296 (2019).
21. Ren, Z., Wang, S. & Zhang, Y. Weakly supervised machine learning. *CAAI Transactions on Intelligence Technology* (2023).
22. Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T. & Engelhardt, B. E. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome biology* **16,** 14 (2015).
23. Akhavan-Niaki, H. & Samadani, A. A. DNA methylation and cancer development: molecular mechanism. *Cell biochemistry and biophysics* **67,** 501–513 (2013).
24. Surace, A. E. A. & Hedrich, C. M. The role of epigenetics in autoimmune/inflammatory disease. *Frontiers in immunology* **10,** 1525 (2019).

25. Yuvaraj, N. & Vivekanandan, P. *An efficient SVM based tumor classification with symmetry non-negative matrix factorization using gene expression data* in *2013 International Conference on Information Communication and Embedded Systems (Icices)* (2013), 761–768.
26. Jazayeri, N. & Sajedi, H. Breast cancer diagnosis based on genomic data and extreme learning machine. *SN Applied Sciences* **2,** 3 (2020).
27. Si, Z., Yu, H. & Ma, Z. Learning deep features for dna methylation data analysis. *IEEE Access* **4,** 2732–2737 (2016).
28. Khwaja, M., Kalofonou, M. & Toumazou, C. A Deep Autoencoder System for Differentiation of Cancer Types Based on DNA Methylation State. *arXiv preprint arXiv:1810.01243* (2018).
29. Titus, A. J., Wilkins, O. M., Bobak, C. A. & Christensen, B. C. Unsupervised deep learning with variational autoencoders applied to breast tumor genome-wide DNA methylation data with biologic feature extraction. *bioRxiv,* 433763 (2018).
30. Liu, B. *et al.* DNA Methylation Markers for Pan-Cancer Prediction by Deep Learning. *Genes* **10,** 778 (2019).
31. Amor, R. d., Colomer, A., Monteagudo, C. & Naranjo, V. A deep embedded refined clustering approach for breast cancer distinction based on DNA methylation. *Neural Computing and Applications,* 1–13 (2021).
32. Del Amor, R. *et al.* *A Deep Embedded Framework for Spitzoid Neoplasm Classification Using DNA Methylation Data* in *2021 29th European Signal Processing Conference (EUSIPCO)* (2021), 1271–1275.
33. Snead, D. R. *et al.* Validation of digital pathology imaging for primary histopathological diagnosis. *Histopathology* **68,** 1063–1072 (2016).
34. Srinidhi, C. L., Ciga, O. & Martel, A. L. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis* **67** (2021).
35. Van der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: the path to the clinic. *Nature medicine* **27,** 775–784 (2021).
36. Del Amor, R. *et al.* *A Self-Contrastive Learning Framework for Skin Cancer Detection Using Histological Images* in *2022 IEEE International Conference on Image Processing (ICIP)* (2022).

37. Launet, L. *et al. Federating Unlabeled Samples: A Semi-supervised Collaborative Framework for Whole Slide Image Analysis* in *Intelligent Data Engineering and Automated Learning–IDEAL 2022: 23rd International Conference, IDEAL 2022, Manchester, UK, November 24–26, 2022, Proceedings* (2022), 64–72.
38. Del Amor, R., Silva-Rodríguez, J. & Naranjo, V. Labeling confidence for uncertainty-aware histology image classification. *Computerized Medical Imaging and Graphics,* 102231 (2023).
39. Iacucci, M. *et al.* Artificial Intelligence enabled histological prediction of remission or activity and clinical outcomes in ulcerative colitis. *Gastroenterology* (2023).
40. Villanacci, V. *et al.* OP15 A new simplified histology artificial intelligence system for accurate assessment of remission in Ulcerative Colitis. *Journal of Crohn's and Colitis* **16,** i015–i017 (2022).
41. Gui, X. *et al.* PICaSSO Histologic Remission Index (PHRI) in Ulcerative Colitis–Development of a Novel Simplified Histological Score for Monitoring Mucosal Healing and Predicting Clinical Outcomes and its Applicability in an Artificial Intelligence System. *Gut* (2022).
42. Del Amor, R. *et al.* Constrained multiple instance learning for ulcerative colitis prediction using histological images. *Computer methods and programs in biomedicine* **224,** 107012 (2022).
43. Pathak, D., Krahenbuhl, P. & Darrell, T. *Constrained convolutional neural networks for weakly supervised segmentation* in *Proceedings of the IEEE international conference on computer vision* (2015), 1796–1804.
44. Jia, Z., Huang, X., Eric, I., Chang, C. & Xu, Y. Constrained deep weak supervision for histopathology image segmentation. *IEEE transactions on medical imaging* **36,** 2376–2388 (2017).
45. Kervadec, H. *et al.* Constrained-CNN losses for weakly supervised segmentation. *Medical image analysis* **54,** 88–99 (2019).
46. Zhu, Y., Tang, S., Quan, L., Jiang, W. & Zhou, L. Extraction method for signal effective component based on extreme-point symmetric mode decomposition and Kullback–Leibler divergence. *Journal of the Brazilian Society of Mechanical Sciences and Engineering* **41,** 1–11 (2019).
47. Xu, Z. *et al.* Anti-interference from Noisy Labels: Mean-Teacher-assisted Confident Learning for Medical Image Segmentation. *IEEE Transactions on Medical Imaging,* 1–13 (2022).

48. Xu, Z. *et al. Noisy labels are treasure: mean-teacher-assisted confident learning for hepatic vessel segmentation* in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2021).
49. Luo, W. & Yang, M. *Semi-supervised semantic segmentation via strong-weak dual-branch network* in *European Conference on Computer Vision (ECCV)* (2020).
50. Del Amor, R. *et al.* An attention-based weakly supervised framework for spitzoid melanocytic lesion diagnosis in whole slide images. *Artificial intelligence in medicine* **121,** 102197 (2021).
51. Sharma, S., Kelly, T. K. & Jones, P. A. Epigenetics in cancer. *Carcinogenesis* **31,** 27–36 (2010).
52. Tsou, J. A., Hagen, J. A., Carpenter, C. L. & Laird-Offringa, I. A. DNA methylation analysis: a powerful new tool for lung cancer diagnosis. *Oncogene* **21,** 5450–5461 (2002).
53. Esteller, M. Epigenetics in cancer. *New England Journal of Medicine* **358,** 1148–1159 (2008).
54. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics* **11,** 587 (2010).
55. Martorell-Marugán, J. *et al.* in *Computational Biology [Internet]* (Codon Publications, 2019).
56. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98,** 288–295 (2011).
57. Laird, P. W. Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics* **11,** 191–203 (2010).
58. Bellman, R. Dynamic programming, princeton univ. *Princeton* (1957).
59. Venkat, N. The curse of dimensionality: Inside out (2018).
60. Tasoulis, S., Pavlidis, N. G. & Roos, T. Nonlinear dimensionality reduction for clustering. *Pattern Recognition* **107,** 107508 (2020).
61. Hofmeyr, D. P. Clustering by minimum cut hyperplanes. *IEEE transactions on pattern analysis and machine intelligence* **39,** 1547–1560 (2016).
62. Cevikalp, H. High-dimensional data clustering by using local affine/convex hulls. *Pattern Recognition Letters* **128,** 427–432 (2019).
63. Araújo, A. F., Antonino, V. O. & Ponce-Guevara, K. L. Self-organizing subspace clustering for high-dimensional and multi-view data. *Neural Networks* **130,** 253–268 (2020).

64. Guo, X., Liu, X., Zhu, E. & Yin, J. *Deep clustering with convolutional autoencoders* in *International conference on neural information processing* (2017), 373–382.

65. Xie, J., Girshick, R. & Farhadi, A. *Unsupervised deep embedding for clustering analysis* in *International conference on machine learning* (2016), 478–487.

66. Guo, X., Zhu, E., Liu, X. & Yin, J. *Deep embedded clustering with data augmentation* in *Asian conference on machine learning* (2018), 550–565.

67. Enguehard, J., O'Halloran, P. & Gholipour, A. Semi-supervised learning with deep embedded clustering for image classification and segmentation. *IEEE Access* **7,** 11093–11104 (2019).

68. Hershey, J. R., Chen, Z., Le Roux, J. & Watanabe, S. *Deep clustering: Discriminative embeddings for segmentation and separation* in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016), 31–35.

69. Prasetio, B. H., Tamura, H. & Tanno, K. *A Deep Time-delay Embedded Algorithm for Unsupervised Stress Speech Clustering* in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (2019), 1193–1198.

70. Tian, T., Wan, J., Song, Q. & Wei, Z. Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nature Machine Intelligence* **1,** 191–198 (2019).

71. GEO. *Epigenome Analysis of Breast Tissue From Women With and ithout Breast Cancer* `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse32393`.

72. Zhuang, J. *et al.* The dynamics and prognostic potential of DNA methylation changes at stem cell gene loci in women's cancer. *PLoS genetics* **8,** e1002517 (2012).

73. Halvorsen, A. R. *et al.* Differential DNA methylation analysis of breast cancer reveals the impact of immune signaling in radiation therapy. *International journal of cancer* **135,** 2085–2095 (2014).

74. Min, E. *et al.* A survey of clustering with deep learning: From the perspective of network architecture. *IEEE Access* **6,** 39501–39514 (2018).

75. Foster, D. *Generative deep learning: teaching machines to paint, write, compose, and play* (O'Reilly Media, 2019).

76. Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9,** 2579–2605 (2008).

77. Hubert, L. & Arabie, P. Comparing partitions. *Journal of classification* **2,** 193–218 (1985).
78. Strehl, A. & Ghosh, J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research* **3,** 583–617 (2002).
79. Apalla, Z., Lallas, A., Sotiriou, E., Lazaridou, E. & Ioannides, D. Epidemiological trends in skin cancer. *Dermatology practical & conceptual* **7,** 1–6 (2017).
80. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2017. *CA: A Cancer Journal for Clinicians* **67,** 7–30 (2017).
81. Wiesner, T. *et al.* Genomic aberrations in spitzoid melanocytic tumours and their implications for diagnosis, prognosis and therapy. *Pathology* **48,** 113–131 (2016).
82. Barnhill, R. L. The Spitzoid lesion: rethinking Spitz tumors, atypical variants,'Spitzoid melanoma'and risk assessment. *Modern pathology* **19,** S21–S33 (2006).
83. Lodha, S., Saggar, S., Celebi, J. T. & Silvers, D. N. Discordance in the histopathologic diagnosis of difficult melanocytic neoplasms in the clinical setting. *Journal of cutaneous pathology* **35,** 349–352 (2008).
84. Gurcan, M. N. *et al.* Histopathological image analysis: A review. *IEEE reviews in biomedical engineering* **2,** 147–171 (2009).
85. Bejnordi, B. E. *et al.* Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama* **318,** 2199–2210 (2017).
86. Rakhlin, A., Shvets, A., Iglovikov, V. & Kalinin, A. A. Deep convolutional neural networks for breast cancer histology image analysis. *international conference image analysis and recognition* **10882,** 737–744 (2018).
87. Litjens, G. *et al.* Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports* **6,** 1–11 (2016).
88. Silva-Rodríguez, J., Colomer, A., Sales, M. A., Molina, R. & Naranjo, V. Going deeper through the Gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer Methods and Programs in Biomedicine* **195,** 105637 (2020).
89. Del Toro, O. J. *et al.* Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score. *Medical Imaging 2017: Digital Pathology* **10140,** 101400O (2017).

90. Yu, K.-H. *et al.* Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature communications* **7,** 1–10 (2016).

91. Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine* **24,** 1559–1567 (2018).

92. Codella, N. C. *et al.* Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM Journal of Research and Development* **61,** 5–1 (2017).

93. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *nature* **542,** 115–118 (2017).

94. Haenssle, H. A. *et al.* Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology* **29,** 1836–1842 (2018).

95. Maron, R. C. *et al.* Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. *European Journal of Cancer* **119,** 57–65 (2019).

96. Brinker, T. J. *et al.* Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer* **113,** 47–54 (2019).

97. Kassani, S. H. & Kassani, P. H. A comparative study of deep learning architectures on melanoma detection. *Tissue and Cell* **58,** 76–83 (2019).

98. Liu, Y. *et al.* A deep learning system for differential diagnosis of skin diseases. *Nature Medicine* **26,** 900–908 (2020).

99. Astorino, A., Fuduli, A., Veltri, P. & Vocaturo, E. Melanoma detection by means of Multiple Instance Learning. *Interdisciplinary Sciences: Computational Life Sciences* **12,** 24–31 (2020).

100. Yu, C. *et al.* Acral melanoma detection using a convolutional neural network for dermoscopy images. *PloS one* **13,** 1–14 (2018).

101. Hekler, A. *et al.* Pathologist-level classification of histopathological melanoma images with deep neural networks. *European Journal of Cancer* **115,** 79–83 (2019).

102. De Logu, F. *et al.* Recognition of Cutaneous Melanoma on Digitized Histopathological Slides via Artificial Intelligence Algorithm. *Frontiers in oncology* **10,** 1559 (2020).

103. Wang, L. *et al.* Automated identification of malignancy in whole-slide pathological images: identification of eyelid malignant melanoma in gigapixel pathological slides using deep learning. *British Journal of Ophthalmology* **104,** 318–323 (2020).

104. Devalland, C. Spitzoid Lesions Diagnosis Based on SMOTE-GA and Stacking Methods. *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019): Volume 2-Advanced Intelligent Systems for Sustainable Development Applied to Agriculture and Health* **1103,** 348 (2020).

105. Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. *Journal of Big data* **3,** 1–40 (2016).

106. Vilalta, R., Giraud-Carrier, C., Brazdil, P. & Soares, C. Inductive Transfer. *Springer US,* 634–683 (2010).

107. Caruana, R. Multitask learning. *Machine learning* **28,** 41–75 (1997).

108. Silver, D. L. & Mercer, R. E. The task rehearsal method of life-long learning: Overcoming impoverished data. *Conference of the Canadian Society for Computational Studies of Intelligence,* 90–101 (2002).

109. Zhang, S. *et al.* Computer-aided diagnosis (CAD) of pulmonary nodule of thoracic CT image using transfer learning. *Journal of digital imaging* **32,** 995–1007 (2019).

110. Tokuoka, Y., Suzuki, S. & Sugawara, Y. An Inductive Transfer Learning Approach using Cycle-consistent Adversarial Domain Adaptation with Application to Brain Tumor Segmentation. *Proceedings of the 2019 6th International Conference on Biomedical and Bioinformatics Engineering,* 44–48 (2019).

111. Zhou, Y., Wang, B., Huang, L., Cui, S. & Shao, L. A benchmark for studying diabetic retinopathy: Segmentation, grading, and transferability. *IEEE Transactions on Medical Imaging,* 818–828 (2020).

112. De Bois, M., El Yacoubi, M. A. & Ammi, M. Adversarial multi-source transfer learning in healthcare: Application to glucose prediction for diabetic people. *Computer Methods and Programs in Biomedicine* **199,** 105874 (2021).

113. Srinidhi, C. L., Ciga, O. & Martel, A. L. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis,* 101813 (2020).

114. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine* **25,** 1301–1309 (2019).

115. Das, K., Conjeti, S., Chatterjee, J. & Sheet, D. Detection of Breast Cancer from Whole Slide Histopathological Images using Deep Multiple Instance CNN. *IEEE Access,* 213502–213511 (2020).

116. Zhao, Y. *et al.* Predicting Lymph Node Metastasis Using Histopathological Images Based on Multiple Instance Learning With Deep Graph Convolution. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 4837–4846 (2020).

117. Silva-Rodriguez, J., Colomer, A., Dolz, J. & Naranjo, V. Self-learning for weakly supervised Gleason grading of local patterns. *IEEE journal of biomedical and health informatics,* 3094–3104 (2021).

118. Openseadragon. *Archivo Situacionista Hispano* 1999.

119. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556,* 213502–213511 (2014).

120. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition,* 770–778 (2016).

121. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition,* 2818–2826 (2016).

122. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition,* 4700–4708 (2017).

123. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition,* 7132–7141 (2018).

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

125. Ng, S. C. *et al.* Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *The Lancet* **390,** 2769–2778 (2017).

126. Jain, D., Warren, B. F. & Riddell, R. H. Inflammatory disorders of the large intestine. *Morson and Dawson's gastrointestinal pathology,* 552–635 (2013).

127. Bryant, R. V. *et al.* Beyond endoscopic mucosal healing in UC: histological remission better predicts corticosteroid use and hospitalisation over 6 years of follow-up. *Gut* **65,** 408–414 (2016).
128. Narang, V. *et al.* Association of endoscopic and histological remission with clinical course in patients of ulcerative colitis. *Intestinal research* **16,** 55 (2018).
129. Ponte, A. *et al.* Impact of histological and endoscopic remissions on clinical recurrence and recurrence-free time in ulcerative colitis. *Inflammatory bowel diseases* **23,** 2238–2244 (2017).
130. Lobatón, T. *et al.* Prognostic value of histological activity in patients with ulcerative colitis in deep remission: A prospective multicenter study. *United European gastroenterology journal* **6,** 765–772 (2018).
131. Römkens, T. E. *et al.* Assessment of histological remission in ulcerative colitis: discrepancies between daily practice and expert opinion. *Journal of Crohn's and Colitis* **12,** 425–431 (2018).
132. Alsoud, D. *et al.* P442 Real-world endoscopic and histologic outcomes are linked to ustekinumab exposure in Ulcerative Colitis. *Journal of Crohn's and Colitis* **16,** i424–i424 (2022).
133. Magro, F. *et al.* Comparison of different histological indexes in the assessment of UC activity and their accuracy regarding endoscopic outcomes and faecal calprotectin levels. *Gut* **68,** 594–603 (2019).
134. Mosli, M. H. *et al.* Histologic evaluation of ulcerative colitis: a systematic review of disease activity indices. *Inflammatory bowel diseases* **20,** 564–575 (2014).
135. Mojtahed, A. *et al.* Assessment of histologic disease activity in Crohn's disease: a systematic review. *Inflammatory bowel diseases* **20,** 2092–2103 (2014).
136. Riley, S., Mani, V., Goodman, M., Dutt, S. & Herd, M. Microscopic activity in ulcerative colitis: what does it mean? *Gut* **32,** 174–178 (1991).
137. Pai, R. K. *et al.* Complete resolution of mucosal neutrophils associates with improved long-term clinical outcomes of patients with ulcerative colitis. *Clinical Gastroenterology and Hepatology* **18,** 2510–2517 (2020).
138. Magro, F. *et al.* ECCO position paper: harmonization of the approach to ulcerative colitis histopathology. *Journal of Crohn's and Colitis* **14,** 1503–1511 (2020).
139. Ma, C. *et al.* An international consensus to standardize integration of histopathology in ulcerative colitis clinical trials. *Gastroenterology* **160,** 2291–2302 (2021).

140. Stidham, R. W. *et al.* Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis. *JAMA network open* **2,** e193963–e193963 (2019).

141. Ozawa, T. *et al.* Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis. *Gastrointestinal endoscopy* **89,** 416–421 (2019).

142. Alammari, A. *et al. Classification of ulcerative colitis severity in colonoscopy videos using cnn* in *Proceedings of the 9th international conference on information management and engineering* (2017), 139–144.

143. Byrne, M. *et al.* DOP13 Artificial Intelligence (AI) in endoscopy-Deep learning for detection and scoring of Ulcerative Colitis (UC) disease activity under multiple scoring systems. *Journal of Crohn's and Colitis* **15,** S051–S052 (2021).

144. Bhambhvani, H. P. & Zamora, A. Deep learning enabled classification of Mayo endoscopic subscore in patients with ulcerative colitis. *European Journal of Gastroenterology & Hepatology* **33,** 645–649 (2021).

145. Casteele, V. *et al.* Utilizing Deep Learning to Analyze Whole Slide Images of Colonic Biopsies for Associations Between Eosinophil Density and Clinicopathologic Features in Active Ulcerative Colitis. *Inflammatory Bowel Diseases* (2021).

146. Li, B., Li, Y. & Eliceiri, K. W. *Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), 14318–14328.

147. Lu, M. Y. *et al.* Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering* **5,** 555–570 (2021).

148. Iacucci, M. *et al.* An international multicenter real-life prospective study of electronic chromoendoscopy score PICaSSO in Ulcerative Colitis. *Gastroenterology* **160,** 1558–1569 (2021).

149. Ström, P. *et al.* Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *The Lancet Oncology* **21,** 222–232 (2020).

150. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine* **25,** 1301–1309 (2019).

151. Silva-Rodríguez, J., Colomer, A. & Naranjo, V. WeGleNet: A weakly-supervised convolutional neural network for the semantic segmentation of Gleason grades in prostate histology images. *Computerized Medical Imaging and Graphics* **88** (2021).

152. Lei, H., Liu, S., Elazab, A., Gong, X. & Lei, B. Attention-Guided Multi-Branch Convolutional Neural Network for Mitosis Detection from Histopathological Images. *IEEE Journal of Biomedical and Health Informatics* **25,** 358–370 (2021).

153. Choi, J. H. & Ro, J. Y. Cutaneous spindle cell neoplasms: pattern-based diagnostic approach. *Archives of pathology & laboratory medicine* **142,** 958–972 (2018).

154. Dolz, J., Desrosiers, C. & Ayed, I. B. *Teach me to segment with mixed supervision: Confident students become masters* in *International Conference on Information Processing in Medical Imaging (IPMI)* (2021).

155. Zhu, H., Shi, J. & Wu, J. *Pick-and-learn: automatic quality evaluation for noisy-labeled image segmentation* in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2019).

156. Zhang, T., Yu, L., Hu, N., Lv, S. & Gu, S. *Robust medical image segmentation from non-expert annotations with tri-network* in *International Conference on medical image computing and computer-assisted intervention (MICCAI)* (2020).

157. Codella, N. C. *et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)* in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)* (2018), 168–172.

158. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *nature* **542,** 115–118 (2017).

159. Wang, L. *et al. Medical Matting: A New Perspective on Medical Segmentation with Uncertainty* in *Medical Image Computing and Computer Assisted Intervention (MICCAI)* (2021).

160. Winnepenninckx, V., De Vos, R., Stas, M. & van den Oord, J. J. New phenotypical and ultrastructural findings in spindle cell (desmoplastic/neurotropic) melanoma. *Applied Immunohistochemistry & Molecular Morphology* **11,** 319–325 (2003).

161. Lai, V., Cranwell, W. & Sinclair, R. Epidemiology of skin cancer in the mature patient. *Clinics in dermatology* **36,** 167–176 (2018).

162. Xu, Z. *et al.* Spindle cell melanoma: Incidence and survival, 1973-2017. *Oncology letters* **16,** 5091–5099 (2018).

163. Ha Lan, T. T. *et al.* Expression of the p40 isoform of p63 has high specificity for cutaneous sarcomatoid squamous cell carcinoma. *Journal of cutaneous pathology* **41,** 831–838 (2014).

164. Ghesu, F. C. *et al.* Quantifying and leveraging predictive uncertainty for medical image assessment. *Medical Image Analysis* **68** (2021).

165. Arvaniti, E. *et al.* Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific Reports* **8,** 1–11 (2018).

166. Galdran, A., Dolz, J., Chakor, H., Lombaert, H. & Ben Ayed, I. *Cost-Sensitive Regularization for Diabetic Retinopathy Grading from Eye Fundus Images* in *Medical Image Computing and Computer Assisted Intervention (MICCAI)* (2020).

167. Ji, W. *et al. Learning Calibrated Medical Image Segmentation via Multi-rater Agreement Modeling* in *IEEE Conference on Computer Vision and Pattern Recoginition (CVPR)* (2021).

168. B, M. H. J., Jørgensen, D. R. & Jalaboi, R. *Improving Convolutional Neural Networks Using Inter-rater Agreement* in *Medical Image Computing and Computer Assisted Intervention (MICCAI)* (2019).

169. Jungo, A. *et al. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation* in *Medical Image Computing and Computer Assisted Intervention (MICCAI)* (2018).

170. Ju, L., Wang, X., Wang, L., Member, G. S. & Mahapatra, D. Improving medical images classification with label noise using dual-uncertainty estimation. *IEEE transactions on medical imaging* **41,** 1533–1546 (2022).

171. Zhang, L. *et al. Disentangling human error from the ground truth in segmentation of medical images* in *Advances in Neural Information Processing Systems (NeurIPS)* (2020).

172. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł. & Hinton, G. *Regularizing neural networks by penalizing confident output distributions* in *International Conference on Learning Representations (ICLR)* (2019).

173. Belharbi, S. *et al.* Deep Interpretable Classification and Weakly-Supervised Segmentation of Histology Images via Max-Min Uncertainty. *IEEE Transactions on Medical Imaging* **41,** 702–714. ISSN: 1558254X (2022).

174. Ayhan, M. S. *et al.* Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. *Medical Image Analysis* **64** (2020).

175. Guo, S. *et al. CurriculumNet: Weakly supervised learning from large-scale web images* in *Medical Image Computing and Computer Assisted Intervention (MICCAI)* (2018).

176. Han, B. *et al. Co-teaching: Robust training of deep neural networks with extremely noisy labels* in *Advances in Neural Information Processing Systems (NeurIPS)* (2018).

177. Jiang, L., Zhou, Z., Leung, T., Li, L. J. & Fei-Fei, L. *Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels* in *International Conference on Learning Representations (ICLR)* (2018).

178. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. *Rethinking the inception architecture for computer vision* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).

179. Liu, B., Ayed, I. B., Galdran, A. & Dolz, J. *The Devil is in the Margin: Margin-based Label Smoothing for Network Calibration* in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2022).

180. Meister, C., Salesky, E. & Cotterell, R. *Generalized Entropy Regularization or: There's Nothing Special about Label Smoothing* in *Annual Meeting of the Association for Computational Linguistics* (2020).

181. Ilse, M., Tomczak, J. M. & Welling, M. *Attention-based deep multiple instance learning* in *35th International Conference on Machine Learning (ICML)* (2018).

182. Grandvalet, Y. & Bengio, Y. *Semi-supervised Learning by Entropy Minimization* in (2004).

183. Boudiaf, M. *et al. Few-Shot Segmentation Without Meta-Learning: A Good Transductive Inference Is All You Need?* in (2021).

184. Simonyan, K. & Zisserman, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition* in *International Conference on Learning Representations (ICLR)* (2014).

185. Deng, J. *et al. ImageNet: A Large-Scale Hierarchical Image Database* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2009).