UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Deep learning strategies for histological image retrieval

Author:
Zahra TABATABAEI

Supervisors:
Prof. Valery NARANJO
Dr. Adrian COLOMER
Dr. Javier OLIVER MOLL

Electronic Engineering Department
Technologies for Health and Well-Being

June, 2, 2024

Declaration of Authorship

I, Zahra TABATABAEI, declare that this thesis titled, "Deep learning strategies for histological image retrieval" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date: June, 2, 2024

Abstract

According to the World Health Organization (WHO), cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020, or about one in six deaths. To prevent and decrease this huge amount of death, an accurate cancer diagnosis is necessary. Deep Learning (DL)-based techniques have advanced Computer-Aided Diagnosis (CAD) to assist doctors with their diagnosis. High-resolution images, such as histopathological slides and medical scans, enhance these techniques. This thesis mainly focuses on histopathological images scanned by the Whole Slide Image (WSI) scanners, aiming to minimize human errors in cancer diagnosis. In this thesis, we propose three Content-Based Medical Image Retrieval (CBMIR) frameworks on histopathological images with two DL-based techniques presented in different scenarios. SICAPv2, BreaKHis, Camelyon 17, Arvaniti, and Spitziod are the main data sets that the experiments are conducted on. These data sets are patches of prostate, breast, breast, prostate, and skin cancer, respectively. Our studies are categorized into three classes based on learning types. Initially, we develop an unsupervised Feature Extractor (FE) for extracting meaningful features from prostate and breast cancer datasets, achieving 70% and 91% accuracy in prostate (SICAPv2) and breast cancer (BreaKHis) datasets, respectively, at the top 5.

In the other contribution of this class of study, we mainly focused on Color Normalization (CN) as a pre-processing step in a Content-Based Histopathological Image Retrieval (CBHIR) framework. The obtained results reported that as the effectiveness of color normalization techniques in reducing intra-center variance improved, the CBHIR results exhibited higher performance levels. The state-of-the-art color normalization technique employed on the patches provides an 18% improvement in accuracy.

In the first class of our studies, we identified potential obstacles that a CBMIR in digital pathology could encounter, including limited power of GPU resources, lack of enough data set, strict data privacy regulations for data sharing, etc.

Regarding these complexities, we focus on federated-based learning in the second class of our research. We combine the concepts of Federated Learning (FL) with a CBMIR framework to mimic a worldwide Federated CBMIR (FedCBMIR) on histological images of breast cancer. This research explores three scenarios. In the first scenario, two medical centers have different breast cancer datasets (Camelyon 17 and BreaKHis 400×) and cannot share their images due to data privacy. This scenario meets pathologists' needs and speeds up training for engineers, reducing training time by 11.44 hours. Pathologists can achieve 97.8% accuracy for BreaKHis and 98.1% for Camelyon 17.

The assumption behind the second scenario is that there are four data sets of breast cancer, and the centers don't have any agreement to share their data sets. Therefore, four well-trained FEs are trained more generalized and in 32.36 hours shorter time than training four distinct models. The proposed FedCBMIR framework can overpass the accuracy of the CBMIR that is training locally with 96%, 92%, 89%, and 94% precision respectively for BreaKHis at 40×, 100×, 200×, and 400×.

The last scenario in this class of study focuses on the condition that pathologists need to measure some important patterns at different magnifications of their tissue although they don't have access to high-level scanners. This can army the pathologist with a tool that they can see the patterns of their query in the other similar retrieved patches at different levels of magnification.

In the last contribution of this thesis, the main focus is a contrastive learning-based strategy. We propose a CBMIR framework that can overpass the previous techniques with top K ($K > 1$) and excels in retrieving images at the top first. This can be a more reliable framework for pathologists since it can provide similar patches to their query even at the first top. It is the first CBHIR framework proposed for Spitzoid cancer, achieving 70% and 81% F1 score in BreaKHis at 400× and skin cancer, respectively. The proposed technique's precision at the top rank for skin cancer is 67% higher than previous methods.

Furthermore, this class of study solves the challenges that pathologists have in grading Spitzoid Tumors of Uncertain Malignant Potential (STUMP). STUMP cases require careful assessment to determine their true nature. To assist pathologists in coping with this complexity, the framework can provide top K similar patches for them with their corresponding labels. Our framework assists pathologists by providing top K similar patches with corresponding labels, allowing them to grade their STUMP query by examining the retrieved images and their histological patterns.

To conclude, the proposed CBMIR and CBHIR frameworks in this thesis contribute to the diagnosis of prostate, breast, and skin cancer from histopathological images by making use of DL-based FEs under different scenarios. These not only enhance the accuracy and efficiency of cancer diagnosis but also hold promise for facilitating early detection and personalized treatment strategies. Leveraging these frameworks in the current cancer diagnosis could ultimately lead to improved patient outcomes, reduced healthcare costs, and enhanced quality of life for individuals affected by prostate, breast, and skin cancer. These advancements have the potential to drive positive societal change and contribute to the global fight against cancer.

Resumen

Según Organización Mundial de la Salud (WHO), el cáncer es una de las principales causas de muerte a nivel mundial, con cerca de 10 millones de fallecimientos en 2020, aproximadamente una de cada seis muertes. Para prevenir y disminuir esta enorme cantidad de muertes, es necesario un diagnóstico preciso del cáncer. Las técnicas basadas en Deep Learning (DL) han avanzado el Diagnóstico Asistido por Computadora (CAD) para ayudar a los médicos con su diagnóstico. Las imágenes de alta resolución, como las láminas histopatológicas y las exploraciones médicas, mejoran estas técnicas. Esta tesis se centra principalmente en imágenes histopatológicas escaneadas por escáneres de Whole Slide Image (WSI), con el objetivo de minimizar los errores humanos en el diagnóstico del cáncer. En esta tesis, proponemos tres marcos de Recuperación de Imágenes Médicas Basada en Contenido (CBMIR) sobre imágenes histopatológicas con dos técnicas basadas en DL presentadas en diferentes escenarios. SICAPv2, BreaKHis, Camelyon 17, Arvaniti y Spitziod son los principales conjuntos de datos en los que se realizan los experimentos. Estos conjuntos de datos son parches de cáncer de próstata, mama, mama, próstata y piel, respectivamente. Nuestros estudios se categorizan en tres clases según los tipos de aprendizaje. Inicialmente, desarrollamos un Extractor de Características (FE) no supervisado para extraer características significativas de los conjuntos de datos de cáncer de próstata y mama, logrando una precisión del 70% y 91% en los conjuntos de datos de cáncer de próstata (SICAPv2) y mama (BreaKHis), respectivamente, en el top 5.

En la otra contribución de esta clase de estudio, nos centramos principalmente en la Normalización del Color (CN) como un paso de preprocesamiento en un marco de Recuperación de Imágenes Histopatológicas Basada en Contenido (CBHIR). Los resultados obtenidos informaron que a medida que la efectividad de las técnicas de normalización del color en la reducción de la variabilidad intra-centro mejoró, los resultados del CBHIR mostraron niveles de rendimiento más altos. La técnica de normalización del color de última generación empleada en los parches proporciona una mejora del 18% en la precisión.

En la primera clase de nuestros estudios, identificamos obstáculos potenciales que un CBMIR en patología digital podría encontrar, incluida la limitación del poder de los recursos de GPU, la falta de suficientes conjuntos de datos, las estrictas regulaciones de privacidad de datos para el intercambio de datos, etc.

En cuanto a estas complejidades, nos centramos en el aprendizaje federado en la segunda clase de nuestra investigación. Combinamos los conceptos de Federated Learning (FL) con un marco CBMIR para imitar un CBMIR Federado Mundial (FedCBMIR) en imágenes histológicas de cáncer de mama. Esta investigación explora tres escenarios. En el primer escenario, dos centros médicos tienen diferentes conjuntos de datos de cáncer de mama (Camelyon 17 y BreaKHis 400×) y no pueden compartir sus imágenes debido a la privacidad de los datos. Este escenario satisface las necesidades de los patólogos y acelera el entrenamiento para los ingenieros, reduciendo el tiempo de entrenamiento en 11,44 horas. Los patólogos pueden lograr una precisión del 97,8% para BreaKHis y del 98,1% para Camelyon 17.

La suposición detrás del segundo escenario es que hay cuatro conjuntos de datos de cáncer de mama y los centros no tienen ningún acuerdo para compartir sus conjuntos de datos. Por lo tanto, se entrenan cuatro FEs bien entrenados de manera más generalizada y en 32,36 horas menos que entrenar cuatro modelos distintos. El marco FedCBMIR propuesto puede superar la precisión del CBMIR que se entrena localmente con un 96%, 92%, 89% y 94% de precisión respectivamente para BreaKHis en 40×, 100×, 200× y 400×. El último escenario en esta clase de estudio

se centra en la condición de que los patólogos necesitan medir algunos patrones importantes a diferentes aumentos de su tejido aunque no tengan acceso a escáneres de alto nivel. Esto puede armar al patólogo con una herramienta que les permita ver los patrones de su consulta en los otros parches similares recuperados a diferentes niveles de aumento.

En la última contribución de esta tesis, el enfoque principal es una estrategia basada en aprendizaje contrastivo. Proponemos un marco CBMIR que puede superar las técnicas anteriores con el top K ($K > 1$) y sobresale en la recuperación de imágenes en el top primero. Esto puede ser un marco más confiable para los patólogos ya que puede proporcionar parches similares a su consulta incluso en el primer top. Es el primer marco CBHIR propuesto para el cáncer de Spitzoid, logrando un 70% y 81% de puntuación F1 en BreaKHis a 400× y cáncer de piel, respectivamente. La precisión de la técnica propuesta en el rango superior para el cáncer de piel es un 67% más alta que los métodos anteriores.

Además, esta clase de estudio resuelve los desafíos que los patólogos tienen al clasificar los Tumores Spitzoides de Potencial Maligno Incierto (STUMP). Los casos de STUMP requieren una evaluación cuidadosa para determinar su verdadera naturaleza. Para ayudar a los patólogos a enfrentar esta complejidad, el marco puede proporcionar los parches similares del top K para ellos con sus etiquetas correspondientes. Nuestro marco ayuda a los patólogos proporcionando los parches similares del top K con las etiquetas correspondientes, permitiéndoles clasificar su consulta de STUMP examinando las imágenes recuperadas y sus patrones histológicos.

En conclusión, los marcos CBMIR y CBHIR propuestos en esta tesis contribuyen al diagnóstico del cáncer de próstata, mama y piel a partir de imágenes histopatológicas mediante el uso de FEs basados en DL en diferentes escenarios. Estos no solo mejoran la precisión y eficiencia del diagnóstico del cáncer, sino que también tienen el potencial de facilitar la detección temprana y las estrategias de tratamiento personalizado. Aprovechar estos marcos en el diagnóstico actual del cáncer podría conducir en última instancia a mejores resultados para los pacientes, menores costos de atención médica y una mayor calidad de vida para las personas afectadas por el cáncer de próstata, mama y piel. Estos avances tienen el potencial de impulsar un cambio social positivo y contribuir a la lucha global contra el cáncer.

Resum

Segons l'Organització Mundial de Salut (WHO), el càncer és una de les principals causes de mort a nivell mundial, amb prop de 10 milions de defuncions en 2020, aproximadament una de cada sis morts. Per a prevenir i disminuir aquesta enorme quantitat de morts, és necessari un diagnòstic precís del càncer. Les tècniques basades en Deep Learning (DL) han avançat el Diagnòstic Assistit per Computadora (CAD) per a ajudar els metges amb el seu diagnòstic. Les imatges d'alta resolució, com les làmines histopatològiques i les exploracions mèdiques, milloren aquestes tècniques. Aquesta tesi se centra principalment en imatges histopatològiques escanejades per escàners de Whole Slide Image (WSI), amb l'objectiu de minimitzar els errors humans en el diagnòstic del càncer. En aquesta tesi, proposem tres marcs de Recuperació d'Imatges Mèdiques Basada en Contingut (CBMIR) sobre imatges histopatològiques amb dues tècniques basades en DL presentades en diferents escenaris. SICAPv2, BreaKHis, Camelyon 17, Arvaniti i Spitziod són els principals conjunts de dades en els quals es realitzen els experiments. Aquests conjunts de dades són trossos de càncer de pròstata, mama, mama, pròstata i pell, respectivament. Els nostres estudis es categoritzen en tres classes segons els tipus d'aprenentatge. Inicialment, desenvolupem un Extractor de Característiques (FE) no supervisat per a extraure característiques significatives dels conjunts de dades de càncer de pròstata i mama, aconseguint una precisió del 70% i 91% en els conjunts de dades de càncer de pròstata (SICAPv2) i mama (BreaKHis), respectivament, en el top 5.

En l'altra contribució d'aquesta classe d'estudi, ens centrem principalment en la Normalització del Color (CN) com un pas de preprocessament en un marc de Recuperació d'Imatges Histopatològiques Basada en Contingut (CBHIR). Els resultats obtinguts van informar que a mesura que l'efectivitat de les tècniques de normalització del color en la reducció de la variabilitat intra-centre va millorar, els resultats del CBHIR van mostrar nivells de rendiment més alts. La tècnica de normalització del color d'última generació emprada en els trossos proporciona una millora del 18% en la precisió.

En la primera classe dels nostres estudis, identifiquem obstacles potencials que un CBMIR en patologia digital podria trobar, inclosa la limitació del poder dels recursos de GPU, la falta de suficients conjunts de dades, les estrictes regulacions de privacitat de dades per a l'intercanvi de dades, etc.

Quant a aquestes complexitats, ens centrem en l'aprenentatge federat en la segona classe de la nostra investigació. Combinem els conceptes de Federated Learning (FL) amb un marc CBMIR per a imitar un CBMIR Federat Mundial (FedCBMIR) en imatges histològiques de càncer de mama. Aquesta investigació explora tres escenaris. En el primer escenari, dos centres mèdics tenen diferents conjunts de dades de càncer de mama (Camelyon 17 i BreaKHis 400×) i no poden compartir les seues imatges a causa de la privacitat de les dades. Aquest escenari satisfà les necessitats dels patòlegs i accelera l'entrenament per als enginyers, reduint el temps d'entrenament en 11,44 hores. Els patòlegs poden aconseguir una precisió del 97,8% per a BreaKHis i del 98,1% per a Camelyon 17.

La suposició darrere del segon escenari és que hi ha quatre conjunts de dades de càncer de mama i els centres no tenen cap acord per a compartir els seus conjunts de dades. Per tant, s'entrenen quatre FEs ben entrenats de manera més generalitzada i en 32,36 hores menys que entrenar quatre models diferents. El marc FedCBMIR proposat pot superar la precisió del CBMIR que s'entrena localment amb un 96%, 92%, 89% i 94% de precisió respectivament per a BreaKHis en 40×, 100×, 200× i 400×. L'últim escenari en aquesta classe d'estudi se centra en la condició que els

patòlegs necessiten mesurar alguns patrons importants a diferents augments del seu teixit encara que no tinguen accés a escàners d'alt nivell. Això pot armar el patòleg amb una eina que els permet veure els patrons de la seua consulta en els altres trossos similars recuperats a diferents nivells d'augment.

En l'última contribució d'aquesta tesi, l'enfocament principal és una estratègia basada en aprenentatge contrastiu. Proposem un marc CBMIR que pot superar les tècniques anteriors amb el top K ($K > 1$) i sobresurt en la recuperació d'imatges en el top primer. Això pot ser un marc més fiable per als patòlegs ja que pot proporcionar trossos similars a la seua consulta fins i tot en el primer top. És el primer marc CBHIR proposat per al càncer de Spitzoid, aconseguint un 70% i 81% de puntuació F1 en BreaKHis a 400× i càncer de pell, respectivament. La precisió de la tècnica proposada en el rang superior per al càncer de pell és un 67% més alta que els mètodes anteriors.

A més, aquesta classe d'estudi resol els desafiaments que els patòlegs tenen a l'hora de classificar els Tumors Spitzoides de Potencial Maligne Incert (STUMP). Els casos de STUMP requereixen una avaluació acurada per a determinar la seua veritable naturalesa. Per a ajudar els patòlegs a afrontar aquesta complexitat, el marc pot proporcionar els trossos similars del top K per a ells amb les seues etiquetes corresponents. El nostre marc ajuda els patòlegs proporcionant els trossos similars del top K amb les etiquetes corresponents, permetent-los classificar la seua consulta de STUMP examinant les imatges recuperades i els seus patrons histològics.

En conclusió, els marcs CBMIR i CBHIR proposats en aquesta tesi contribueixen al diagnòstic del càncer de pròstata, mama i pell a partir d'imatges histopatològiques mitjançant l'ús de FEs basats en DL en diferents escenaris. Aquests no sols milloren la precisió i eficiència del diagnòstic del càncer, sinó que també tenen el potencial de facilitar la detecció primerenca i les estratègies de tractament personalitzat. Aprofitar aquests marcs en el diagnòstic actual del càncer podria conduir en última instància a millors resultats per als pacients, menors costos d'atenció mèdica i una major qualitat de vida per a les persones afectades pel càncer de pròstata, mama i pell. Aquests avanços tenen el potencial d'impulsar un canvi social positiu i contribuir a la lluita global contra el càncer.

Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

Accuracy: ACC
Artificial Intelligence: AI
Area Under the Curve: AUC
Auto-Encoders: AE
Bayesian K- Singular Value Decomposition: BKSVD
Bag-Of-Features: BOF
CAMELYON17 challenge: CAM17
CLoud ARtificial Intelligence For pathologY: CLARIFY
Color Normalization: CN
Convolutional Neural Network: CNN
Computer Aide Diagnosis: CAD
Confusion Matrix: CM
Content-Based Histopathological Image Retrieval: CBHIR
Content-Based Medical Image Retrieval: CBMIR
Convolutional Auto Encoder: CAE
Deep Convolutional Gaussian Mixture Model: DCGMM
Deep Learning: DL
Feature Extractor: FE
Federated Content-Based Medical Image Retrieval: FedCBMIR
Federated Learning: FL
F1Score: F1S
Generative Adversarial Networks: GAN
Hematoxylin and Eosin: H&E
Immunohistochemistry: IHC
Magnetic Resonance Imaging: MRI
Macenko: Mac
Normalized Median Intensity: NMI
Normalized Median Intensity Standard Deviation: NMI SD
Normalized Median Intensity Coefficient of Variation: NMI CV
Peak Signal to Noise Ratio: PSNR
Regions Of Interest: ROI
Singular Value Decomposition: SVD
Tensorflow Federated: TFF
Unsupervised Features Learning: UFL
Variational Auto Encoder: VAE
Vahadane: Vah
Whole Slide Images: WSIs

To my lovely parents, whose boundless love and unwavering support have been my guiding light throughout the journey of life. Despite the miles that separate us, your smiles have bridged the distance, filling my heart with warmth and courage. Your unwavering belief in me has been the foundation upon which I have built my dreams. This thesis is a tribute to your enduring love, resilience, and sacrifice, which have shaped me into the person I am today. . . .

Chapter 1

Introduction

In this chapter, the motivation and the main contributions of this PhD thesis are introduced. This chapter also includes the thesis outline.

1.1 Motivation

Artificial Intelligence (AI) was founded at a workshop held on the campus of Dartmouth College, USA in the summer of 1956 [1]. Since then, AI has been developed in a wide diversity of research fields including medical purposes and finance to autonomous vehicles and recommendation systems. Xu et al. in [2] declared that the application of AI-based tools in healthcare can enhance human abilities and improve the process of medical treatment. Since the 1970s, with the development of AI tools for data collection and storage, plenty of hospitals have been provided with well-equipped systems to gather and share a large amount of information. Despite these advancements, many clinics and hospitals still suffer from digitalization problems and do not have access to enough equipment to join the world of digital pathology. However, the available amount of data can be useful for both clinicians and AI experts to train their models and provide methods with higher accuracy for clinicians. Since then, a wide diversity of research has been dedicated to this area of study [3]. Over time, in the case of more complicated use cases, Deep Learning (DL)-based techniques play a pivotal role in enhancing AI's ability to automate tasks and analyze the images to make predictions. So, this can advance AI's problem-solving capabilities. This feature of DL-based techniques can make a bridge between AI and the medical domain. DL-based techniques leverage sophisticated neural network architectures to automatically extract intricate patterns and features from complex medical images, such as MRI scans, CT scans, and histopathological slides. By learning hierarchical representations directly from data, these tools enhance the accuracy and efficiency of disease detection, aiding healthcare professionals in making informed decisions. This can reduce the rate of human errors in cancer diagnosis.

Human errors can range from misinterpretations of subtle patterns to oversight in extensive data sets, potentially leading to misdiagnoses. For instance, assume that a pathologist has to review 100 WSIs in a day. The first 70 WSIs were benign and the 71th WSI is a tissue in grade 2 [4]. Since his eyes used to see normal biopsies, by seeing the abnormalities he might grade the tissue as malignant although it is not. This rate of human errors might be decreased by the experiences that pathologists have. By harnessing human errors, patient care in the realm of cancer diagnosis can be improved. In traditional cancer diagnosis, there is a reference book (a.k.a. Atlas [1]) which contains some samples for cancer types and includes sections of risk factor and taking action. In the cases where pathologists cannot conclude the grade of cancer, they can refer to this book and compare the histopathological patterns of their query with patterns in the images of Atlas [5]. Consequently, they can make up their decision on the grade of their query. Although this assists pathologists, it is time-consuming and it has limited cases that might not be enough for the pathologist.

An alternative approach in the current cancer diagnosis in complicated cases is that pathologists send their tissue to other centers that might be located in different countries or other cities to consult with their peers [6]. This workflow might increase the accuracy of grading and diagnosis of cancer but it is time-consuming, expensive, and risky. Some accidents might occur to the tissue such as breaking, losing, etc. To prevent these troubles, a Content-Based Histopathological Image Retrieval (CBHIR) mechanism mimics this workflow by developing an accurate Feature Extractor (FE) and a high-performance search engine [7]. CBHIR not only searches for images but also compares the different types of histopathological images based on their patterns.

[1] https://canceratlas.cancer.org

The main motivation of this thesis is to propose a CBHIR framework to increase the accuracy of cancer diagnosis, speed up the searching workflow, and decrease the workload of pathologists. In the current thesis, we propose some CBHIR frameworks to assist pathologists in the diagnosis of prostate, breast, and skin cancer to help pathologists in decision-making. The proposed models train with data sets that were stained by Hematoxylin and Eosin (H&E) which provide pink and purple images [8]. In this Ph.D. thesis, we propose different cutting-edge DL-based CBHIR frameworks that improve the accuracy of cancer diagnosis. To explore the potential use case of the proposed methods in diagnostic imaging, we address several scenarios to cover the challenges that pathologists face while analyzing the tissue.

In the second chapter of this thesis, we propose a novel CBHIR framework that mainly focuses on prostate cancer and breast cancer as multi-class and binary data sets. In Chapter 3, we propose to employ a CN technique as a pre-processing step in a CBHIR framework as a result of analyzing the obtained results. With the advent of running DL-based techniques on GPUs, a new range of promising possibilities opened a new range of potential approaches that can increase the feasibility of using DL techniques. In Chapter 4, we go further to tackle the difficulties related to storage problems, resource limitations, and medical data privacy thanks to GPUs and Federated Learning (FL). In the last chapter (Chapter 5) of this thesis, we propose a novel FE capable of returning the top first similar patches with high accuracy. This can provide a more confidential tool for the pathologist and there are no other studies that can have high performance at top first retrieved images. In this chapter, the proposed CBHIR framework can assist pathologists in grading the STUMP tissue of skin cancer data set which was not done in other studies.

1.1.1 Big players

In the world of digital pathology, doctors and researchers from different countries come together to collaborate to improve the cancer diagnosis workflow. To do so, the CBHIR framework brings computer scientists, doctors, and industry experts into one room to solve complex medical puzzles. This represents a paradigm shift in medical image analysis, particularly in the context of pathology. Companies and research centers can leverage CBHIR to encourage collaboration and knowledge sharing among researchers, clinicians, and data scientists, creating platforms that facilitate collaborative efforts. CBHIR offers data management to companies by reducing reliance on text-based queries and streamlining the retrieval process. Also, the integration of cross-modal retrieval, where information is retrieved from different modalities, extends the advancement of CBHIR beyond conventional image retrieval, offering innovative solutions for pathological diagnosis and research. This integration and utilization of the histology foundation models underscore the potential for groundbreaking advancements in medical imaging and pathology. So, the interest in CBHIR is a driving force for both researchers and companies and it makes an opportunity for the collaboration between academia and industry.

According to Scopus search results by considering **content** AND **based** **medical** AND **image** AND **retrieval** as the keywords from 2013 to 2023 (a period of 10 years), there are 921 journal and conference papers in English. We limited the subject area to **computer science**, **engineering**, and **medicine**. This illustrates that the recent impact of CBHIR is reflected in academia through the vast number of publications.

CBHIR is a search engine for medical images which motivates big players in the research world, such as top universities and institutes, to dive deep into this topic. Kimia lab [2] at the University of Waterloo, in Canada is one of the research centers which dedicated many pieces of research to this topic [9, 10, 11, 12, 13, 14]. CBHIR is not just about finding similar patches but also about providing similar histopathological patterns for pathologists for a better and faster cancer diagnosis. This cross-disciplinary framework can deepen the understanding of diseases and foster innovative solutions.

Philips [3] and Google Health [4] are playing a role as a big player in this field aiming to assist pathologists through DL-based techniques. The other important company in the field of digital pathology is Bigpicture[5]. This company has a dedicated part in the field of CBHIR and federated learning.

The interest in CBHIR extends to its practical applications in clinical settings, facilitating more accurate and efficient pathology diagnosis. Companies specializing in medical imaging solutions, such as Siemens Healthineers[6] and GE Healthcare[7], recognize the transformative impact of CBHIR on streamlining diagnostic workflows.

1.2 Clinical use case

The primary data set of this thesis is cancerous images since according to the reports from the World Health Organization (WHO), every year, the world is witness to more than 6 million cancer deaths of the 10 million new cases each year. The point is that more than half occurs in developing countries since the developing countries succeed in achieving lifestyles similar to Europe, North America, Australia, New Zealand, and Japan, they will also encounter much higher cancer rates, particularly cancers of the breast, colon, prostate, and uterus (endometrial carcinoma) [15]. Based on the estimations reported by WHO, these figures will only worsen in the next twenty years; increasing to 10 million deaths and 15 million new cases annually.

Among men, prostate cancer is largely seen in more developed countries. Among women, breast cancer is one of the most prevalent cancer types with 2.3 million women diagnosed with this cancer and 685,000 deaths globally in 2020 [16]. In both genders, over 1.5 million skin cancers are diagnosed every year as a result of the ultraviolet radiation from the sun. This means that one in every three cancers diagnosed is a skin cancer and as ozone levels are depleted, it is estimated that a 10% decrease in ozone level will result in an additional 300,00 non-melanoma and 4,500 melanoma skin cancer cases [17]. Based on these reports, breast, prostate, and skin cancer are chosen as the main cancer types that this thesis focuses on [18]. We select these cancer types since they are more prevalent and we want to focus on gender-based and non-gender-based cancers. As far as the rate of patients struggling with cancer is increasing, the role of digital pathology and CAD tools become more highlighted. So, in this thesis, we proposed CBHIR approaches with different frameworks to assist pathologists and increase the diagnosis accuracy to decrease the rate of death due to cancer [19].

The primary users of the proposed approach in this thesis are pathologists with different levels of expertise, catering to their varied needs in diagnosing cancerous tissues. In addition to the clinicians, this framework extends its benefits to society

[2] https://tizhoosh.com/labs/kimia-lab/

[3] https://www.philips.com.au/healthcare/solutions/pathology

[4] https://health.google/consumers/search/

[5] https://www.linkedin.com/company/bigpicture-project/

[6] https://www.siemens-healthineers.com/

[7] https://www.gehealthcare.com/

and the patient indirectly. There are plenty of use cases to demonstrate how the CBHIR enhances clinical workflows, supports decision-making, and contributes to advancements in cancer diagnosis and treatment. Some of these use cases are as follows:

- CBHIR assists pathologists in diagnosing cancer by retrieving top K images with similar histopathological patterns. This helps pathologists, dermatologists, and radiologists identify the extent and nature of abnormalities.
- CBHIR facilitates medical research by providing quick access to a diverse set of cancer images. This aims to assist clinicians in tracking disease progression over time by comparing the query tissue with historical cases.
- CBHIR supports early detection of cancer by retrieving similar visual histopathological patterns.
- CBHIR allows pathologists to explain the condition of the tissue to the patient with the aim of patient education.
- CBHIR promotes and facilitates the decision-making of multidisciplinary teams, allowing experts from all over the world and in different centers to analyze the images and contribute to patient care.

In addition to the above-mentioned benefits of CBHIR, retrieving top K similar patches to the query is an automatic Atlas book while different medical centers can adjust the amount of K based on their requirements. Furthermore, the flexibility of CBHIR in retrieving K-similar patches makes it more comprehensive instead of focusing on only one top result or resulting only with a label. The most powerful property of this kind of approach is the transparency and feedback that it provides to specialists instead of a categorical classification. This benefit might serve as an educational tool for medical professionals, allowing professors to teach medical students by providing some comparable cases with their queries.

It is noteworthy to mention that the expertise and experience of pathologists play a significant role in the accuracy and reliability of cancer diagnosis and tissue grading. So, newly graduated pathologists not only need more time to diagnose the cancer grade and cancer type but also need more samples to be able to compare the histopathological patterns from the Atlas book or any other reference books. CBHIR offers similar patches to their query in a shorter time and from all the collaborative centers. Therefore, it can transfer the knowledge and the guidance provided by seasoned experts contributes to the continuous development of skills in the field.

Moreover, not only medical students can learn better and deeper under the umbrella of CBHIR, but researchers can also reach a rich source of annotated images for the development and validation of new CAD methods in order to improve diagnostic methods and treatment options.

Apart from all the benefits of CBHIR in digital pathology, it brings some benefits for the patients besides they can get better diagnosis and treatment. The anxiety and stress that the person suffering from cancer is tolerating might affect his family, society, his performance in his job, etc. The case that the time of diagnosis decreases this stress and pressure on the person can be diminished and all the consequences.

In essence, the application of this CBHIR transcends individual medical practices, creating a ripple effect that positively influences broader healthcare dynamics and societal well-being.

1.3 Main objective

The main objective of this Ph.D. thesis is to research on deep learning strategies for CBHIR. Several DL-based frameworks are designed, developed and validated to assist experts in diagnosing and grading a tissue. We aim to address the challenges that pathologists face during cancer detection and grading by offering a CBHIR technique that works as an automatic Atlas book. In particular, we set our sights on DL-based approaches to propose CBHIR frameworks for the searching task. This thesis intends to frame these approaches covering different learning paradigms and scenarios to reply to the demands of pathologists. In pursuit of this, we aim to propose new CBHIR frameworks based on DL algorithms to reach the deeper features of tissue in order to find similar histopathological patterns for pathologists. To do so, we categorize our investigations into unsupervised learning, federated learning, and contrastive learning methodologies. In this thesis, we conduct an extensive comparison between these types of learning to reach the framework that excels in terms of training time and accuracy, catering to the needs of both pathologists and engineers. This framework provides clinicians with a search engine that assists them in characterizing complicated and challenging cases.

In this thesis, we propose a search engine that can solve computer vision and medical challenges using DL-based algorithms. We aim to provide CBHIR frameworks for the clinician's demands with more explainable DL-based methods in terms of transparency, trustworthiness, and reliability for pathologists. In particular, the main objectives of this thesis are as follows:

- Proposing an unsupervised approach for detection of similar patterns on binary and multi-class data sets;
- Studying a CBHIR framework with an unsupervised feature extractor that takes Color Normalization (CN) into account as a pre-processing step. This draws attention to the relevance of color variation and its impact on CBHIR;
- Proposing a novel international FL-based Content-Based Medical Image Retrieval (CBMIR). This is a decentralized and confidential unsupervised technique on varying data sets distributed among individual clients;
- Designing and implementing a custom-built Siamese network to address inter-class variations and large intra-class variances by applying contrastive loss;
- Formulating a CBHIR approach to tackle the challenges in grading Spitzoid Tumors of Uncertain Malignant Potential (STUMP) by providing deep insights into the complexities.

To deal with these main objectives, we split the goals into four chapters. In each of these chapters, a new CBHIR framework with different scenarios is presented. The main cancer types that these frameworks target are breast, prostate, and skin cancer.

Our investigations on unsupervised learning techniques propose a fully unsupervised FE conducted on prostate and breast cancer to tackle the lack of annotated data and provide top K similar images for pathologists. Then, these studies move further to offer a pre-processing step on patches, aiming to assess and analyze the effects of various CN techniques on the ultimate outcomes of CBMIR.

In the federated learning-based studies, a worldwide CBHIR framework is proposed for breast cancer by leveraging FL into the CBHIR tool. Specifically, we propose a novel CBHIR framework that trains with distributed data sets to tackle the lack of enough annotated data sets and data storage. To the best of our knowledge, there are no other studies that offer a Federated CBHIR on breast cancer.

Contrastive learning-based studies mainly focused on retrieving the first top similar images with high accuracy for breast and skin cancer data sets. There are zero studies with this high performance at the first top retrieval to the best of the author's knowledge. The second goal of contrastive learning-based techniques is to retrieve similar histopathological patterns for the STUMP tissues of skin cancer which is the first study in this area to the best of our belief.

In short, this thesis aims to propose a novel approach that can integrate CAD DL-based systems into daily clinical practice. The novel proposed strategies were developed to facilitate cancer diagnosis and grading for pathologists without requiring a huge search time or a laborious annotation process.

To achieve these objectives, each proposed approach in each chapter must undergo a sequential protocol based on the CRISP-ML(Q) methodology [20] such as below:

- **Data understanding:** An in-depth exploration of the histopathological patterns of prostate, breast, and skin cancer is needed. Similarly, a deep analysis of cancer diagnosis workflow in medical centers should be performed with the aim of understanding the pathologists' challenges in daily work.

- **Data pre-processing:** Each of the data sets corresponding to each particular cancer type should be pre-processed before feeding to the proposed DL-based CBHIR tools. Histopathological images mainly suffer from color variation due to the staining process. This might affect the final results of a DL-CBHIR which has to be tackled.

- **Modeling:** This stage includes the cutting-edge design and development of innovative DL-based CBHIR frames to reply to the demands of pathologists. A wide diversity of paradigms should be taken into account to provide the most sufficient technique according to the histopathological features of each cancer type.

- **Evaluation:** In this stage to compare the performance of the proposed DL-structures, the state-of-the-art methods should be explored in terms of quantitative and qualitative.

Monitoring and Maintenance as the last steps of the CRISP-ML(Q) methodology are not presented in this PhD thesis.

1.4 Main contributions

As mentioned above, our primary focus has been dedicated to prostate, breast, and skin cancer within different scenarios to assist pathologists in the cancer diagnosis process in terms of speed, accuracy, feasibility, decision-making, etc. These cancer types were selected because prostate and breast cancer are two leading causes of cancer death in men and women as the most common types of cancers in the entire population [21, 22]. Also, according to the World Health Organization, nearly one in three diagnosed cancers worldwide is a skin cancer [23]

CAMELYON17 challenge (CAM17), BreaKHis, SICAPv2, and skin CLARIFY are the histopathological images in the used data sets. The images in these data sets were stained with H&E to clarify the histopathological patterns of the patches.

In the course of this Ph.D., DL-based techniques are applied through these data sets to propose different CBMIR frameworks. These histopathological images are chosen because of their prevalence in cancerous patches.

In [24], we proposed an unsupervised CBHIR on prostate cancer using deep learning approaches to find similar histopathological patterns to the query in the largest pixel-annotated prostate data set [3]. In [25], we carried out a comparison between different evaluation techniques in the CBHIR studies to identify the best way of performance evaluation of CBHIR tools. This study was conducted on prostate and breast cancer data sets in order to report a promising comparison with the previous studies. In [26], our proposed unsupervised DL-CBHIR algorithm faces the effects of color variation of histopathological images on CBHIR performance. To do so, three different color normalization techniques were employed in the pre-processing stage to do some experiments on breast cancer data sets. In [27], we extended our DL knowledge by leveraging FL concepts to the CBHIR framework, for the first time, to train the DL-model with distributed breast cancer data sets. In [28] we conducted our experiments with a self-supervised classification tool to tackle the challenges of grading prostate cancer specifically into G3 and G4. Further on, we conducted distance-based training of algorithms on breast and skin cancer data sets in [29]. As a novelty in [29], for the first time, we explored the potential solution of grading STUMP tissues as one of the most challenging cancer types of skin cancer.

In the following subsection, we are going to provide a condensed overview of the technical contributions from each chapter, focusing on the learning strategies addressed in each paper.

1.4.1 *Unsupervised learning strategy*

In Chapter 2, we propose an Unsupervised Content-Based Medical Image Retrieval (UCBMIR) tool in an unsupervised manner for prostate and breast gradation problems that achieved comparable performance to fully supervised methods. We demonstrate our experiments on the largest pixel-was annotated prostate data set [3], BreaKHis, and Arvaniti. The proposed innovative approach addresses the problem of a lack of a pool of annotated images in DL-CBHIR tools. To do so, a custom-built Convolutional Auto Encoder (CAE) is developed to extract the meaningful and deep features of the patches. Euclidean distance is the similarity measure function that is applied to the extracted features to find similar patches and rank them. By reviewing the state-of-the-art studies, we conclude that there are two widely used numerical techniques in CBMIR and visual evaluations. Therefore, we measured the performance of the proposed technique with the numerical techniques and the visual evaluations. Furthermore, to determine if the retrieved images belong to the same cancer grade as the query we made a comparison with some cutting-edge classification studies. The main aim of this comparison was to evaluate the performance of the proposed CBHIR tool in discriminating different grades. To have a comprehensive evaluation, an external evaluation demonstrated the performance and generalization of UCBMIR, by training the model on SICAPv2 and testing it on the Arvaniti data set.

In Chapter 3, following the approach outlined in [30], three CN techniques including classic and contemporary methods were employed on CAM17. These normalized data sets were then fed into the proposed CAE as a distinct data set. This work explores the influence of the performance of CN techniques on the final results of a CBHIR framework through an extensive comparison between different results of the framework with inputs from each CN technique.

1.4.2 *Federated learning based strategy*

In Chapter 4, we mainly focus on breast cancer in distributed centers. Thanks to the CLoud ARtificial Intelligence For pathologY (CLARIFY) project we had a great opportunity to distribute our data set into two countries and four centers including companies and universities. These centers were TYris software company (TY), Valencia, Spain, University van Amsterdam (UvA), Amsterdam, The Netherlands, UPV Universitat Politècnica de València (UPV), Valencia, Spain, and University of Granada (UGR), Granada, Spain.

This contribution is presented in three scenarios and two different breast cancer data sets including BreaKHis and CAM17. In Chapter 4, we proposed to mimic a WWfedCBMIR within 2 and 4 centers to train an unsupervised FE. One custom-built CAE is introduced as an unsupervised FE to be trained under each scenario of this chapter. First, we distributed CAM17 and BreaKHis at 400× into TY and UvA centers to train the model with two distinct data sets. In the second scenario, we distributed each magnification of BreaKHis into a center to mimic the case that four clients connect to the proposed FedCBMIR. In the last scenario of this study, we assumed that pathologists need to retrieve similar patches to their query at different levels of magnification. To do so, we assumed that centers have an agreement and they can share similar patches. So, it is handy to retrieve similar patches at different levels of magnifications to be able to analyze the patterns with more detail.

1.4.3 *Contrastive learning based strategy*

Breast and skin cancer are the two main data sets in Chapter 5. In this Chapter, a Siamese network is presented for each data set and the obtained results overpassed not only the previous CBHIR tools but also the the classifiers. These proposed Siamese networks are robust to imbalanced data sets and address the shortcomings of histopathological images.

In this work, we made an extensive evaluation and comparison between the proposed CBHIR tool and the recently published techniques to compare the performance of the proposed tool in retrieving top K similar patches. To the best of our belief, it is the first CBHIR tool on breast and skin data set that can reach high performance with top first retrieved patches. In the other part of the evaluation, we compared our results with state-of-the-art classifiers to evaluate to what extent the proposed CBHIR tool succeeds in retrieving patches with the same class label as the query. The performance of the proposed CBHIR framework was extensively evaluated visually and numerically. We show the Gradient-weighted Class Activation Mapping (Grad-CAM) figures as explainable Skin-twins to provide interpretability to the uncifrable STUMP cases

The other challenge of pathologists that we considered in our study is grading STUMP tissues in skin cancer. To the best of our knowledge, it is the first time that a study dedicates its main objective to tackling this challenge and gives a hand to the pathologist in grading this case.

1.5 Framework

This PhD thesis is framed within a research project named CLoud ARtificial Intelligence For pathologY (CLARIFY)[8]. CLARIFY is an innovative, multinational, multisectorial, and multidisciplinary research and training program that makes a bridge

[8] https://www.clarify-project.eu/

between engineering and medicine with a focus on digital pathology. The main goal of this project is to increase the benefits of digital pathology and CAD to assist pathologists in their daily workflow. To do so, it creates a research infrastructure based on histopathological image processing, DL, and cloud computing to enable pathologists to share their knowledge and reach better-informed decisions.

The CBHIR framework in CAD is based on DL and cloud-computing algorithms to improve workflow efficiency, stimulate collaboration, and increase diagnostic confidence at pathology labs without the dependencies on the lab's locations.

1.6 Outline

This thesis comprises six chapters. The current chapter introduces the research motivation, defines the objectives, and highlights the primary contributions. Furthermore, this chapter provides an overview of the framework and outlines the structure of the thesis. This thesis is structured by four journal papers into four chapters.

Chapter 2 corresponds to a paper titled **Towards More Transparent and Accurate Cancer Diagnosis with an Unsupervised CAE Approach** which is published in ***IEEE Access*** [25] belonging to the editorial IEEE. The paper was published in 2023, but the following details in this journal correspond to 2022. This journal had an impact factor of 3.9 in 2022 and 4.1 in the 5-year impact factor. The top ranking was in the category of *ENGINEERING, ELECTRICAL & ELECTRONIC* with a percentile of 63.8 (Q2).

Chapter 3 corresponds to a paper titled **Advancing CBHIR Pre-processing: Comparative Analysis of the Effects of Color Normalization Techniques on Content-Based Histopathological Image Retrieval** which is published in ***Applied Science journal*** [26] belonging to the editorial MDPI. The paper was published in 2024. Applied science had an impact factor of 2.2 in 2022 and 2.9 in the 5-year impact factor. Applied Science is considered a Q2 journal in 2022 by having a 53.9 JIF percentile in the category of *ENGINEERING, MULTIDISCIPLINARY*.

Chapter 4 corresponds to the paper titled **WWFedCBMIR: World-Wide Federated Content-Based Medical Image Retrieval** which was published in ***Bioengineering journal*** [27] belonging to the editorial MDPI. The paper was published in 2023. Bioengineering had an impact factor of 4.6 in 2022 and 3.900 in the 5-year impact factor. The journal by having a 68.1 JIF percentile was counted as a Q2 journal in 2022.

Chapter 5 corresponds to the paper titled **Siamese Content-based Search Engine for a More Transparent Skin and Breast Cancer Diagnosis through Histological Imaging** which is under review by ***Computers in Biology and Medicine*** belonging to the editorial Elsevier. The first draft of the paper was published in 2024 on Arxiv [29]. Computers in Biology and Medicine journal had an impact factor of 7.7 in 2022 and 6.9 in the 5-year impact factor. The journal by having an 88.6 JIF percentile was counted as a Q1 journal in 2022.

Note that Chapters 2, 3, 4, and 5 follow a consistent communication structure. Initially, each chapter presents an abstract, followed by an introduction including a review of the related literature and outlining the key contributions of the proposed work. The subsequent sections of these chapters delve into a material section to illustrate the data sets used to train the proposed DL-based techniques which are detailed in the methodology section. Subsequently, the proposed methods undergo evaluation using specified methods outlined in the evaluation section. Then, an extensive comparison is conducted to report the performance of the proposed frameworks in comparison with other state-of-the-art and classic methods. Each chapter concludes with a succinct summary, recapping the main results and contributions. Finally, in the final section, they provide some future lines according to the main contributions for further investigations.

In Chapter 6, we align the findings of each chapter with the global aim of this Ph.D. thesis. We provide a comprehensive synthesis of the results and suggest future research lines. In the Merit section of this chapter, we highlight the academic achievements, including journal publications, and participation in national and international conferences. Finally, the Bibliography is presented.

Chapter 2

Towards More Transparent and Accurate Cancer Diagnosis

This Chapter corresponds to the author's version of the following published paper:

Tabatabaei, Zahra, et al. "Toward More Transparent and Accurate Cancer Diagnosis With an Unsupervised CAE Approach." IEEE Access, vol. 11, pp. 143387-143401, 2023.

Towards More Transparent and Accurate Cancer Diagnosis with an Unsupervised CAE Approach

Zahra Tabatabaei[1,2], Adrián Colomer[2], Javier Oliver Moll[1], and Valery Naranjo[2]

[1] Dept. of Artificial Intelligence, Tyris Tech S.L., Valencia, Spain.
[2] Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano, HUMAN-tech, Universitat Politècnica de València, Spain.

Abstract

According to the Global Cancer Observatory, 2020, breast cancer is the most prevalent cancer type in both genders (11.7%), while prostate cancer is the second most common cancer type in men (14.1%). In digital pathology, Content-Based Medical Image Retrieval (CBMIR) is a powerful tool for improving cancer diagnosis by searching for similar histopathological Whole Slide Images (WSIs). CBMIR empowers pathologists to explore similar patches to their query, enhancing diagnostic reliability and accuracy. In this paper, a customized unsupervised Convolutional Auto Encoder (CAE) was developed in the proposed Unsupervised CBMIR (UCBMIR) to replicate the traditional cancer diagnosis workflow, offering the potential to enhance diagnostic accuracy and efficiency by reducing pathologists' workload. Furthermore, it provides a more transparent supporting tool for pathologists in cancer diagnosis. UCBMIR was evaluated using two widely used numerical techniques in CBMIR, visual techniques, and compared with a classifier. Validation encompassed three data sets, including an external evaluation to demonstrate its effectiveness. UCBMIR achieved 99% and 80% top 5 recalls on BreaKHis and SICAPv2 with the first evaluation technique while using the second technique, it reached 91% and 70% precision for BreaKHis and SICAPv2, respectively. Moreover, UCBMIR displayed a strong capability to identify diverse patterns, yielding 81% accuracy in the top 5 predictions on an external image from Arvaniti. The proposed unsupervised CBMIR tool delivered 83% accuracy in retrieving images with the same cancer type.

2.1 Introduction

Cancer is a leading cause of death worldwide, with nearly 10 million deaths reported in 2020, as per the World Health Organization (WHO) [31]. In 2020, breast and prostate cancer affected 2.26 million and 1.41 million cases, respectively. Accurate cancer diagnosis is critical for effective treatment because each cancer type requires a specific treatment regimen. However, diagnostic errors are prevalent, affecting approximately 5.08% of cases, which translates to around 12 million adults in the United States [32]. This significant percentage of human error in a large number of cancer cases poses significant drawbacks for society and the quality of human lives.

Moreover, there is a significant disparity in treatment availability between countries with varying income levels. In high-income countries, comprehensive treatment is available in over 90% of cases, but this figure drops to less than 15% in low-income countries [33]. In this context, there is an urgent need to develop reliable and accurate diagnostic tools that can assist medical professionals in making accurate and timely diagnoses, regardless of their location or income level.

Computer-Aided Diagnosis (CAD) models play a vital role in reducing the incidence of human errors and providing an inclusive worldwide platform for individuals with varying incomes. CAD offers multiple approaches under the umbrella of "digital pathology" to enhance conventional cancer diagnosis. Digital pathology has garnered significant attention due to its ability to provide a definitive diagnosis at the pathology level, taking into account factors such as size, complexity, and color [13]. The challenges and opportunities presented in digital pathology are explained in [34]. Despite the challenges, digital pathology can serve as a bridge toward the discovery of histopathological imaging and enable more accurate prognostic predictions for disease aggressiveness and patient outcomes.

The following subsections cover a brief literature review of digital pathology on WSIs.

2.1.1 Segmentation

Automatic detection of irrelevant regions of tissue may bring a more reliable prediction. In [35], they proposed a multi-scale model to detect invasive cancerous area patterns in WSIs of bladder cancer. Similarly, [36] focuses on detecting blood and damaged tissue as problematic artifacts in bladder tumors. In [37], the authors apply the DenseRes-Unet model to multi-organ histopathological images to segment overlapped/clustered nuclei. A binary threshold is set to detect the contour of the extracted nuclei in the images, as the morphological characteristics of the cells are critical to grading the cancers. Moreover, the two-stage nuclei segmentation strategy proposed in [38] based on watershed segmentation is used to distinguish between carcinoma and non-carcinoma recognition in the Bio-imaging 2015 data set. Additionally, [39] introduced a novel approach to detect nuclei in breast cancer histopathological images using a stacked sparse Auto Encoder (AE).

Segmentation techniques have been studied extensively to quantify cell nucleus form and dispersion, which may improve accuracy in classification and grading [40]. However, these methods do not offer direct benefits to pathologists. Though Deep Learning (DL) has shown promise in improving segmentation, it relies heavily on large annotated data sets [41], limiting its impact [42]. Innovative approaches are needed to develop new techniques that can benefit pathologists and improve disease diagnosis.

2.1.2 Classification

Classification of input images is a critical task in medical image analysis, where an optimal classifier is expected to provide accurate labels for each input [28]. This can significantly aid pathologists in their daily analysis of tissue grading. For instance, in the realm of cancer diagnosis, saliency maps computation has improved the diagnostic process, both in radiological [43] and histopathological [44] images.

[3] validated an end-to-end pixel-level prediction of Gleason grades and scored the entire biopsy. Similarly, [45] proposed an AE using Siamese network aimed at learning image features by minimizing the distance between input and output. Another approach was proposed by [46], who aimed to decrease the rate of diagnostic errors by performing patch-based transfer learning. However, patch-level data sets extracted from Whole Slide Images (WSIs) often contain mislabeled patches, which may lead the classifiers to miss important information. To address this problem, [47] proposed DenseNet121-AnoGAN, which employs unsupervised anomaly detection with generative adversarial networks (AnoGAN) to prevent missing mislabeled patches. This approach has been successfully applied to classify the BreaKHis data set into benign and malignant.

The author in [48] fed an Inception Recurrent Residual Convolutional Neural Network (IRRCNN) model with two breast cancer data sets to have binary and multi-class classifiers. The authors in [49] conducted their experiments on DenseNet with SENet IDSNet and BreaKHis data set. They fine-tuned DenseNet-121 to propose an accurate classifier. A deep Feature Extractor (FE) from a pre-trained network and a classifier are used in [50] to classify BreaKHis. 16 pre-trained networks and 7 classifiers were tested in this paper.

In brief, classifier architectures have been proposed for use in diagnostic pathology to aid pathologists in making more accurate cancer diagnoses. Many studies have reported high classification accuracy, which has been validated in engineering laboratories. However, despite their potential importance, these measurements have not yet led to a significant change in diagnostic imaging.

While classification and segmentation have proven to be valuable tools, they have not drastically transformed the diagnostic process. This may be attributed to their inability to reduce ambiguity and boost the confidence of pathologists in their diagnoses. In essence, these methods do not provide any additional information to aid pathologists in their report writing during the diagnostic process. For instance, CAMs and saliency maps provide clinicians with information exclusively about the case under study. These are visual information from the input image that CNNs are paying attention to perform classification. While CBMIR goes deeper in providing transparency and reliability to pathologists. It is retrieving similar cases from the previously diagnosed cases. This approach is similar to the pipeline followed by pathologists who consult with reference books in pathology, such as Atlas. So, CBHIR is a digital intelligent Atlas book that can speed up the search process and be more accurate.

In regards to developing methods for specific applications, it is possible to achieve higher results for the intended objective. However, creating and implementing unique methods for each potential task of interest is impractical. As an alternative approach, CBMIR has established a reliable framework for quality control. While it may have poorer accuracy than an application-specific instrument, having a multipurpose general-purpose tool like CBMIR can still be useful.

2.1.3 Content-Based Medical Image Retrieval (CBMIR)

Automated medical imaging has been growing dramatically to improve clinical treatment and intervention in medical diagnosis. This yields an exigent demand for developing highly effective CAD systems. CBMIR is an active area of research with significant applications in routine clinical diagnostic aid, medical education, and research. In CBMIR, the end user targets retrieving the most relevant images. So, pathologists will trust this outcome easier because not only will they have a second opinion on their tissue (label), but they can also look for the same patterns in the previous tissues. In the classification task, they can get a label for their new tissue without knowing the reason. But in CBMIR, they can see the similarity between their tissues and the retrieved ones. Moreover, the explainable nature of CBMIR allows clinicians to understand how the system arrived at a particular diagnosis or recommendation, promoting transparency and trust in AI-assisted medical decision-making. Most notably, CBMIR is pathologist-centric; in contrast to classification, it is essentially an attempt to make decisions on behalf of the pathologists.

In DL, similar patterns mean similar features and representations. Humans can properly describe and interpret image contents, while digital machines can provide fewer semantic words for the same image. Machines provide a numerical description of the images with a wide gap compared to the human interpretation of the same image. This gap is named "*semantic gap*," and this broadly limits the performance of retrieval tasks [51]. The semantic gap is the main reason CBMIR has not made it into the daily laboratories workflow, yet. Indeed, this is arguably the paramount challenge in adopting CBMIR into the laboratories' workflow. Pathologists face numerous challenges in the current diagnostic paradigm, with time being a common factor. However, the impact of these challenges extends beyond just medical professionals and patients; it can also affect society as a whole. This can lead to emotional distress and other adverse effects on the well-being of patients and their families. Digital pathology, through the use of CBMIR, can mitigate the impact of these changes and enhance the accuracy of diagnoses.

CBMIR in virtual telepathology offers a reliable framework for achieving quality control through computational consensus-building, ensuring that diagnoses are accurate and consistent across different pathologists and healthcare institutions. By utilizing a vast database of reference images and advanced algorithms, CBMIR enhances the accuracy of diagnoses, potentially decreasing the need for additional studies and speeding up the diagnostic process. This can lead to better patient outcomes and a more efficient healthcare system. In recent years, CBMIR has gone through a renaissance with the promise of revolution. In a previous study [52], a CNN-based AE was applied to the BreaKHis data set with the aim of minimizing misinformation and evaluating the performance of CBMIR in a binary scenario. However, the reconstructed images produced by this method were found to be blurry, indicating that the extracted features by the AE were not robust enough to reconstruct the original image. In addition, the scope of this study was limited to detecting breast cancer using a two-class data set, without considering other diseases. These limitations highlight the need for further research to improve the quality of feature extraction in CBMIR systems. In [10], the CBMIR performance was improved in a supervised manner using a Hybrid feature-based ICNN model. The model was trained by adding three Fully Connected (FC) layers to accommodate the classification of cancer subtypes from TCGA. The researchers in [11] aimed to replicate the process of detecting morphological features used by pathologists in cancer diagnosis by incorporating different magnification levels into their CBMIR system.

Specifically, they trained their system using a subset of TCGA data set in three magnification levels: 20×, 10×, and 5×. To address the differences in features that might exist at these different magnification levels, the last DenseNet-121 block [53] was re-trained using 10× and 5× magnification patches. This supervised approach improved the adaptability of the FE and resulted in better overall performance of the CBMIR system. KimiaNet reported two types of image search: horizontal search and vertical search. In the horizontal search, the query is applied to the entire data set to find similar whole slide images (WSI) with a self-supervised model [12], while in the study by Fashi et al. [54], the vertical search approach is designed to identify similar types of malignancies in a specific organ. This is achieved by utilizing pretrained models with openly provided weights from the Keras library. The problem with supervised CBMIR is that it requires a large amount of labeled data, which can be time-consuming and costly to obtain. On the other hand, the problem with self-supervised CBMIR is that it may not perform as well as supervised methods and may require more complex models. Indeed, many researches [55, 56, 57, 58, 59, 60, 61, 62, 14, 63] have been dedicated to CBMIR, but the overall performance of the existing systems is not high enough due to the growing medical images and digital pathology.

The main contributions of this paper in proposing an Unsupervised CBMIR (UCBMIR) are:

- Proposing a new unsupervised approach for prostate and breast cancer gradation problem using CBMIR that achieves performance comparable to fully supervised methods.
- Extensively validating the proposed UCBMIR approach on three databases, including BreaKHis for a binary scenario and SICAPv2, which is the largest pixel-wise annotated prostate data set, and Arvaniti for multi-class grading problem, which is more challenging.
- Conducting an external evaluation to demonstrate the performance and generalization of UCBMIR, by training the model on SICAPv2 and testing it on the Arvaniti data set.
- A comprehensive evaluation of UCBMIR is presented, encompassing various numerical and visual performance metrics.

In addition, the paper addresses two major problems in traditional cancer diagnosis: inexperienced pathologists requiring more ancillary studies for diagnosis and the time-consuming process of differentiating between cancer grades. The UCBMIR model proposed in the paper provides a vast database of images that pathologists can use as a reference for diagnosis, allowing them to make more accurate diagnoses even if they are inexperienced. Additionally, the proposed tool enables pathologists to access annotated image databases instantly, leading to a faster diagnosis and skipping time-consuming reading and searching processes in "*ExpertPath*" and "*PathologyOutlines*" or an Atlas book.

In order to reach the primary objective of this study we introduce an unsupervised image search tool for pathologists to facilitate the efficient retrieval of similar images from previous cases. The initial stage involves training a customized CAE that includes a skip layer between the encoder and the decoder, as well as an attention block in the bottleneck. This CAE is trained to reconstruct images and learn effective data representations while simultaneously ignoring the noise. The encoder with the bottleneck of the trained CAE serves as our FE in the search stage. We represent the complete training set of the data set as in previous cases and carry out patch-by-patch retrieval to obtain diagnosis-relevant patches for each query in the test set. Our ranking algorithm, which utilizes Euclidean distance, identifies the retrieved patches, which are then presented to the pathologists as the output of the proposed UCBMIR. Our study showcases the practicality of our approach in enhancing the efficiency and accuracy of image retrieval for pathologists and engineers. As a result, our method can accelerate cancer diagnosis for pathologists, and the deep layers in the custom-built CAE can learn image features in an unsupervised manner, circumventing the issue of insufficient training images.

2.2 Material

The study evaluates the performance of the UCBMIR on two of the largest labeled histopathological images in breast and prostate cancer, namely BreaKHis and SICAPv2, respectively. The Arvaniti data set is utilized as the third and an external data set in order to validate the model performance. These two cancers, prostate and breast cancer, are selected as they are prevalent in society.

BreaKHis: breast tissue biopsy slides were stained with Hematoxylin and Eosin (H & E) and labeled by pathologists at the P&D medical laboratory in Brazil [64]. This data set is composed of 7909 microscopic images of breast tumor tissues collected from patients using magnifying factors of 40×, 100×, 200×, and 400× in the size of 224 × 224 × 3. This binary data set contains 588 benign and 1232 malignant images in 400×.

SICAPv2: prostate samples were sliced, stained in H &E, and digitized at 40× magnification. Images were divided into 512 × 512×3 and down-sampled to 10×, which is commonly used for evaluating images. This multi-class data set contains 155 WSIs in total: 4417 non-cancerous patches, of which 1635 are labeled as Grade 3 (G3), 3622 as Grade 4 (G4), and 665 as Grade 5 (G5), Table 1. Images labeled by a group of expert urogenital pathologists at Hospital Clínico of Valencia. SICAPv2 is the largest publicly available data set that includes pixel-level annotations of Gleason grading, providing detailed information on the presence of cribriform patterns[3].

TABLE 1: SICAPv2 data set description.

Grades	NC	G3	G4	G5
WSIs	37	60	69	16
Patches	4417	1636	3622	655

In order to validate the generalization capability of the UCBMIR to find similar images, Arvaniti, an external data set containing pixel-level annotations of Gleason grades, is used.

Arvaniti: the data set was shared by Arvaniti et al. [65], which contains 625 patches of prostate histology images at 40× magnification. Regarding a fair comparison with SICAPv2, in [3], some configurations were applied to re-sample images to 512 × 512 at 10× magnification. To normalize the color distribution of Arvaniti, the author in [3] applied a histogram match to the re-sampled images and set the images in SICAP and Panda[66] as the reference images. These re-sampled images are used as the third data set and the external evaluations in this paper. Arvaniti is employed for performance evaluation alongside normalization by both Panda and SICAP, in

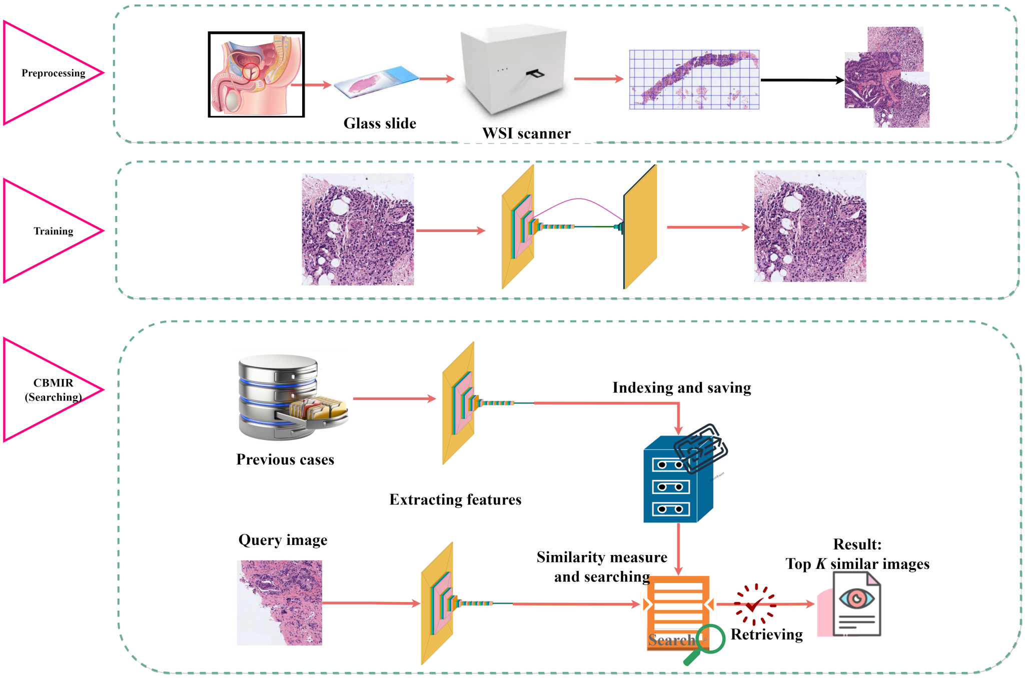


FIGURE 1: an overview of the UCBMIR for retrieving similar cases to a given query. The preprocessing stage involves extracting tissue from the patient's body and dividing the whole slide images (WSIs) into patches of a specific size[3]. In the training stage, these patches are used to train the proposed unsupervised CAE to extract feature representations. The trained encoder and bottleneck layers are then used to extract feature embeddings FEs that are used in the CBMIR section. In the CBMIR stage, the search engine computes the embedding features of the training set and stores them in a dictionary. When a query image is selected from the test set, the FE computes the embedding of that query and compares it with those in the dictionary. The model then returns the K most similar patches based on the pathologists' needs.

addition to the external evaluation. This is discussed in detail in the following sections of this journal paper.

2.3 Methodology

A CBMIR contains four subsections: 1. training, 2. indexing and saving, 3. searching, and 4. evaluating. The search tool in CBMIR uses the contents within each pixel of the images instead of using annotations or metadata. Consequently, similar images are retrieved from a large data set that matches the contents of the queried image. Also, it is often impractical to manually annotate images in a large data set, thus an unsupervised FE is developed in this study to address this issue.

The four phases of the proposed UCBMIR are described in-depth in the following subsections with Figure 2 and Figure 1, in accordance with SICAPv2, as it is a complex multi-class data set.

2.3.1 Training

The only training part of the proposed UCBMIR is training the proposed CAE. CAE aims to reconstruct the output as equal to the input. CAE could learn effective features with unlabeled data in an unsupervised manner. Using a CAE approach offers several benefits, such as its ability to capture spatial information through convolutional layers, which are well-suited for processing image data. Moreover, CAE employs multiple layers to extract advanced image representations, resulting in higher-level feature recognition. These multi-layered models have fewer free parameters, making them simpler and faster to train, reducing the cost and resources required for training. Figure 2 exhibits an illustration of the proposed CAE architecture. It contains three main parts:

- **Encoder**: it captures the structural attributes of the input images across a feature vector per image with 200 elements. The sizes of the convolutional filters in the encoder are $[16, 32, 64, 128, 256]$. Histopathological images are highly detailed; using operations such as pooling layers will cause them to lose lots of information. Because of that, in the proposed CAE, the size of the images decreases by passing through convolutional layers without any pooling layer.
- **Bottleneck**: it contains 200 extracted features per image. As can be seen in Figure 2, a residual block in the bottleneck contains four filters in the size of $[64, 32, 1, 256]$.
- **Decoder**: it reconstructs the input from its 200 intermediate feature vectors. Consequently, to the encoder, the filters in the decoder part are in the size of $[128, 64, 32, 16, 3]$.

The main objective of the proposed CAE is to find the most discriminative feature vectors to describe the images without supervision. Briefly, it compresses input image patches (of dimensions width × height × channels) into a fixed-length vector. The performance of the FE is directly related to the depth of the learning model, as deeper and more complex models might result in overfitting. To address this issue, the proposed CAE employs a residual block in the bottleneck to increase the network depth and improve end-to-end mapping.

To get better performance gain, inspired by highway networks[67] and deep residual networks[68], we add a skip connection between two corresponding convolutional and deconvolutional layers. When the network goes deeper, image details can be lost, making deconvolution weaker at recovering the input. Skip connections benefit by back-propagating the gradient to the bottom layers, making training a

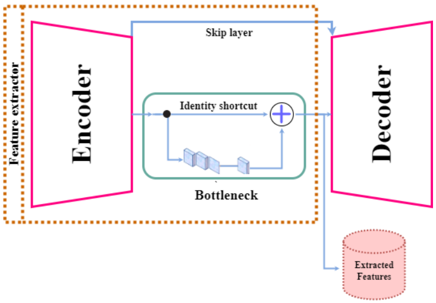


FIGURE 2: proposed CAE architecture with kernel size of 3 throughout the model, the stride of 2 in the encoder and decoder, and 1 in the bottleneck layer.

deeper network much more accessible. In other words, skip connections pass gradients backward, which helps find a better local minimum.

Mean Square Error (MSE) is the loss function that updates the network weights during the training phase. The minimum amount of MSE means more similarity between the input and the output. The greater the similarity between the input and the reconstructed images, the more meaningful features are in the bottleneck. In practice, we find that using Adam with a learning rate $5 \times 10^{(-5)}$ can train the model in 10 epochs. Then, the decoder part of the trained CAE is discarded, and the remaining sections, including the encoder and the bottleneck, play a role as an FE.

2.3.2 Indexing and saving

The indexing and saving stage is a crucial step in CBMIR as it enables efficient storage and organization of extracted features. This, in turn, enables fast and accurate retrieval of relevant medical images during the search stage, thus improving the diagnosis and treatment of medical conditions. In our study, we used n images from both the validation and train sets of each data set as input to the FE, resulting in n feature vectors that represent each image in a $200-$ dimensional latent space. These n feature vectors were then stored in a dictionary, $D_i = [F_1, F_2, ..., F_n]$, where each F_i contains the features for a single image.

During the retrieval stage, we utilized this dictionary as a reference for comparison with the query image. For the SICAPv2 data set, there were 2122 query images. By organizing and storing the extracted features in this manner, we aimed to improve the efficiency and accuracy of our CBMIR system. This approach enabled us to retrieve medical images that were relevant to a given query, thereby aiding in faster and more accurate diagnosis and treatment of medical conditions. Figure 1 illustrates the process of organizing and storing the extracted features for efficient retrieval of medical images during the search stage.

The extracted feature vectors of the database and the index of the related image were stored in a pickle file with the columns of $("indexes" : indexes, "features" : features$ by importing the pickle library in Python. Since extracting, indexing, and storing need high computational power, we implemented the work on GPU with the *NVIDIA GeForce RTX 3090*.

2.3.3 Searching

The searching process in CBMIR involves three key steps: similarity calculation, ranking and retrieval, and visualization and presentation. During similarity calculation, the search engine uses similarity measures such as Euclidean distance, cosine, Manhattan, and Haversine to calculate the similarity between the query image and other images in the database. The images in the database are then ranked based on their similarity to the query image, and the top-ranked images are retrieved and presented to the user for further analysis.

In this paper, we experimented with both Cosine and Euclidean distances, and based on our results, we concluded that the Euclidean distance was the more suitable choice. We use Euclidean Distance to measure the similarity of two feature vectors. Specifically, we calculate the distance by each query feature F_Q with all the feature vectors in D_i, and the smaller Euclidean value corresponds to more similar images. Our experimental findings suggest that Euclidean Distance is an effective metric for measuring similarity in CBMIR systems. By accurately measuring the similarity between images, the search engine can more effectively retrieve relevant medical images, leading to improved diagnosis and treatment outcomes.

2.3.4 Evaluation

It is worth considering what "accuracy" means in the context of a CBMIR. The accuracy of CBMIR depends on what we are looking for and what is displayed by the search engine. The use case determines whether the search is looking for images with the same stain, comparable stain intensity, same histologic feature, or similar grade; hence, this objective is ambiguous. To address this lack of awareness of the intent of the search engine, top K score at retrieving images of the same histologic features and Gleason grades engaged in the prior research to determine the performance of their experiments. To the best of the author's knowledge, there are two most-used strategies for calculating the top K score described in the recent articles:

1. If there is only one correct retrieved image, this has been shown as a correct answer [24]. In this paper, we set $K = 3, 5, 7$, which evaluates the performance of our model to correctly present at least one correct result in the top K retrieved images. In this paper, we name this method as "EV 1" regarding the report of the results in the following tables.

$$ACC@K = \frac{1}{N}\sum_{i}^{N}\varepsilon(\alpha_i, TOP(ans[:K])) \tag{2.1}$$

   In this equation, N denotes the number of query patches, and α_i represents the label of the i-th query patch. The function $TOP(ans_i[:K])$ retrieves the top k most similar results for the query and outputs 1 if any of these results match with the query and 0 otherwise. In other words, if $TOP(ans_i[:K])$ belongs to the set of labels of the i-th query, denoted by α_i, the function $\varepsilon()$ returns 1.

2. Precision (2.2) and recall (2.3) are the two selected indicators to evaluate the results. In this study, this is termed "EV 2".

$$Precision = \frac{R_v}{n} \tag{2.2}$$

$$Recall = \frac{R_v}{M} \tag{2.3}$$

The relevancy of the query and retrieved images has to be measured by considering the provided labels for each patch in the ground truth. Herein, R_v denotes the set of retrieved images that are considered relevant, while n signifies the total number of captured images. Moreover, the number of relevant images present in the data set is explicitly annotated as M. The proposed UCBMIR is evaluated based on both top-ranking image retrieval strategies to mimic the standard search process.

It is worth noting that this is the only place where the grand truth provided by expert pathologists was used. In other words, the labels were exclusively used to assess how effectively the model could retrieve images with similar histopathological patterns.

2.4 Discussion and results

Matching pairs to the image is the core of any search engine, in which an image is compared to a database to determine similarities. Numerous studies have been conducted on CBMIR in a binary manner, as it is more challenging with multi-class data sets.

Breast cancer is a prevalent malignancy affecting women globally. In the domain of CAD, the BreaKHis data set is a popular choice for evaluating the performance of algorithms in CBMIR. In this study, we employed the BreaKHis to assess the efficacy of our UCBMIR approach. Our method demonstrated superior performance in matching image pairs, as evidenced by the results presented in Table 2. Specifically, our approach outperformed two previously reported methods, namely [52] and [69], with precision scores of 92% and 91%, respectively, for both evaluation criteria (EV1 and EV2). These results suggest that our approach is highly effective in accurately identifying patterns in breast cancer images. Besides, according to the obtained results in Table 2, EV 1 works better in assessing the performance of a CBMIR tool since it succeeds in reporting the recall while EV 2 suffers from a notable decrease in measuring the recall. So, in the following experiments, EV 1 evaluates the CBMIR performance for prostate cancer as the multi-class data set.

TABLE 2: comparative results on BreaKHis 400× at $K = 5$. We measure the precision and recall with both EV 1 and EV 2.

Type of evaluation	Method	Precision	Recall
EV 1	UCBMIR	**0.92**	**0.99**
EV 2	UCBMIR	**0.91**	**0.50**
	Minarno [52]	0.70	0.31
	Gu [69]	0.63	-

In order to evaluate the effectiveness of our CBMIR method on a multi-class data set, we utilized SICAPv2 and Arvaniti data sets, both containing four classes. Given the global prevalence of prostate cancer, we selected this type of cancer for our experiments and used SICAPv2 as the largest pixel-wise annotated data set. The Arvaniti data set was re-sampled by referencing Panda and SICAP, as stated in [3], and was used in two experiments of this paper to demonstrate the robustness of our methodology. Table 3 presents the results obtained using our approach with $K = 5$ and EV1 as the evaluation criteria. To demonstrate the efficacy of our methodology, we conducted two experiments using the Arvaniti data set. In the first experiment, to ensure a fair comparison, we trained the model using Arvaniti normalized based on both SICAP and Panda. The results of these two trained models, obtained by conducting the entire training and searching steps, are reported in Table 3. These findings demonstrate the superior performance of our approach in accurately identifying and classifying prostate cancer images in multi-class data sets, thereby potentially contributing to the development of improved diagnostic tools and clinical decision-making processes.

In a study by Hegde [70], Scale-Invariant Feature Transform (SIFT) [71] was used as a traditional FE, along with SMILY, to report the accuracy of retrieving images with the correct Gleason patterns from prostate specimens in TCGA. Our UCBMIR, as shown in Table 3, achieved an accuracy of 80%, surpassing SMILY's accuracy of 73%.

TABLE 3: model quality results on SICAPv2 with top 5 retrieved images. The reported results are obtained by EV 1. The metrics are precision, recall, and accuracy. All other studies reported their results at the top 5 images.

Method	Data set	Precision	Recall	Accuracy
UCBMIR	SICAPv2	0.79	**0.80**	0.79
UCBMIR	Patches normalized Arvaniti (SICAP)	0.71	0.75	**0.80**
UCBMIR	Patches normalized Arvaniti (Panda)	**0.80**	0.68	0.78
Hegde [70] *(SMILY)*	TCGA	-	-	0.73
Hegde [70] *(SIFT)*	TCGA	-	-	0.62
VGG16 *(ImageNet)*)	SICAPv2	**0.80**	**0.80**	**0.80**

To provide an interpretive perspective for the quantitative results, we incorporated a pre-trained VGG16 [72] (ImageNet) as a backbone to extract histological features from the images. We added a GlobalMaxPooling2D (GMP) and two dense layers $[200, 4]$ to train the model as a classifier in a fully-supervised manner. After training the model, we removed the last layer (Dense (4)) and used the remaining layers as an FE to extract 200 features per image, which were then fed into the search engine component of UCBMIR. Figure 3 illustrates how we integrated the pre-trained VGG16 into our CBMIR. So, as can be seen in Figure 3, first, the VGG16 trained as a classifier by adding two dense layer with 200 and 4 nodes. Subsequently, the well-trained VGG16, in conjunction with the dense layer with 200 features, was moved to the offline and online sessions of the proposed CBMIR platform to extract the features of the data set and the query. Following the retrieval of the top K images, it is time to display and evaluate the performance of the CBMIR platform with a supervised FE (VGG16). Comparing the results shown in Table 3, UCBMIR achieved a comparable performance as the supervised method with EV1.

We conducted experiments using SICAPv2 and Arvaniti data sets to evaluate and compare the effectiveness of our UCBMIR with respect to the fully-supervised VGG16 in identifying and classifying prostate cancer images. As shown in Tables 3 our unsupervised method achieved similar results to the supervised VGG16, with a slightly lower precision score in both EV 1 (by 0.01). This suggests that our unsupervised method is a promising approach for CBMIR in the context of prostate cancer, potentially reducing the need for manual annotation and supervision.

In this study, Figure 4a and Figure 4b were used to present the results of the experiments. The bar charts were used to depict the number of similar images out of $K = 5$ retrieved images for BreaKHis and SICAPv2, respectively. For the BreaKHis data set, 545 images in the test set were used as query images. Based on Figure 4a, the model failed to find at least one similar image for 29 queries, while it could find three and four similar images for 114 and 170 queries, respectively. In the case of SICAPv2 data set, the model could retrieve one similar image among the top K for 628 image queries, according to the results shown in Figure 4b. The model was able to retrieve two images out of 5 in the same class label as the query in 101 cases of BreaKHis. In the case of SICAPv2, the model could retrieve two images with the same class label for 628 queries out of the top 5 retrieved images. So, the proposed

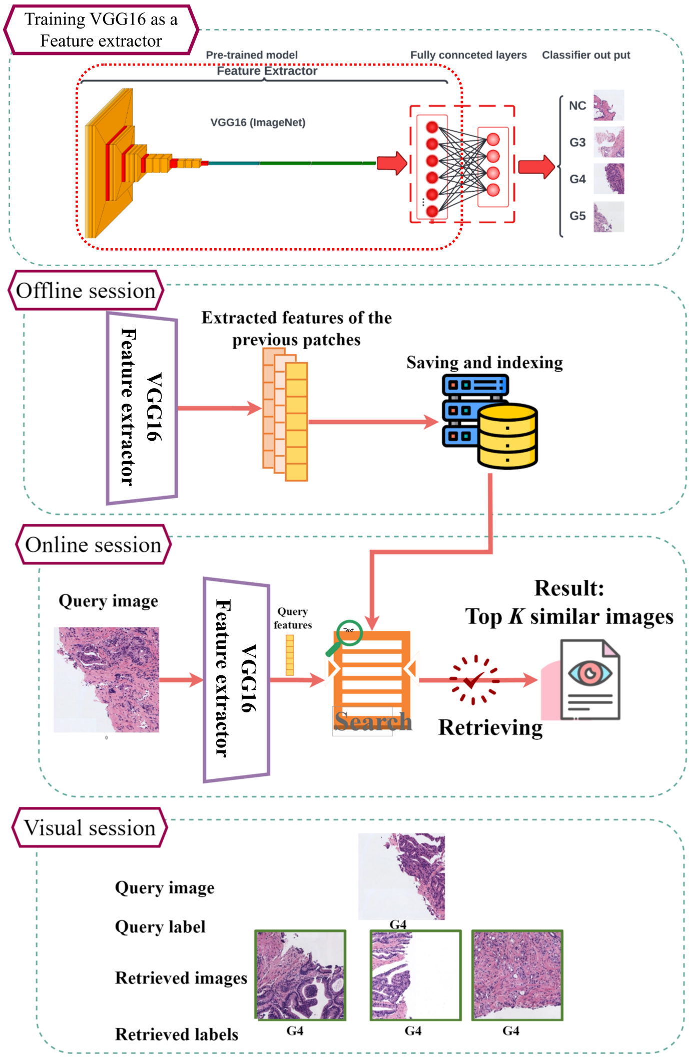
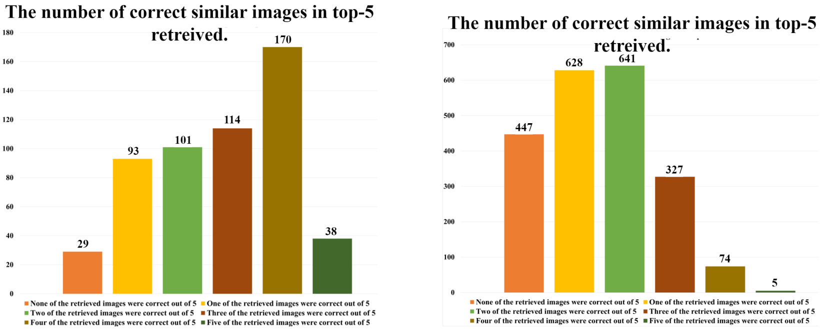


FIGURE 3: the structure of training VGG as a multi-class classification on SICAPv2 and delivering 200 features per image to the following CBMIR steps. The Conv2D layers are shown in "orange", MaxPooling2D in "red", Dense layers in "green", and Flatten in "teal". By discarding the last dense layer of the well-trained VGG16, the FE can be moved to the next sessions to extract the features of the images. This is just used as a fully supervised baseline method to compare the performance of the proposed unsupervised FE in extracting features with respect to the one extracted from the classification-task VGG training.

UCBMIR approach can correctly retrieve at least one sample belonging to the same category as the query in 78.9% of cases for the SICAPv2 data set. In the BreaKHis data set, this retrieval success rate is even higher at 94.7%. To the best of the authors' knowledge, these results are very promising in the field of automatic atlases.



(A)

(B)

FIGURE 4: evaluation of the UCBMIR at k = 5 on BreaKHis 4a and SICAPv2 4b. From 2122 query images in SICAPv2, for 447 cases, the model could not find at least one correct similar image according to their labels, while it retrieved two similar images at 5 top for 641 cases.

Due to the well-known variability between pathologists in Gleason grading and variations in histology sample preparation, it is a difficult challenge to distinguish between different grades of prostate cancer. These factors may contribute to the differences in results. Differentiating between G3 and G4 in prostate cancer requires highly experienced pathologists, takes time, and has limited inter-pathologist repeatability. However, Figure 5 demonstrates the impressive ability of our UCBMIR to identify similar patterns between G3 and G4. Each row and column in Figure 5 corresponds to three different values of K in Arvaniti (Panda), SICAPv2, and Arvaniti (SICAP), respectively.

These results highlight the potential of our approach to aid in the accurate identification and CBMIR of prostate cancer images, thereby facilitating diagnosis and improving patient outcomes. Further research is needed to validate these findings on larger and more diverse data sets. Due to this, our model is also verified on an external data set with the intention of evaluating the trained model's capacity for generalization.

2.4.1 Validation on an external data set

In order to validate the performance of our trained model on an external data set, we utilized the SICAPv2-trained model to make predictions on the re-sampled Arvaniti data set (normalized by SICAP). The results of this evaluation are reported in Table 4. The obtained accuracy and precision results are slightly better than those obtained from the test set on SICAPv2, while the recall is slightly lower by 0.07. It is important to mention that this validation process is crucial in demonstrating the generalization ability of our UCBMIR beyond the original training data set. These results further confirm the robustness of our approach and its capability to provide accurate retrieval results across multiple data sets. hl

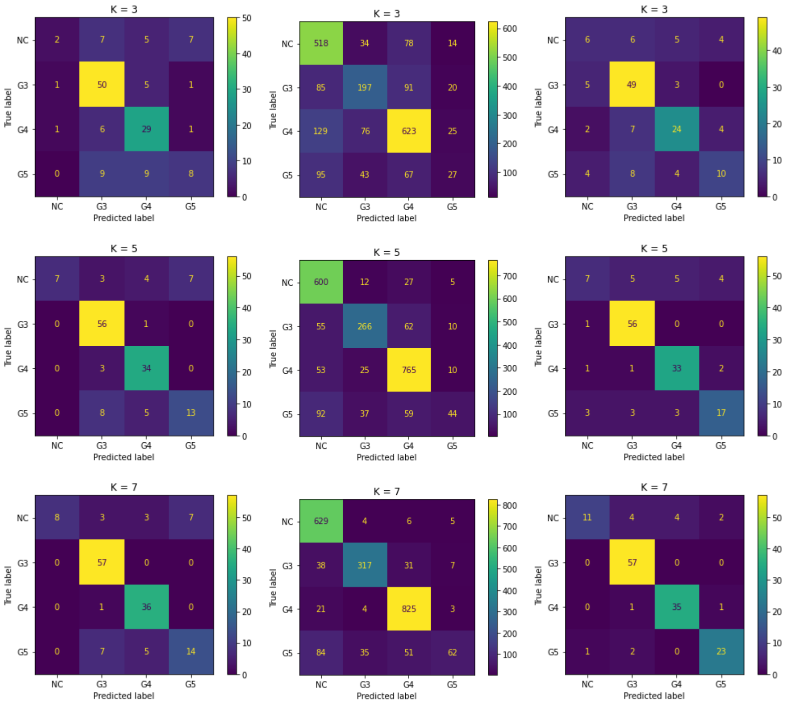


FIGURE 5: confusion matrix of UCBMIR on the different test cohorts at K = 3, 5, 7. a. ARVANITI (Panda), b. SICAPv2, c. ARAVNITI (SICAP).

2.4.2 Visual evaluation

We have included three figures, Figure 6, Figure 7, and Figure 8, which showcase the results of our experiments. These three figures illustrate the interpretability of the proposed UCBMIR.

The purpose of these figures is to enhance comprehension of the comparison by utilizing visual aids. Each of these figures comprises rows that correspond to a

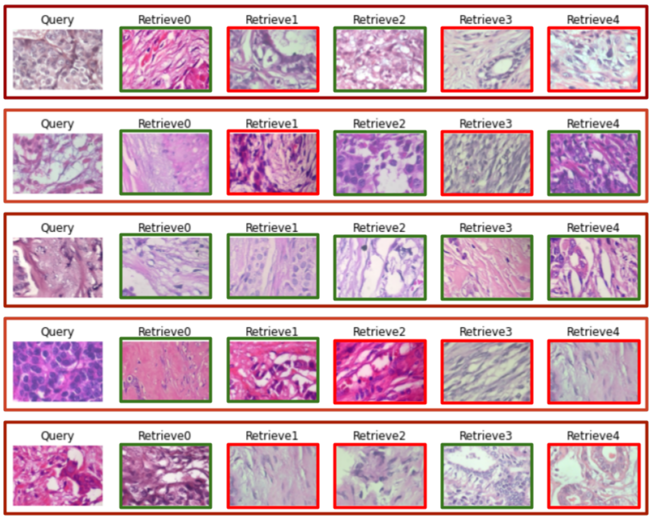


FIGURE 6: the top 5 images retrieved from the BreakHis data set for five randomly selected query images, with the true and false retrieval results depicted in green and red boxes, respectively.

random query from the test set of BreaKHis×400, SICAPv2, and Arvaniti, respectively. The subsequent images in each row exhibit the top 5 retrieved images from the training set of the relevant data set. We have implemented a color-coding scheme to facilitate the interpretation of the results. In particular, a green border surrounds the correct retrieved image, which possesses the same label as the query, whereas a red border highlights the mis-retrieved images that have different labels than the query.

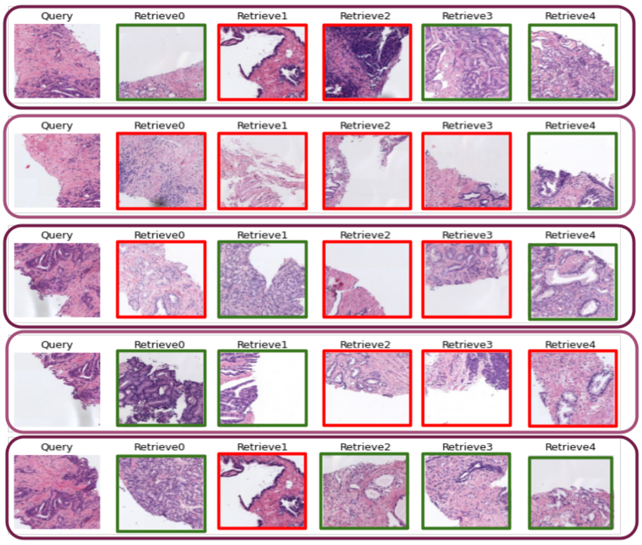


FIGURE 7: the top 5 retrieved images from the SICAPv2 data set for five query images selected at random, where true and false retrieval results are respectively indicated with green and red boxes.

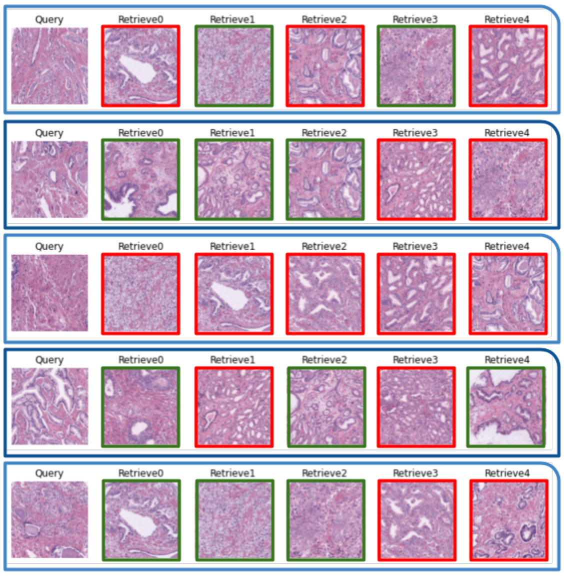


FIGURE 8: the top 5 images retrieved from the Arvaniti data set, which have been normalized by SICAP, for five query images that were randomly selected. The figure visually demonstrates the results of the external validation, utilizing a well-trained model with the SICAPv2 data set and applying it in the search stage of Arvaniti. The green and red boxes, respectively, indicate the true and false retrieval results.

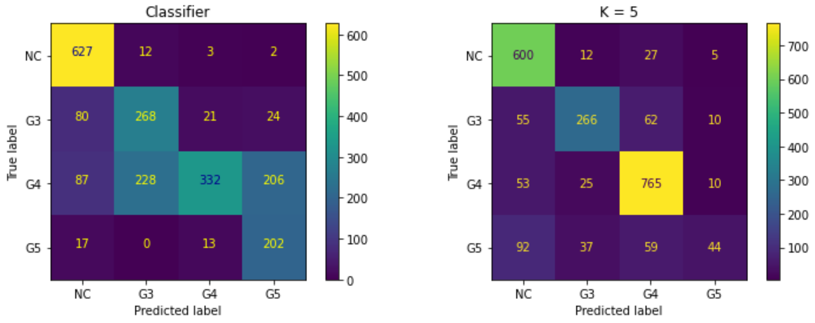
TABLE 4: results of Arvaniti (normalized by SICAP) as data set for the external experiments with top $K = 3, 5$, and 7, EV 1

K	Precision	Recall	Accuracy
3	0.66	0.58	0.66
5	0.80	0.73	0.81
7	0.89	0.86	0.91

Figure 7 demonstrates that one of the challenges we encountered in our experiments with SICAPv2 was the presence of a white background in the images. To determine whether the patches in SICAPv2 contained meaningful patterns for pathologists to analyze, we enlisted the help of an expert pathologist to review them. Our pathologist confirmed that despite the presence of a white background in the images, there was still enough tissue for pathologists to evaluate and compare the patterns in the query tissue with the retrieved patches. There are some bad cases as shown in line 2 of Figure 7 which most of the retrieved images are not highlighted with the red border. This means that the tool had challenges in retrieving similar patches for this query. The reason is mainly due to the high similarity of histopathological features of G3 and G4 of prostate cancer.

In addition to validating the approach of UCBMIR using an external data set and demonstrating the generalization capability of our method, we selected the Arvaniti data set for another reason: it does not have a white background. Figure 8 shows the top 5 retrieved images resulting from our external validation experiment, where a well-trained model with SICAPv2 was used to retrieve images for five random queries from the Arvaniti data set. Our external validation experiment not only validates our proposed UCBMIR for use with external data sets but also demonstrates the generalization capability of our method.

Through our visual evaluation, we aim to present a clearer understanding of the effectiveness of our approach. Observing the retrieved images alongside their labels can be useful to evaluate the performance of our method and assess its strengths and limitations. These figures are an essential component of our evaluation and will contribute significantly to understanding our methodology. Overall, this evaluation can provide valuable insights into the performance of our approach and make informed judgments regarding its effectiveness.



(A) (B)

FIGURE 9: The confusion matrix presented in [3] is displayed in Figure 9a, while the matrix for the retrieved labels is shown in Figure 9b. It can be observed that the UCBMIR model results in less conflict between the challenging grades (G3 and G4) compared to the classifier.

2.4.3 Comparing UCBMIR with a classifier

CBMIR and classification are two different approaches in medical image analysis. CBMIR aims to retrieve similar images from a database based on the content features of a query image, while classification aims to categorize images into pre-defined classes or labels. The only mutual output of the classification tool and CBMIR is the output labels corresponding to the output patches. In the above sections, it is mentioned that UCBMIR achieved comparable results with supervised CBMIR techniques, proven by the reported accuracies in Table 3. Regarding comparing the predicted labels in terms of retrieving similar images belonging to the same cancer type, we provide Table 5. This table compares the accuracy of UCBMIR with a classifier in [3] in both the validation and the test set of SICAPv2.

The proposed unsupervised CBMIR model was found to be highly effective in distinguishing between different cancer grades, especially between the challenging Gleason grades G3 and G4. This observation was evident from the confusion matrix shown in Figure 9. The proposed method's success can be attributed to its ability to identify and utilize subtle features and patterns in the images that may be missed by human observers or conventional supervised models.

TABLE 5: shows a comparison between the performance of UCBMIR with EV 1 and the classification introduced in [3]. SICAPv2 is the data set under study.

Data set	Model	Accuracy
Validation set	UCBMIR	**0.83**
	Classification[3]	0.76
Test set	UCBMIR	**0.79**
	Classification[3]	0.67

2.4.4 Limitations

The limitation of the pre-processing step and the medical session can directly affect the final results of the DL-based UCBMIR. For instance, noisy images and low-quality scanners can diminish the accuracy of the retrieval task. However, CAE was chosen to tackle the noisy images while training; the low quality of images might still affect the final results. Color variation as a result of different staining processes, different scanners, and laboratory conditions might fool the DL-based models. In this paper, adding a histogram equalization technique during the image loading process, could tackle this issue, and as can be seen in Figure 9b and Figure 6, the proposed framework is robust against this issue. Due to the computational limitation, large-scale retrieval is a limitation that hinders the performance of the CBHIR tools in searching through a vast database of biopsies. The proposed tool can retrieve the images at high speed. In the SICAPv2 experience, it can retrieve the top 5 images in almost 0.28 seconds.

Although some challenges have been solved by the proposed tool, some challenges still remain such as semantic gap, scalability, cross-domain retrieval, privacy and security, etc. For instance, medical images need high data protection due to the personal information. This affects extending the CBMIR domain as a worldwide tool across different hospitals. Another important limitation is the semantic gap between the extracted features of the low-level histopathological features and high-level semantic concepts. To address this issue, FEs have to train effectively to extract the most representative features of the patches, fast and accurately. However, limitation or inconsistent annotations makes the well-training of the DL-based FEs difficult. Despite all these limitations, the CBHIR tool provides a promising improvement in cancer diagnosis, research, and treatment. Several pieces of research aim to address many of these issues and further enhance the capability of the CBHIR tool and provide more transparency, robustness, and trustworthiness.

2.5 Conclusion

In summary, this paper introduces a highly qualified Unsupervised CBMIR (UCBMIR) model that can be used for both binary and multi-class data sets. The model was evaluated on three different data sets, as well as an external validation set. Using the two most-used evaluation techniques, the proposed method achieved 79% precision in EV 1 on SICAPv2 as a multi-class data set. Notably, the unsupervised method was able to differentiate between challenging Gleason grades of prostate cancer. In addition to numerical evaluation, visual assessments were conducted to demonstrate the effectiveness of the UCBMIR. The results show that UCBMIR has good generalizability and can be effectively applied to other types of cancer. UCBMIR has the potential to improve laboratory productivity, increase pathologists' diagnostic confidence, and contribute to the advancement of cancer diagnosis and treatment.

The UCBMIR model not only addresses the needs and challenges of pathologists but also addresses the problem of engineers who face a lack of sufficient images for training models. Future research in this field could build on these findings and further enhance the performance of CBMIR models for cancer diagnosis.

2.6 Future work

In the world of CBMIR, there is a vast range of possibilities for enhancing and optimizing laboratory productivity. With a large archive of diagnosed patients and corresponding data, including images and treatment and monitoring reports, it should be possible to identify and retrieve images that are either anatomically or pathologically similar to the biopsy sample of the patient being examined, as well as the annotated data for each case. CBMIR has the potential to be applicable to many types of cancer, which would increase its utility.

Furthermore, pathologists' reports contain the medical knowledge of many other pathologists for similar cases, making them a treasure trove of high-quality diagnostic information. In the future of CBMIR, it may be possible to make the raw information directly available to the pathologist or to merge the important information in retrieved reports. This would make the diagnosis process more efficient, accurate, and informative for both the pathologist and the patient. Additionally, expanding the use of CBMIR to other types of medical imaging and diagnostic data could provide valuable insights for a range of medical specialties.

In order to integrate CAD tools into daily clinician routines, reliability, trustworthiness, and transparency are the most critical needs. By incorporating eXplainable AI (XAI) methodologies like filter activations, the decision of the DL-based tools can be demystified. In regard to integrating these tools with the clinical workflows, it is necessary to harmonize DL-based tools alignment with the diagnostic precision and the quality of patient care.

2.7 Acknowledgment

I sincerely thank Umay Kiraz and Andrés David Mosquera Zamudio for their collaboration and efforts in image annotation. Their expertise and dedication were invaluable, and I am grateful for the opportunity to work with them.

Chapter 3

Impacts of color normalization on CBMIR

This Chapter corresponds to the author's version of the following published paper:

Tabatabaei, Zahra, et al. "Advancing Content-Based Histopathological Image Retrieval Pre-Processing: A Comparative Analysis of the Effects of Color Normalization Techniques." Applied Sciences 14.5 (2024): 2063.

Advancing CBHIR Pre-processing: Comparative Analysis of the Effects of Color Normalization Techniques on Content-Based Histopathological Image Retrieval

Zahra Tabatabaei [1,2†,*], Fernando Pérez Bueno [3], Adrián Colomer [2],Javier Oliver Moll [1], Rafael Molina [4], and Valery Naranjo [2]

[1] Dept. of Artificial Intelligence, Tyris Tech S.L., Valencia, Spain, 46021.
[2] Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano, HUMAN-tech, Universitat Politècnica de València, Spain, 46022.
[3] Basque Center on Cognition, Brain and Language, 20009 San Sebastián, Spain; f.perezbueno@bcbl.eu
[4] Dpto. Ciencias de la Computación e Inteligencia Artificial. Universidad de Granada, 18014.

Abstract

Content-Based Histopathological Image Retrieval (CBHIR) is a searching technique based on visual content and histopathological features of Whole Slide Images (WSIs). CBHIR tools assist pathologists to have faster and more accurate cancer diagnosis. Stain variation between hospitals hampers the performance of the CBHIR tools. This paper explores the effects of Color Normalization (CN) in a recently proposed CBHIR approach to tackle this issue. In this paper, three different CN techniques were used on the CAMELYON17 (CAM17) data set, which is a breast cancer data set. CAM17 was produced by different staining protocols and scanners in five hospitals. Our experiments reveal that a proper CN technique, which can transfer the color version into the most similar median values, has a positive impact on the retrieval performance in the proposed CBHIR framework. According to the obtained results, using CN as a pre-processing step can improve the accuracy of the proposed CBHIR framework to 97% by 14% increasing, compared to working with the original images.

3.1 Introduction

Breast cancer is one of the most prevalent cancer types with 2.3 million women diagnosed with this cancer and 685,000 deaths globally in 2020 [22]. For this large amount of patients, medication should be accurate with little to zero margins for errors, otherwise, the consequence of a wrong diagnosis could be fatal [73]. Also, since breast cancer is one of the leading causes of death for women globally, making precise detection and ensuring timely treatment can enhance the chance of recovery [74]. Computer Aided Diagnosis (CAD) provides some Deep Learning (DL)-based techniques in digital pathology that can assist pathologists in more accurate cancer diagnosis [28]. These techniques need to be trained by medical images such as Magnetic Resonance Imaging (MRI) [75] and histopathological images [76]. A unique feature of histopathological images is that they are typically much larger than other medical images [70].

Histopathological images play a crucial role in the realm of medical image processing, allowing the integration of image information and pathologists' expertise to improve diagnosis[77]. Medical images have witnessed a rapid expansion in quantity, content, and dimension [36]. Due to an enormous increase in the number of diverse clinical exams and the availability of a wide range of image modalities, demand for efficient medical image data retrieval and management has increased [78]. Current approaches to medical image retrieval often rely on alphanumeric keywords assigned by human experts, enabling retrieval at a conceptual level. However, this text-based search methodology falls short in capturing the intricate visual features inherent in image content [79].

Recent Content-Based Histopathological Image Retrieval (CBHIR) methods support full retrieval by visual content and histopathological patterns of the tissue. These advanced CBHIR tools facilitate searching at a perceptual level [80]. It is noteworthy to mention that a single pathology image may contain just basic patterns of a tissue such as epithelium and connective tissue. However, the actual number of patterns in the DL-based technique's point of view is almost infinite.

CBHIR explores a database to find visually similar images to provide clinicians with comparable lesions. In the diagnostic workflow, pathologists utilize this search engine to reach top k similar to their queries to determine if a histological feature is malignant or benign [59].

Mainly, CBHIR tools work on the extracted features of the images, including color, texture, shape, etc. Color is a visual feature that plays an important role in CBHIR techniques due to its invariance with respect to image scaling, translation, and rotation [24]. The use of color improves the capturing of distinctive histopathological features. This provides valuable information about the distribution and arrangement of different tissue components in histopathology [81].

In histopathology, tissues must be stained using various dyes, including Hematoxylin and Eosin (H&E), in order to be readable to pathologists [27], [36]. In digital pathology, these tissues must be scanned as Whole Slide Images (WSIs)[35]. In addition to the use of different scanners and staining manufacturers, lab conditions and temperatures may cause color variation in WSIs[82]. Color variation in WSIs can arise from both inter- or/and intra- laboratory factors in the acquisition procedure [81]. Fig 3.1 shows the color variation across five different collaborating hospitals involved in the collection of CAMELYON17 challenge (CAM17) data set [83]. This diverse color variation misleads the model and potentially deceives the Feature Extractors (FEs) into extracting erroneous features. Also, the performance of CAD tools may be significantly influenced by these variations in WSIs [84].

In this paper, we propose a CBHIR framework that includes Color Normalization (CN) as a pre-processing step to address this issue. CN methods have been developed with the aim of transferring the color interval of the WSIs in a data set to a common color range [85, 82, 86].

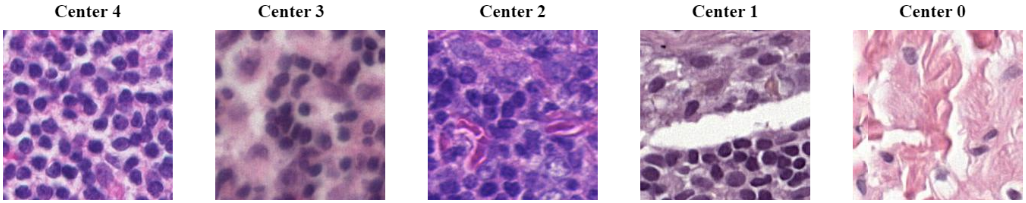


FIGURE 10: image examples from five hospitals that had collaborated to collect CAM17 as a single data set.

3.2 Related work

3.2.1 Content-Based Image Retrieval

There are different approaches to Content-Based Histopathological/Medical Image Retrieval techniques in histopathology [87], which we briefly review here. [88] proposed a framework for size-scalable query Regions Of Interest (ROI), including epithelial breast tumors of Motic data set. This work reached 96% precision in retrieving the top 20 similar images. In [88], the authors applied the proposed technique on the Motic database [1] with the original color of 145 stained WSIs. Authors in [63] applied a supervised kernel hashing technique on several thousand histopathological images from breast microscopic tissue. The precision results for the top 30 retrieved images were reported as 77.0%. In this study, they considered the gradient of pixels around the detected regions to make the model robust to subtle changes in the color of patches. A combination of Unsupervised Features Learning (UFL) with the classical Bag-Of-Features (BOF) was introduced in [89]. Their proposed method was evaluated in particular histopathological images to show that the learned representation has a positive impact on the retrieval performance. A CBHIR based on multi-scale, multichannel decoded local ternary pattern features and VLAD coding was presented in [90]. The authors evaluated their method on KIMIA Path960 data set [91] by retrieving top 10 matching images from the data set. The authors in [24] presented a modified Convolutional Auto Encoder (CAE) to extract features of patches from SICAPv2 as the largest pixel-annotated prostate data set[3]. The experimental results of this paper demonstrated 85% and 78% accuracy at top 7 and top 5, respectively. However, one of the challenges that the author had with the SICAPv2 was the color variation of the images in the data set. To deal with this variation, they applied a simple histogram equalization to the patches, but still, this issue affected the results.

Among the previously mentioned studies, only [63] and [24] applied CN techniques to tackle the color variation, while the rest did not address this issue. Another challenge CBHIR is the lack of annotated images, which the mentioned papers above struggled with.

[1] Motic (Xiamen) Medical Diagnostic Systems Co. Ltd., Xiamen 361101, China.

3.2.2 Color normalization

CN is the most used technique to deal with color variation [82]. CN can standardize images by referencing an image and simulating a chosen staining procedure. CN methods normalize the images with different techniques, including histogram matching, color transfer, and spectral matching [82]. Histogram matching disregards stain separation, color transfer modifies colors based on statistical correspondences between histological regions, and spectral matching estimates stain concentrations and color properties [92].

Among a variety of CN techniques, some are more popular, such as those proposed in [93] and [94]. The method proposed by Macenko et al. [93] introduced the Mac [2] technique in 2009, assuming that the amount of protein or nucleic acids is a random variable. Mac utilizes Singular Value Decomposition (SVD) to separate H&E channels. Within this technique, the concentration intensity of both source and target images is scaled using the 99th percentile to compute a robust estimation of the maximum.

Vahadane et al. [94] published the Vah [3] technique in 2016 to model the physical phenomena that define tissue structures. In this technique, there is a preferable stain color based on pathologists' point of view and a stain density map. Using an unsupervised approach, the method decomposes images into stain density maps. When Vah is applied to a specific image, it combines the relevant stain density maps based on a pathologist's preference for stain color. This process selectively modifies the color while preserving the underlying structure as described by the maps.

The most recent techniques in CN are the Auto-Encoders (AE), Generative Adversarial Networks (GAN), and Bayesian K- Singular Value Decomposition (BKSVD) [95, 96, 30]. Bentaieb et al. [95] applied GAN to combine the color normalization and classification of WSIs. In this method, the generator is employed as a stain transfer network, while the discriminator separates the classes and original and normalized images. In order to map unpaired images between two scanners, the cycleGAN was used by StainGAN [92]. In [96], the authors used three different Convolutional Neural Network (CNN) models for CN purposes: Variational Auto Encoder (VAE), GAN, and Deep Convolutional Gaussian Mixture Model (DCGMM). In [97], a CN network is fed by a heavily augmented data set and trained to reconstruct the original appearance of WSIs. Notably, Pix2pix conditional GAN framework and CycleGAN are the other noticeable CNN architecture.

BKSVD [30], a cutting-edge technique that was proposed in 2022, utilizes an unsupervised estimation of the stain concentration that preserves histological structures with variational and empirical Bayes.

In this paper, we provide a deep understanding of the effects of CN on the extracted features in the proposed CBHIR tool. Our experiments were conducted by applying three CN techniques in the pre-processing step in order to explore how CN influences the extracted features within the CBHIR tool. Among CN techniques, Mac [93] and Vah[94], were selected as the two most used CN techniques, and BKSVD was applied as a recent CN technique. Furthermore, in this work, we applied an unsupervised FE to tackle the need for a pool of histopathological images to train the deep-learning model.

The main contributions of this paper are as follows:

1. Proposal of a new CBHIR framework with an unsupervised feature extractor that takes color normalization into account as a pre-processing step;

[2] The proposed method in [93] is named Mac in this paper.

[3] The proposed method in [94] is named Vah in this paper.

2. Analysis of CBHIR performance when using normalized images in comparison with original images;

3. We draw attention to the relevance of color variation and its impact on CBHIR

4. We provide a comprehensive performance assessment of the proposed method. This evaluation employs a large breast cancer database scored from five distinct laboratories. This evaluation has a more restrictive K-top accuracy assessment compared to the recent state-of-the-art studies and also involves an in-depth analysis of retrieving patches with the same cancer label.

3.3 Methodology

This section introduces the three levels of the proposed CBHIR tool in detail. Fig 11 provides an overview of the proposed CBHIR tool, comprising three levels: pre-processing, training, and searching.

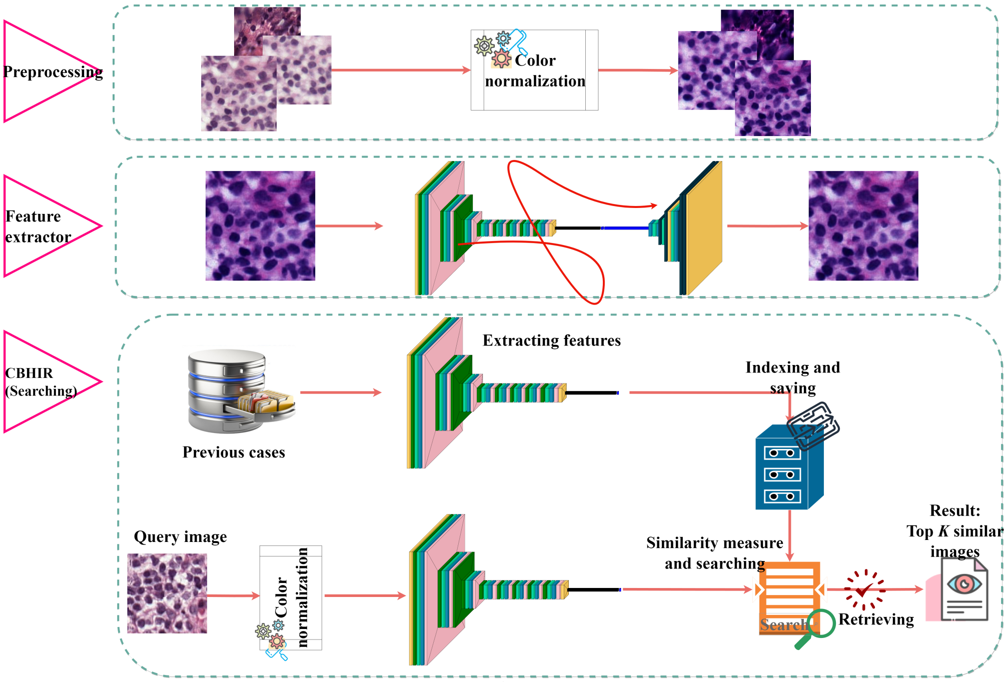


FIGURE 11: three main stages of CBHIR, including pre-processing, training, and searching. The pre-processing was done, followed by [30], [93], and [94]. The training stage contains the CAE training. Here, each layer of the CAE is presented in a different color. Conv2D, Dropout, Dense, and Flatten are shown with green, pink, blue, and black, respectively. The searching stage contains extracting features, indexing, searching, and displaying the top K similar retrieved images.

3.3.1 Pre-processing

In this paper, Mac [93], Vah [94], and BKSVD [30] are three CN techniques spanning from 2009-2022, that were investigated to normalize the data set. These three CNs were applied to the original images of the CAM17, and the results were collected as a separate data set. The aim is to gain insight into the effectiveness of these CN techniques in tackling the effects of color variation on the results of the CBHIR tool.

It is important to establish and maintain uniformity in color normalization throughout the whole framework. As can be seen in Fig 11, the input query needs to undergo the pre-processing step before feeding to the FE. Furthermore, a crucial consideration is that the CN technique, which applies to the data set in the pre-processing step, should match the CN technique employed in the query. Ensuring consistency in the CN technique is imperative for accurate processing and analysis.

3.3.2 Feature extractor

In this paper, a CAE is utilized to extract the representative histopathological features of the patches. The CAE aims to solve the back-propagation problem in an unsupervised manner by only relying on the input image as a teacher by itself [98]. It has the in-built ability to compress the data efficiently by extracting the important features and removing noise in the data [99]. These behaviors bring benefits to CBHIR, such as ignoring the noises in WSIs that might be noticeable due to the scanners [100]. Also, it can reduce the demand for annotated images for training a FE, which is expensive and time-consuming [101].

Fig 11 and Fig 12 illustrate the structure of the proposed custom-built CAE. In this CAE, a skip connection between a layer in the encoder and the corresponding layer in the decoder can improve the gradient propagation and increase the performance of capturing complex patterns of the WSIs. The proposed CAE comprises an encoder with three convolutional layers $[16, 32, 64]$. Moving to the bottleneck, an attention block with the filter size of $[32, 16, 1, 64]$ is introduced to enhance the feature vectors by robustness to noise or occlusions. The decoder is made up of three Conv2DTranspose with $[32, 16, 3]$. The kernel size in this structure was fixed to 3 in all the layers.

The primary objective of the CAE is to reconstruct input images in the output by minimizing the Mean Squared Error (MSE). The CAE workflow involves feeding an input image (X) to the encoder, which compresses the input. Subsequently, the bottleneck compresses the output of the encoder to obtain a feature vector with 200 meaningful features (F). Finally, the decoder reconstructs the output by receiving the feature vector ($Y = Decoder(F)$). In an ideal CAE, the output is identical to the input (*ideal*,$X = Y$). To achieve this ideal goal, the model endeavors to minimize the MSE by comparing the input (X) and the output (Y) [28].

In this paper, the model was trained on GPU with the *NVIDIA GeForce RTX 3090*, over 10 epochs, utilizing a batch size of 32 and a learning rate of 0.000001. By discarding the decoder from the CAE, the well-trained FE can be yielded for further experiments.

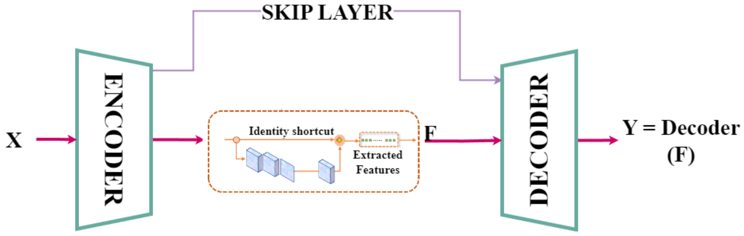


FIGURE 12: the customized CAE which was utilized as a FE.

3.3.3 Searching

The searching process in CBMIR involves three key steps: similarity calculation, ranking and retrieval, and visualization and presentation. In this paper, at this stage, the well-trained CAE extracts the features of the images of the validation and training sets. These feature vectors are subsequently indexed and saved as the features of the database for the searching process [102].

Upon receiving an input query, the proposed CBHIR passes it to the same CN technique, normalizing the whole database. For each query input, a feature vector is required, encompassing representative features of the query. The feature vector's size is fixed at 200 features, aligning with the number of elements in the feature vectors associated with the database.

After extracting the query's features, the next step involves measuring the similarity between the query feature vector and the indexed feature vectors of the database. To achieve this, a distance function is required. In this paper, Euclidean distance was selected as one of the most used distance functions in CBHIR [27]. In this context, distance is inversely related to similarity. Therefore, the most similar images have the smallest distance [24]. Consequently, the top K images with the smallest distance are retrieved and displayed to the pathologists for further analysis.

3.4 Material

CAMELYON17 challenge (CAM17) [83] is a breast cancer metastasis detection in the lymph node sections. This contains WSIs from five different hospitals. Each hospital contributes 20 patients and five slides per patient. The tissues were stained with H&E. Following [30], only the train set of CAM17 was used in the conducted experiments since the WSIs in the test set are not labeled yet.

For the experiments conducted in this paper, the first four hospitals of CAM17 were used as the train and validation set, while the 5th hospital, exhibiting a more significant color difference, was used as the test set [30]. The total number of images for training and validating the FE is 48000 and 12000, respectively. Then, in the search part, there are 25406 query images as the test set. Following [103], the experiments in this paper were performed on non-overlapping patches extracted from CAM17, each sized 224 × 224 pixels in RGB color space. The images were sampled from WSIs containing at least 70% of tissue to represent enough histopathological patterns.

3.5 Experiments and Results

In this section, the procedure is applied for each color version of the data set to report and evaluate the performance of the proposed CBHIR framework.

3.5.1 Pre-processing

Fig 13 shows the Normalized Median Intensity (NMI) information for each hospital and method, which is plotted as a violin plot. NMI is used for assessing the effectiveness of each CN technique in normalizing the data set. It involves calculating the median intensity value of pixel values within an image and then normalizing this value. NMI values were computed for individual images within the data set, and metrics such as the Standard Deviation (NMI SD) and Coefficient of Variation (NMI

CV) were employed. A lower NMI SD and NMI CV suggest a more consistent normalization across the dataset. [30] presented numerical results for BKSVD, Mac, and Vah across individual clinical centers as well as the entire CAM 17 dataset. In this study, Fig 13 visually displays the distribution, variability, and potential outliers of values across different groups, facilitating comparisons between these groups. [93]

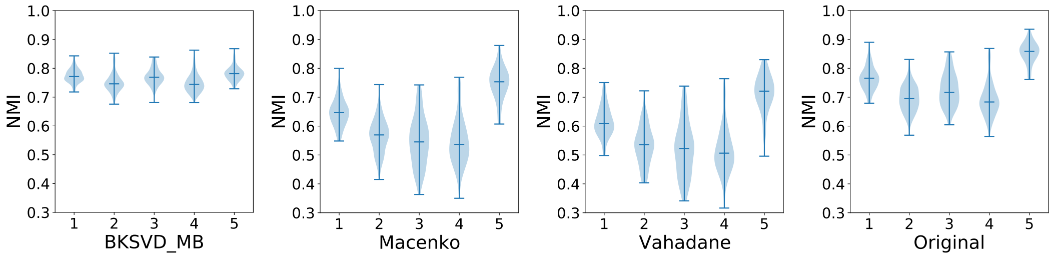


FIGURE 13: violin plots of NMI values for each center in different color versions. The histogram of NMI values for each plot is represented by the blue shadow. Bars mark the maximum, median, and minimum NMI values for each plot. The x-axis corresponds to the hospital numbers.

and [94] could transfer the color of images in each hospital to a similar distribution but in a larger inter/intra-center than the original images' distribution. Among these three CNs, BKSVD had the best performance of normalizing the images of all five centers to the same NMI interval, approximately. This means BKSVD transforms images of each hospital with the lowest intra-center variance and most similar median values, which makes its outputs more interpretable and reliable. To provide a more detailed statistical analysis to strengthen the evidence for the effectiveness of BKSVD compared with Vah and Mac Table 6 [30] provides a quantitative comparison, based on the Peak Signal to Noise Ratio (PSNR) which shows that BKSVD outperforms the rest of methods obtaining a higher mean PSNR while requiring a similar time to Vah methods.

TABLE 6: PSNR for the normalized CAM17 data set.

CN techniques	**BKSVD** [30]	**Vah** [94]	**Mac** [93]
PSNR	19.54	12.74	13.80

Fig 14 provides visual information about the impact of each CN technique on the color distribution of five random images. Randomly chosen images from all five centers exhibit noticeable color variations, as depicted in the initial line showing the original images. Notably, BKSVD effectively mitigated this color variation, achieving a harmonized and consistent color version that surpassed the performance of Vah and Mac as the classical methods.

3.5.2 CBMIR results

The most usable strategy to measure the performance of the CBHIR tool is named “top K accuracy." In this strategy, the CBHIR displays top K-matched patches. If there are one or more correct retrieved images among them, the CBHIR tool performs well [70].

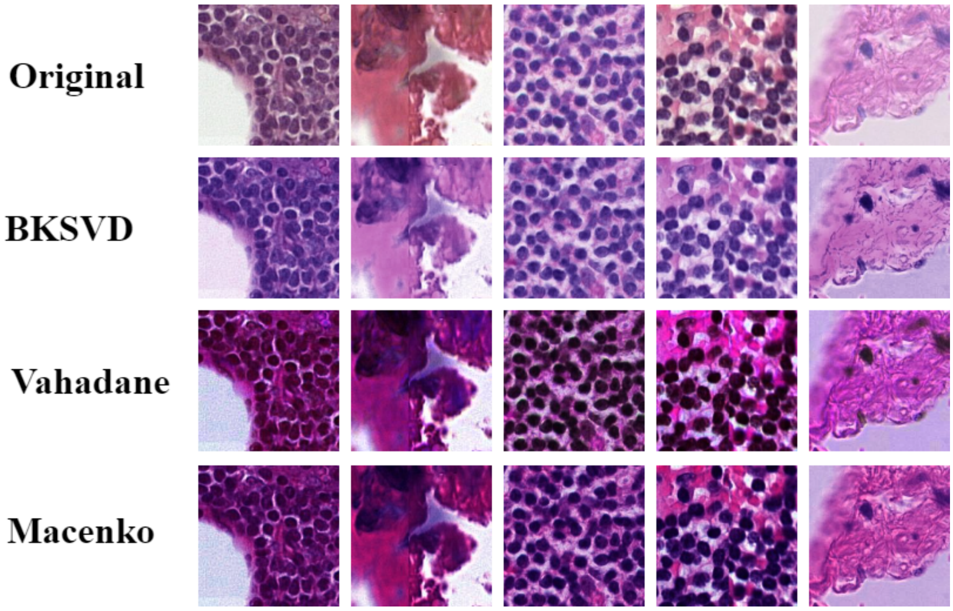


FIGURE 14: line 1 illustrates the original color version of CAM17. Lines $2-4$ contain five random images of CAM17 as a result of CN techniques, BKSVD, Vah, and Mac. These images correspond to all hospitals.

$$\begin{cases} ACC = 1 & \text{,If any of the K-top retrieved images match with the query,} \\ ACC = 0 & \text{,Otherwise} \end{cases}$$

It is noteworthy to mention that in this paper, the results are reported at the top $3, 5$, while in the recent papers [88], [13], and [104], K is $20, 100, 200$, and 400. While the amount of K in this paper is notably lower compared to the other studies, it shows the performance of the model to achieve impressive accuracy. This highlights the model's reliability for pathologists, as it can retrieve similar patches even with a smaller set of retrieved images.

Table 7 reports the results when K = 3 and 5 for CAM17 in the four color versions. As a result of matching Table 7 and Fig 13, less intra-center variance can improve the search performance. According to Table 7, CN as a pre-processing might have a negative impact on the final results. The obtained accuracy at top 3 and 5 for the experiment on the original images is higher than the accuracy of the experiments with the Vah and Mac versions of CAM17. This means that utilizing a non-sufficient CN technique not only cannot improve the final results but also can decrease the performance of the main model. Fig 27 illustrates four confusion matrices for the four color versions of CAM17 at the top 5.

TABLE 7: the achieved top K accuracy in the proposed CBIR on color-normalized images from the CAM17 data set. K = 3 and 5.

K	Original	BKSVD	Vah [94]	Mac [93]
3	0.73	**0.91**	0.66	0.68
5	0.83	**0.97**	0.79	0.81

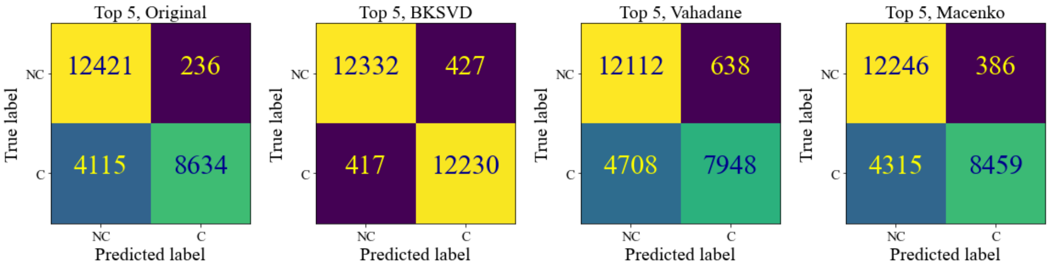


FIGURE 15: Confusion Matrices show the effects of BKSVD, Vah, and Mac on the performance of a CBHIR in retrieving the patches with the most similarity. "*C*" and "*NC*" stand for cancerous and non-cancerous tissue

According to the obtained results, thanks to CN techniques, by reducing the negative impacts of color variation, the features focus on the texture, shape, and histologic features of each patch. Therefore, the Euclidean function ranks the images which are more similar in the sense of relevant features.

3.5.3 Visual evaluation

A visual evaluation with some sample queries and their retrieved patches provides transparency and insight into the functioning of the CBMIR system. This allows pathologists to understand how the CBMIR framework responds to different queries. The results of feeding the proposed framework with the original and normalized images are presented in the four figures below.

Fig 16 illustrates four random queries in their original color space and the top 5 retrieved images. As can be seen, for each of the queries, the retrieved images have different colors not only with the corresponding query but also among themselves. This can highlight the need for color normalization as a pre-processing step for the searching framework.

Fig 17 and 18 show the same random queries as Fig 16, but they were normalized by Mac and Vah, respectively. These figures clarify the numerically reported results and confirm that the effects of Vah as a CN technique can even decrease the accuracy of retrieval.

Fig 19 corresponds to the same four queries as Fig 18 and their top 5 retrieved patches from the data set. In this figure, images were normalized by BKSVD. As can be understood from the figure, normalizing the images by BKSVD can enhance the performance of the search engine in finding similar histopathological patterns without the negative impact of color variation.

In these figures, the images were compared with their queries based on the labels provided by expert pathologists in the ground truth of the data set. The returned patches with the same label as the query were considered the correct retrieved patches. However, the images with different labels with the query were surrounded by a red square in order to clarify the miss-retrieved patches based on their labels.

3.5.4 Comparing the results of CBHIR with a classifier

Indeed, CBHIR and classification are two different CAD tools for pathologists. CBHIR provides similar patches for pathologists based on the content features of the query and the database. However, the main objective of classification is categorizing

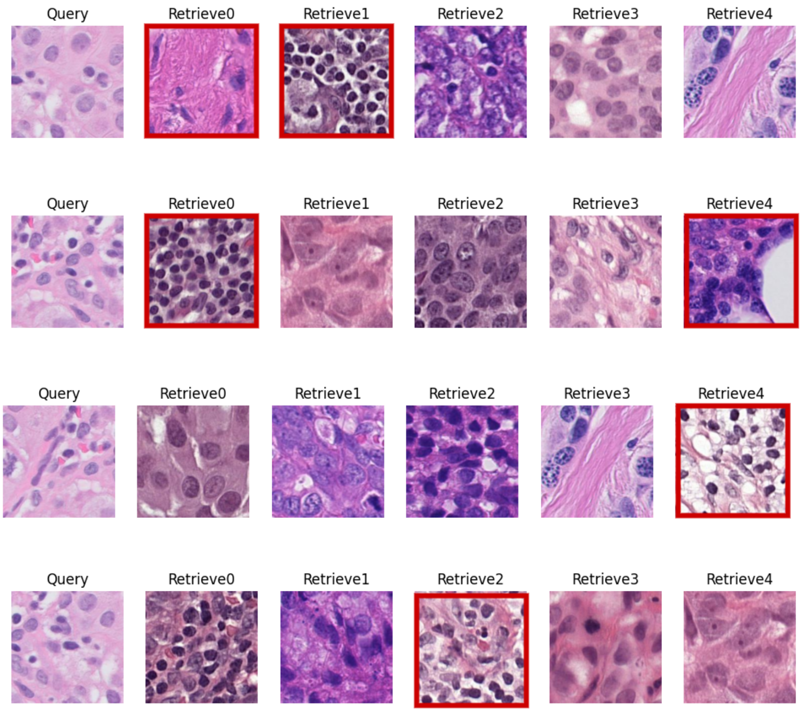


FIGURE 16: For each random query, the 5 top similar patches were presented. The red lines show the miss-retrieved patches. The patches were in their original colors.

images into pre-defined labels. In CBHIR, in addition to the corresponding label of the query, top K similar patches are shown to the pathologists. This enhances the reliability of CBHIR for pathologists as it avoids being a completely black box. CBHIR gives pathologists an opportunity to compare the histological patterns of the query with similar patches besides knowing the corresponding label.

Table 8 reports the amount of Area Under the Curve (AUC) of the unsupervised CBHIR as a result of applying all CN techniques. These results are compared with the results of applying VGG19 on CAM17 as a classifier, which is reported in [30]. Following [25], the main objective of this comparison with the supervised classifier is to evaluate the unsupervised CBHIR performance in terms of retrieving images belonging to the same cancer grade [25]. According to Table 8, BKSVD-VGG19 with 98.17% of AUC had the highest performance among (Vah, Mac, Original)-VGG19. The unsupervised BKSVD-CBHIR with 96.31% of AUC has comparable performance to the fully supervised.

TABLE 8: comparison between the performance of VGG19 as a classifier and the unsupervised CBHIR at top 5. The reported metric is AUC.

	Original	**BKSVD**	**Vah** [94]	**Mac** [93]
VGG19 (supervised)	0.941	**0.9817**	0.7985	0.9499
CBHIR (unsupervised)	0.9631	0.9754	0.9429	0.9632

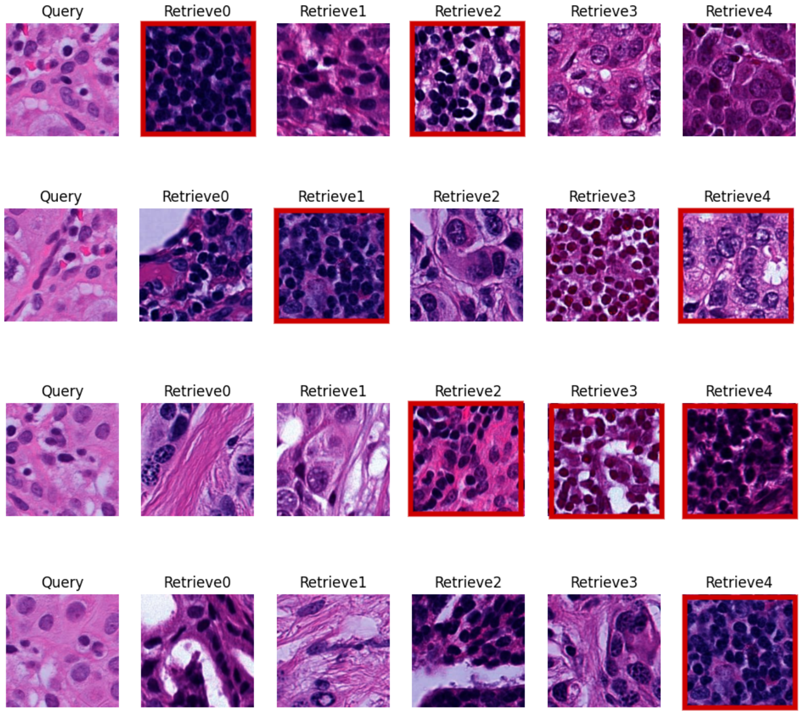


FIGURE 17: Four random patches of breast cancer data set with their corresponding top 5 retrieved images. The images were normalized by Macenko and the red lines mention the non-similar patches based on their labels.

3.6 Conclusion

In this paper, we have proposed a novel CBHIR framework for histopathological images; based on an unsupervised feature extractor and color normalization. We utilize a custom-built Convolutional-Auto Encoder (CAE) to extract the features in an unsupervised manner to tackle the challenges of lack of annotated data sets. In this feature extractor, a skip connection between a layer in the encoder with the corresponding layer in the decoder and a residual block in the bottleneck provided the meaningful features of the data set for the search engine.

The proposed framework is designed to work with data sets with intra- or inter-laboratory color variation since it solves the dependency of CBHIR on the color variation of the data set. We analyzed the effect of using color-normalized images to feed a CBHIR tool. We observed that as the effectiveness of color normalization techniques in reducing intra-center variance improved, the CBHIR results exhibited higher performance levels.

In this paper, we provided a visual evaluation in order to illustrate the results of the proposed framework visually. With this type of assessment, users can visually evaluate the quality of the retrieved patches, which might identify potential areas for improvement. Furthermore, reporting these figures can be valuable for educational purposes, enabling users to grasp the functionality and potential applications. It also facilitates communication between developers, researchers, and end-users, fostering collaboration and improving CBMIR technologies.

Finally, We have compared the results of the proposed unsupervised CBHIR

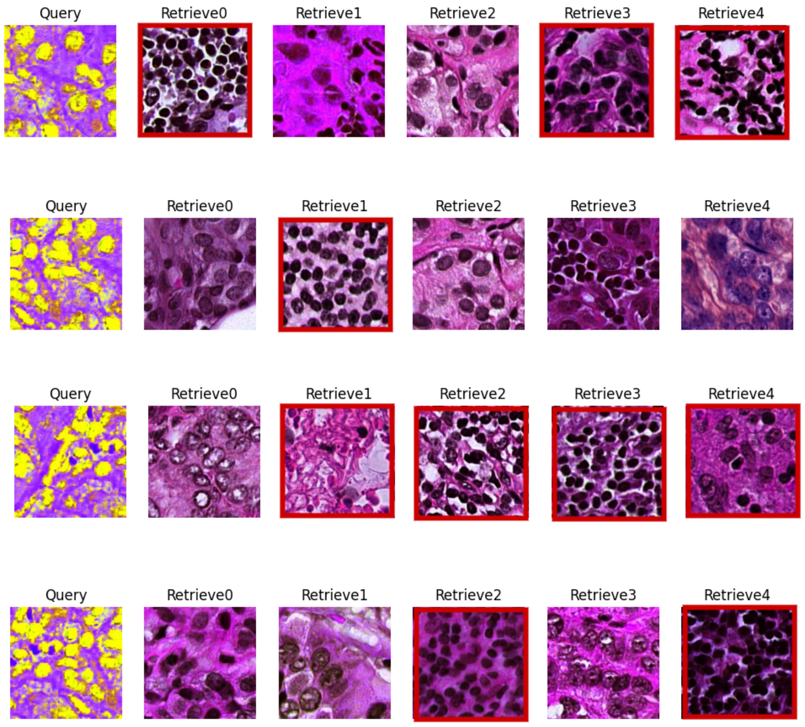


FIGURE 18: Four random breast queries with their 5 top retrieved images. The images were normalized by Vahadane. The miss-retrieved images are marked in red.

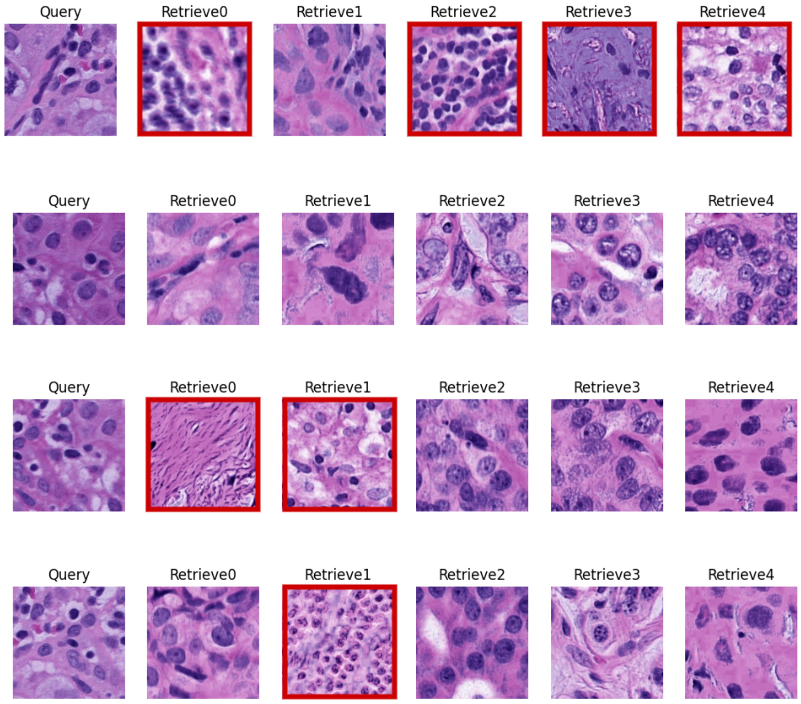


FIGURE 19: Four random queries from the normalized data set by BKSVD with their 5 top retrieved images.

framework with VGG19 classifiers to evaluate the performance of the proposed unsupervised CBHIR framework in order to retrieve images with the same cancer type. The proposed framework was found highly effective in discriminating the grades of the tissues. This observation clarifies the success of the proposed unsupervised CBHIR framework in identifying the correct histopathological features in the contents of the images.

3.7 Future work

CBMIR, a recent framework in digital pathology proposed by CAD, plays a vital role in reducing the incidence of human errors and provides an inclusive worldwide platform for pathologists with varying levels of expertise. Acting as a bridge between medicine and engineering, CBMIR fills the gaps between these two fields, offering future opportunities in both realms. From an engineering perspective, despite achieving high predictive accuracy, this study has its limitations. DL-based FEs should be trained on an extensive database, prompting the need for future works with more data and a prospective approach.

On the medical front, the entire framework can be tested in various hospitals and integrated into traditional cancer diagnosis workflows to analyze its pros and cons in real-world practice. This testing can provide a clearer understanding of a CBMIR framework and serve as a guide for subsequent steps in training DL-based methods.

CRediT authorship contribution statement

Zahra Tabatabaei: Conceptualization, Methodology, Analyzing, Writing, Reviewing & editing, Formal analysis, visualization;
Fernando Pérez Bueno: Review & Editing;
Adrián Colomer: Supervision & Review;
Javier Oliver: Supervision;
Rafael Molina: Supervision & Review;
Valery Naranjo: Supervision.

Acknowledgment

This study is funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 860627 (CLARIFY Project). The work of Adrián Colomer has been supported by Ayuda a Primeros Proyectos de Investigación (PAID-06-22), Vicerrectorado de Investigacion de la Universitat Politecnica de Valencia (UPV).

Chapter 4

Federated Content-Based Medical Image Retrieval

This Chapter corresponds to the author′s version of the following published paper: Tabatabaei, Zahra, et al. "WWFedCBMIR: World-Wide Federated Content-Based Medical Image Retrieval." Bioengineering 10.10 (2023): 1144.

WWFedCBMIR: World-Wide Federated Content-Based Medical Image Retrieval

Zahra Tabatabaei [1,2†,*], Yuandou Wang [3], Adrián Colomer [2,4], Javier Oliver Moll [1], Zhiming Zhao [3], and Valery Naranjo [2]

[1] Dept. of Artificial Intelligence, Tyris Tech S.L., Valencia, Spain.
[2] Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano, HUMAN-tech, Universitat Politècnica de València, Spain.
[3] Universiteit van Amsterdam, The Netherlands.
[4] ValgrAI – Valencian Graduate School and Research Network for Artificial Intelligence.

Abstract

The paper proposes a Federated Content-Based Medical Image Retrieval (FedCBMIR) tool that utilizes Federated Learning (FL) to address the challenges of acquiring a diverse medical data set for training CBMIR models. CBMIR is a tool to find the most similar cases in the data set to assist pathologists. Training such a tool necessitates a pool of Whole Slide Images (WSIs) to train the feature extractor (FE) to extract an optimal embedding vector. The strict regulations surrounding data sharing in hospitals makes it difficult to collect a rich data set. FedCBMIR distributes an unsupervised FE to collaborative centers for training without sharing the data set, resulting in shorter training times and higher performance. FedCBMIR was evaluated by mimicking two experiments, including two clients with two different breast cancer data sets, such as BreaKHis and Camelyon17 (CAM17), and four clients with BreaKHis data set at four different magnifications. FedCBMIR increases the F1-Score (F1S) of each client from 96% to 98.1% in CAM17 and from 95% to 98.4% in BreaKHis, with 11.44 hours less in training time. FedCBMIR provides 98%, 96%, 94%, and 97% of F1S in the BreaKHis experiment with a generalized model and does so in 25.53 hours less training.

4.1 Introduction

Breast cancer accounts for 25% of all cancers in women worldwide. According to the American Cancer Society, a woman is diagnosed with breast cancer in the world every 14 seconds. In the year 2020, approximately 2.3 million women were diagnosed with breast cancer globally, and 685,000 lost their lives due to it[105]. Histopathology is commonly used in the diagnosis and treatment of various diseases, including cancer. A biopsy, which is the removal of a small piece of tissue from the body, is usually required for histopathological examination [106]. Human error in histopathology refers to mistakes or inaccuracies made during the process of examining tissues or cells under a microscope [107]. Some examples of human errors in histopathology include, sampling errors, processing errors, technical errors, interpretation errors, and reporting errors [108]. To minimize human errors in histopathology, it is essential to follow strict protocols and guidelines, perform regular quality control checks, and ensure that all personnel involved in the process are properly trained and competent [109]. Authors in [110] analyzed the accuracy of breast cancer diagnosis in 102 cases and found that there were diagnostic errors in 15.7% of cases. The most common types of errors were misclassification of tumor type and misinterpretation of pathology slides. Digital pathology could help pathologists to improve the accuracy and efficiency of cancer diagnosis, reduce the risk of errors, and enhance patient care.

Digital pathology is a technology that uses digital images of tissues and cells to aid in the diagnosis and management of diseases [111]. Deep Learning (DL) has revolutionized Computer-Aided Diagnosis (CAD) in digital pathology and has opened the door to improve cancer diagnosis while decreasing the pathologist's workload [35].

Content-Based Medical Image Retrieval (CBMIR) is a recent DL-based methodology that allows pathologists for a quick and precise search in previously diagnosed and treated cases [88]. In CBMIR, image features such as texture, shape, color, and intensity are extracted from the query and data set; then, a similarity measure is applied to compare the query features with the features of the database [112]. The retrieved images are ranked according to their similarity to the query image, and the most relevant images are displayed to the user.

To further illustrate the advantages and practicality of CBMIR in the field of histopathology and cancer diagnosis, consider a scenario where a patient is diagnosed with cancer, and grading it accurately poses a challenge for pathologists. In

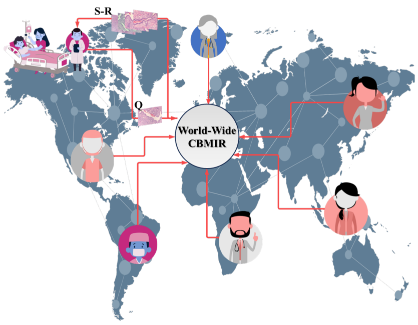


FIGURE 20: An overview of the use case of a worldwide CBMIR. Pathologists send their Query (**Q**) to the worldwide CBMIR since they need a second opinion to make a more confident decision. Then, the model retrieved top K similar images (**S-R**), and the pathologists can get a second opinion from whole over the world.

traditional cancer diagnosis methods, the pathologist would need to physically send the glass slide containing the tissue sample to another hospital, which could be located in a different city or even a different country. This process is not only expensive and time-consuming but also carries inherent risks, such as the loss or damage of the glass slide during transportation. Moreover, it adds additional stress to the patient's already difficult situation.

By implementing a World-Wide Content-Based Medical Image Retrieval (WWCBMIR), these challenges can be effectively addressed, and the process of a cancer diagnosis can be significantly expedited without compromising accuracy. Through the use of digital pathology, where Whole Slide Images (WSIs) are digitized and stored electronically, pathologists can access and analyze the images remotely [106]. The WWCBMIR enables pathologists to retrieve similar cases and relevant information from a vast database of histopathological images without the need for the physical transfer of slides. This approach not only reduces costs and saves time but also minimizes the potential risks associated with the transportation of delicate tissue samples. Figure 20 shows how a WWCBMIR can provide unprecedented access to K number of patches with the most similar patterns, allowing the pathologists to make a more confident diagnosis.

One of the advantages of CBMIR from the pathologist's (user) perspective is that it is not a completely black box for them. CBMIR allows pathologists to find similar patterns among the retrieved images and the queries based on their knowledge. This provides more reliable information than a label for pathologists, which makes CBMIR more beneficial for pathologists than a classification.

An actual context needs a global CBMIR, which demands a generalized data set with a variety of images of different quality, magnification, color, size, etc. The performance of CBMIR relies on a vast amount of data, which is difficult to collect in the medical field due to patient privacy and time costs. In order to create a vast centralized data set, DL experts need to transfer their WSIs. However, these images are gigapixels with high storage sizes. In addition to the challenges of transferring a heavy data set for DL experts, patient privacy policies and other regulatory obstacles on the medical side make it more challenging to create a sufficient data set.

Federated Learning (FL) represents a possible solution to tackle this problem by collaboratively training DL models without transferring WSIs [113]. Multiple institutions can safely co-train DL models in digital pathology using FL, achieving cutting-edge performance with privacy assurances [114]. FL brings an opportunity to share the weights for multi-institutional training without sharing patient data and images. However, there are still some privacy risks since the training parameters and model weights are distributed among collaborators [115].

DL models give information that goes beyond the scope of human vision, and FL solves the problem of data sparsity by connecting international hospitals while complying with the data privacy policies, irrespective of the country of origin. This benefit can remedy the health care limitations due to the lack of facilities (staining materials, scanners, etc.) and experience (students, recently graduated pathologists, etc.). Moreover, it can tackle the lack of data sets of labeled WSIs because of data privacy.

In this paper, we minimized the WWCBMIR to an international CBMIR by leveraging FL. The experiments were conducted through the collaboration of two countries and three cities to examine the feasibility and challenges associated with implementing a WWCBMIR. This international CBMIR was trained with the data collected from different hospitals and answered the needs of clients. Clients might be expert pathologists or a student. Our main contributions include the follows:

- We proposed a novel international FL-based CBMIR, which is named FedCBMIR, to aid pathologists in breast cancer diagnosis.
- An unsupervised network was used as a Feature Extractor (FE) to extract the features of the images for the tasks trained with scanty data sets.
- We proposed a custom-built Convolutional Auto Encoder (CAE) to learn the dependencies and extract the features of the images with higher discriminating values.
- In order to address patient data privacy concerns, we employed the privacy preservation capability of FL. This approach ensures that the data in each institution remains decentralized and confidential, as there is no need to be shared with a central server.
- Through extensive tests on varying data set distributions among individual clients, we verified the robustness of our proposed solution. It proved to be independent of the data quality held by each client.

4.2 Related work

Recently, researchers have directed their attention toward both FL and CBMIR and have invested their efforts in exploring these fields. This section provides a succinct overview of some of the notable studies.

4.2.1 Content-Based Medical Image Retrieval (CBMIR)

CBMIR has been a subject of extensive research since the advent of large-scale databases nearly two decades ago, as noted by Wang [116]. Several studies have made significant contributions to this field. Tabatabaei [24] achieved an accuracy rate of 84% in CBMIR using the largest patch-annotated data set in prostate cancer. Kalra [13] proposed Yottixel, a method for representing The Cancer Genome Atlas Whole Slide Images (TCGA WSIs) compactly to facilitate millions of high-accuracy searches with low storage requirements in real-time. Conversely, Mehta [117] proposed a CBMIR system for sub-images in high-resolution digital pathology images, utilizing scale-invariant feature extraction. Lowe [118] utilized Scale-Invariant Feature Transform (SIFT) to index sub-images and reported an 80% accuracy rate for the top 5 retrieved images. Lowe's experiments were conducted on 50 ImmunohHistoChemistry (IHC) stained pathology images at eight different resolutions. Additionally, Hegde [70] used a manually annotated data set pre-trained on a Deep Neural Network (DNN) to achieve top 5 scores for patch-based CBMIR at different magnification levels. The primary focus of recent studies has been on enhancing the performance of CBMIR in different types of cancer; however, there are still several challenges that can impede its effectiveness. These challenges include data privacy, as medical data is confidential and subject to strict privacy regulations, making it arduous to share and access large data sets for model training. FL can alleviate this issue by facilitating distributed model training on local data without compromising privacy. Another challenge is data distribution, as medical data is frequently dispersed across numerous locations, it is difficult to train models on a centralized data set. FL enables the training of models across multiple distributed data sets without aggregating the data in a central location. In addition, medical data sets can be heterogeneous, varying in terms of imaging modalities, quality, and annotation protocols, which can impede the development of robust and accurate models. FL can mitigate this challenge by allowing models to be trained on diverse data sets in different qualities, improving their performance and generalization ability. Furthermore, medical data sets can be large and complex, necessitating significant computational resources to train models. FL can distribute the computational workload across multiple devices and locations, enhancing scalability and reducing training time.

4.2.2 Federated Learning (FL)

In recent years, FL has achieved impressive progress that enhances a wide adoption of DL from decentralized data [113, 119, 120]. FL is a distributed machine learning approach that can effectively handle decentralized data without raw data exchange to train a joint model by aggregating and distributing local training. Many existing algorithms can be adopted to aggregate updates from distributed clients. Typical examples include FederatedAveraging *-viz* FedAvg [113], and adaptive federated optimization methods [120], e.g., FedAdagrad, FedYogi, and FedAdam. Some popular FL frameworks such as TensorFlow Federated (TFF) [1], PySyft [121], and Flower [122] provides a set of the robust set of tools for building privacy-preserving ML models. Besides, Jupyter Notebook-based tools such as [123] also help simplify the FL setup and enable its deployment of a cross-country federated environment in only a few minutes. Daniel Truhn in [124] employed Homomorphic encryption to protect the model's performance while training by encrypting the weight updates before sharing them with the central server. Firas Khader in [125] presented a technique of "learnable synergy", where the model only chooses pertinent interactions between data modalities and maintains an"internal memory" of key information. Micah J. Sheller [115] investigated that FL among ten institutions is 99% as efficient as those derived using centralized data. One recent work related to content-based image retrieval is introduced in [126], where FLSIR was proposed, and it enables secure image retrieval based on FL and additive secret sharing. Nevertheless, it is not for clinical applications. Although the combination of CBMIR and FL is a relatively new area of research, it has the potential to greatly improve healthcare outcomes. By offering healthcare professionals quick access to accurate and relevant medical image data while maintaining patient privacy, the integration of these techniques can have a significant impact on the field.

The following sections address how the proposed FedCBMIR approach can revolutionize how medical images are searched and utilized, leading to improved diagnoses and treatment plans.

4.3 Experiments

In this section, the proposed FedCBMIR tool is introduced along with the training details and the two data sets used in our study. Figure 21 [2] provides an overview of the CBMIR workflow, starting from the initial stage at a hospital and concluding with the presentation of the top K similar patches to the user. In the medical session, a cancer patient's tissue is obtained, scanned, and divided into patches for storage. In the offline session, the FE is trained, and the extracted features from the database are saved and indexed. In the online session, a pathologist uploads an image to the CBMIR model, where the well-trained FE extracts its features. These features

[1] https://www.tensorflow.org/federated

[2] BreaKHis images

are then used by the search engine to retrieve the top K similar patches from the stored database in the medical session. Finally, in the visual session, the pathologist can reach similar patches and their corresponding labels for further investigation based on their knowledge. In this paper, the proposed FedCBMIR, as shown in Algorithm 1 [3], addresses the described challenges and provides a second opinion for pathologists in writing their reports for a cancer diagnosis. FedCBMIR is inspired by a great vision of a WWCBMIR that effectively manages decentralized medical images by utilizing local training for multiple tasks while avoiding the need for raw data exchange. FedCBMIR takes advantage of FL since it can give CBMIR a higher chance of generalizing its capabilities by accessing multi-central images from different hospitals. A generalized CBMIR framework needs more effective content of the images as the key factor in the field of CBMIR.

In this paper, we cope with the challenges of CBMIR with two different experiments and evaluate it in three scenarios. In our first experiment (EXP 1), we mimic a case of two institutions that have different breast cancer WSIs in completely different image preparation processes. This case occurs when two institutions have a limited number of images, but they need a well-trained model to obtain a supportive idea on their query tissue. This experiment was assessed this experiment on CAMELYON17 (CAM17) and BreaKHis at 40× magnification. Then, in the second experiment (EXP 2), we extended our work with patches at different magnifications by feeding our FedCBMIR framework with BreakHis data set at 40×, 100×, 200×, and 400× magnification. The magnification problem in WSI analysis is the subject of our second experiment. Algorithm 1 shows FedCBMIR step by step. The novelty of this work relies on providing well-trained models that can retrieve similar patches for each client in different countries. Regarding the use of FL in CBMIR, all clients, regardless of their data privacy policies, can train the model with a limited number of patches and find similar patches to their queries more accurately than local training.

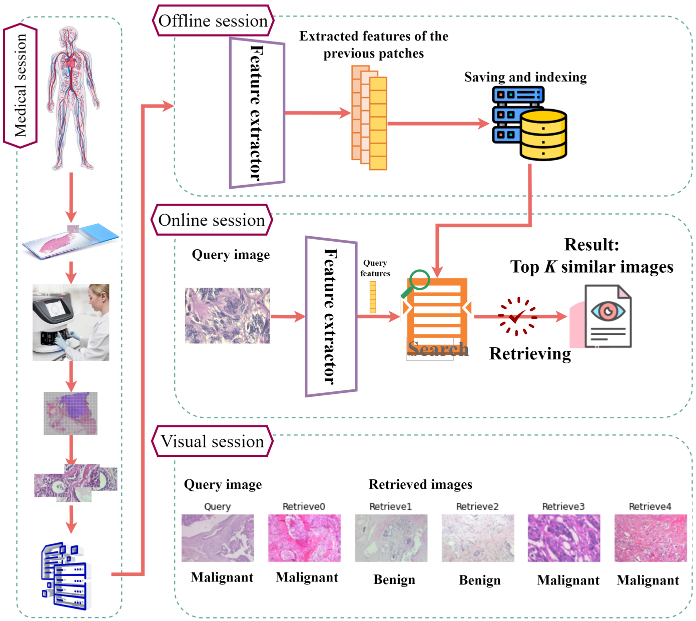


FIGURE 21: A comprehensive illustration of the entire process in a CBMIR, demonstrating the utilization of DL models to acquire images from a hospital and offer a second opinion for pathologists.

[3] More information: https://flower.dev/docs/framework/how-to-implement-strategies.html

The performance of FedCBMIR was validated on histopathological images using a CAE in a cross-institutional distributed environment. FL was used as a collaborative learning paradigm, in which the CAE can be trained across different institutions without explicitly sharing data sets.

4.3.1 Materials

Hematoxylin and Eosin (H&E) is a type of histopathological staining. H&E has been popular for almost a century because it may indicate morphological changes [127]. The images in the used data sets in this paper were stained by H&E.

BreaKHis

BreaKHis contains 7,909 histopathological images of breast tumor tissues that were provided by a collaboration with the P&D Laboratory—Pathological Anatomy and Cytopathology, Parana, Brazil. This data set was collected from 82 patients at four magnifications (40×, 100×, 200×, and 400×) with 2,480 benign and 5,429 malignant cases. As can be understood in Table 9, the number of images in benign and malignant cases is imbalanced. The most considerable portion of the data set belongs to the images at 100× magnification [4].

TABLE 9: The distribution of BreakHis data set.

Magnification	Benign	Malignant	Total
40×	625	1370	1995
100×	644	1437	2081
200×	623	1390	2013
400×	588	1232	1820
Total	2480	5429	7909

CAMELYON17 (CAM17)

The CAM17 data set belonging to the CAMELYON17 challenge, as described by [128], is designed to detect breast cancer metastasis in lymph node sections. It comprises 1000 WSIs obtained from five distinct hospitals. Each hospital contributed data from 20 patients, with five slides per patient, and annotations for cancer regions were provided for a subset of 50 WSIs. In this paper, images from four hospitals were used for training and validating the model, and the images of Hospital 5 were fed into the model as a test set. Non-overlapping 224×224 (at 40×) pixel patches with at least 70% tissue were used for experiments on this data set. In the experiments of this paper, the data set was considered as a binary data set, including Cancerous (annotated) and Non-Cancerous (not annotated).

[4] https://www.kaggle.com/datasets/ambarish/breakhis

Algorithm 1 FedCBMIR(FedAvg)

```
Server (Aggregator)                                   Client (CBMIR)

Initialization .......................................................................

$M$ ◁ The number of clients
$R$ ◁ The communication rounds
$E$ ◁ The local epochs
$B$ ◁ The local batch size                            $H$              ◁ hyperparameters
$\eta$ ◁ The local learning rate                      $\mathcal{M}$    ◁ model structure
$\omega_0$ ◁ weights                                  $D_m$            ◁ local data set of client m

Phase 1 ..............................................................................

for all round $r = 1, 2, ..., R$ do
    $S_r =$ (random set of $M$ clients)
    for all client $m \in S_r$ do
        $\omega_{r+1}^m = \text{ClientUpdate}(m, \omega_r)$
                                                      ClientUpdate:    ◁ execute on client m
                                                          train $D_m$ with model $\mathcal{M}$ structure
                                                          $\beta \leftarrow$ (split $P_m$ into batches of size $B$)
                                                          for all local epoch $i = 1, 2, ..., E$ do
                                                              for all batch $b \in \beta$ do
                                                                  $\omega = \omega - \eta \nabla \iota(\omega; b)$
                                                          return $\omega$ to server

Phase 2 ..............................................................................
FederatedAveraging: ◁ execute on server
    $\omega_{r+1} = \sum_{m=1}^{M} \frac{n_m}{n} \omega_{r+1}^m$
```

4.3.2 Data distribution

CLoud ARtificial Intelligence For pathologY (CLARIFY) project [5] has a multi-institutional paradigm. In this work, according to the connections between different institutions in CLARIFY, four institutions (three universities and one company) in three cities in two countries gathered to mimic the practical situation of FL in CBMIR.

In EXP 1, in order to distribute the data into two nodes, we assume that Tyris (TY)[6] and the Universiteit van Amsterdam (UvA) [7] have CAM17 and BreaKHis 40×, respectively. As can be seen in Table 10, TY caries out training the FedCBMIR on a GPU resource in the type of NVIDIA GeForce RTX 3090. The used GPU in UvA is the NVIDIA Tesla T4, which has fewer CUDA cores, slower memory clock speed, and lower memory bandwidth compared to the used GPU in TY. These different GPUs are chosen to mimic the real condition that different hospitals or research centers have different GPU performances.

In EXP 2, regarding mimicking the real-world data limitation, the four magnifications of the data set were distributed into four nodes. To do so, each institution (client) in this paper has BreakHis at only one magnification to train their model (Table 11). Universidad de Granada (UGR)[8], TY, UvA, and Universidad Politécnica de Valencia (UPV)[9] trained the custom-built CAE with BreakHis 40×, 100×, 200×, and 400×, respectively. To replicate real-world conditions where clients may not have access to high-performance GPUs, our experiment includes three distinct GPU types across four institutions. This ensures alignment with practical scenarios and provides a comprehensive evaluation of different GPU capabilities.

TABLE 10: Data distribution in EXP 1. The information of each institution participating in EXP 1, including their location, the name of their center, the data associated with their data distribution, and the GPUs employed by each client for training and searching tasks.

Client	Region	Institution	Data set	GPU type
1	Valencia, Spain	TY	CAM17	NVIDIA GeForce RTX 3090
2	Amsterdam, The Netherlands	UvA	BreakHis 40×	NVIDIA Tesla T4

4.3.3 Training Convolutional Auto Encoder in each node

One of the most crucial elements of CBMIR that influences search engine results is the FE. The objective of content-based image search is to efficiently compare an extracted feature from a query image to every image in a database to identify the matches that are most similar.

Lack of annotated images and bias are the two major challenges that need to be considered in the integration of DL into cancer diagnosis. Three factors have the potential to make bias in medical studies: data-driven, algorithmic, and human bias. To tackle these obstacles, a custom-built CAE is configured as the FE in this paper as a generative model where it is trained to reconstruct its input in an unsupervised

[5] http://www.clarify-project.eu/

[6] Spain, Valencia

[7] The Netherlands, Amsterdam

[8] Spain, Granada

[9] Spain, Valencia

TABLE 11: Collects information on each client in EXP 2, including their country and city, the name of the center, the related data due to the data distribution, and the GPUs used for training and search tasks by each client.

Client	Region	Institution	Magnification	GPU type
1	Granada, Spain	UGR	40×	NVIDIA GeForce RTX 3090
2	Valencia, Spain	TY	100×	NVIDIA GeForce RTX 3090
3	Amsterdam, The Netherlands	UvA	200×	NVIDIA Tesla T4
4	Valencia, Spain	UPV	400×	NVIDIA TITAN V

way. The proposed structure of CAE contains a skip layer to jump over the layers to not only lead the model to converge faster and minimize the training errors but also boost the representation power and tackle the vanishing problem. Also, it has a residual block in its bottleneck to enable the training of deeper and more accurate CAE.

Figure 22 shows its architecture with convolutional filters in the size of $[32, 64, 128, 256]$ in the encoder and, respectively, $[128, 64, 32, 3]$ in the decoder. In this custom-built CAE, a residual block with the filter size of $[64, 32, 1, 256]$ takes place between the encoder and decoder. This takes the originally extracted features from the backbone as its input and provides a new feature map that contains the context relations between its feature input. In our experiments a skip layer connect a layer in the encoder to the corresponding layer in the decoder. The bottleneck delivers one feature vector with 200 features ($F_i = \{f_1, f_2, f_3, ..., f_{200}\}$) from each encoded input image i. The model aims to achieve the lowest Mean Squared Error (MSE) by comparing Input (I) and Output (O) and is penalized if the reconstruction O differs from I. Once the unsupervised training is completed by discarding the decoder part, a powerful automatic FE is available to extract the desired features.

4.3.4 Local Training

Figure 23[10] explains the whole pipeline of the proposed CBMIR that each institution must follow to retrieve similar patches. In the offline session, images in the training and validation set are fed into the FE to extract and save their features as in the

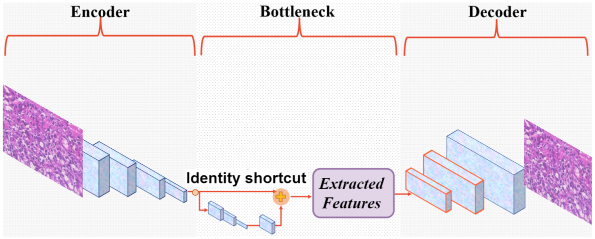


FIGURE 22: The structure of the custom-built CAE. The stride in the encoder = $[1, 2, 2, 2]$, in the bottleneck = $[1, 1, 1, 1]$, in the decoder related to the encoder = $[2, 2, 2, 1]$. The kernel size of the layers in all parts of the structure and for each layer is 3.

[10] BreaKHis images are used to plot the figure.

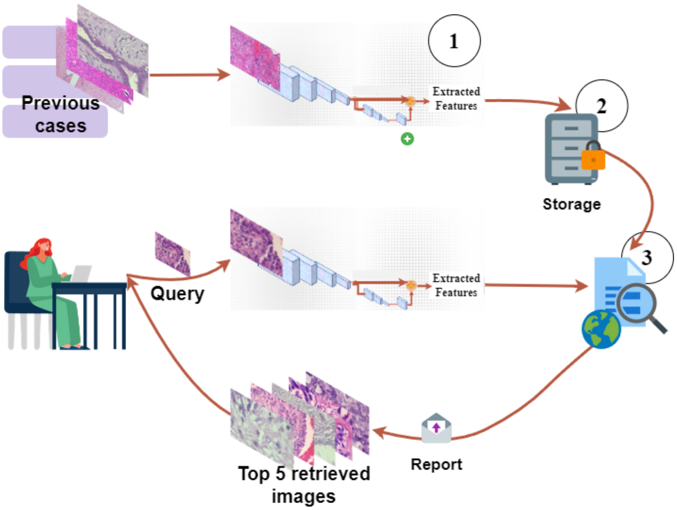


FIGURE 23: The pipeline of CBMIR. It contains three important sections such as 1) FE, 2) indexing and saving, and 3) similarity measure and search.

previous cases. All the F_is are collected in a dictionary $D = [F_1, F_2, ..., F_n]$ in the middle of this figure.

In the online session, pathologists upload their patch as a query image (Q) and expect to receive top K similar patches. In practice, each Q needs to feed to the FE and map to its feature vector F_Q. Then, F_Q feeds to the distance metrics in order to compare with the F_is saved in D. To do so, in our experiments, as soon as the pathologists upload their Q, the Q image is fed to the FE to extract F_Q with 200 features. Then, the Euclidean function applies on both F_Q and the F_is in D to measure their similarity and deliver top K similar patches.

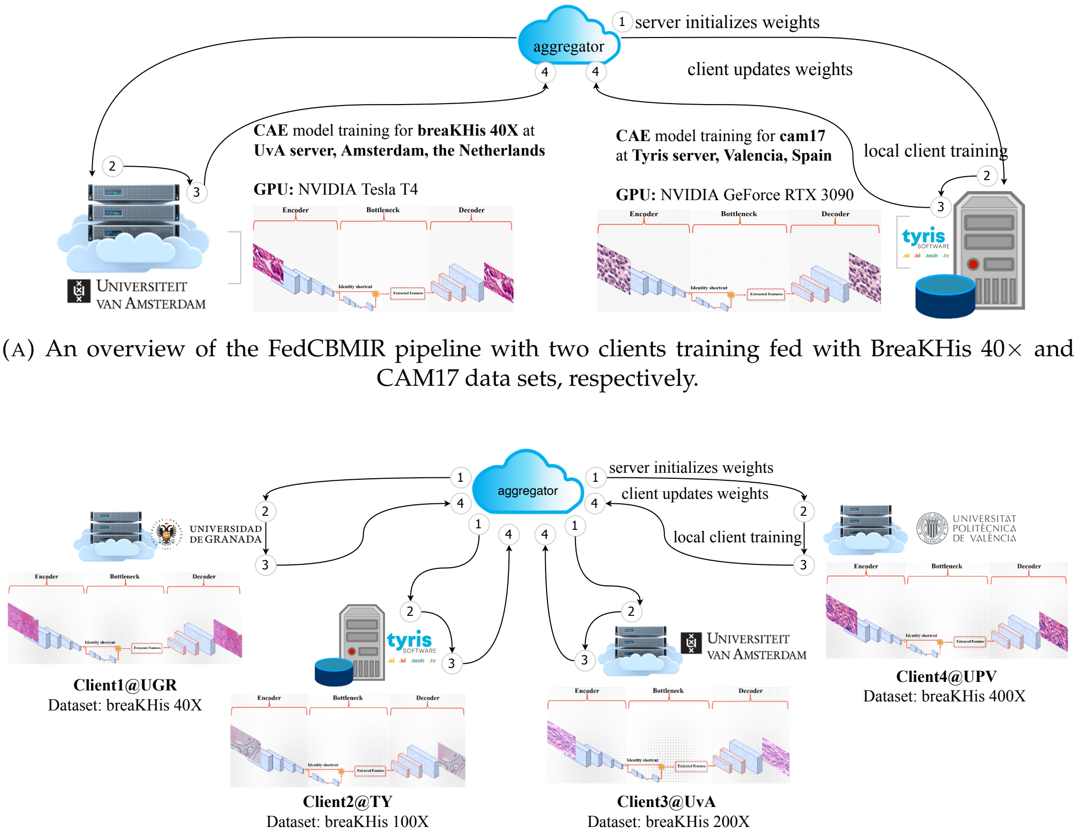
4.3.5 Federated learning configuration

In order to train the CBMIR following a federated strategy, different experiments have been conducted on FedAvg and FedAdagrad. In our cases, with some experiments, it is found that FedAvg performs better than FedAdagrad. Thus, this work adopts FedAvg to aggregate distributed updates from local clients, as shown in Algorithm 1:

$$\omega_{r+1} = \sum_{m=1}^{M} \frac{n_m}{n} \omega_{r+1}^{m} \tag{4.1}$$

where M indicates the number of clients, r presents the communication round. For a client m with n_m samples, the local updates are arbitrary ω_{r+1}^{m}.

FLOWER [122] as a primary framework is applied to configure the FL experiments. Two FL experiments were conducted, as shown in Figure 24a and Figure 24b. The first experiment consists of two distributed training nodes located in TY and UvA, respectively (see Figure 24a). In the two communication rounds, the learning rate is set as 0.000001, 5 local epochs for CAM17 per round, and 100 local epochs for BreakHis 40×. Also, FedCBMIR is extended with more clients, as shown in Figure 24b. The system consists of four separate nodes, each of which is trained using the BreaKHis data set at different magnifications. The training process involves three communication rounds, a learning rate of 0.000001, and each client performs 100 local epochs per round. Table. 11 lists all four distributed processing nodes' information in the training phase.



(A) An overview of the FedCBMIR pipeline with two clients training fed with BreaKHis 40× and CAM17 data sets, respectively.

(B) An overview of the FedCBMIR pipeline with four clients training over clusters at universities and companies with BreaKHis in four different magnifications.

FIGURE 24: The FedCBMIR pipeline consists of four main steps. Step 1: the server initializes weights, and then sends to client for local training, step 2: client starts local training, step 3: client updates local weights to the server side, and step 4: the server side aggregates and updates the distributed weights.

4.4 Discussion and Results

4.4.1 Evaluation

To allow for an adequate comparison of the model's performance, three metrics were selected: Accuracy (ACC), Precision, and F1Score (F1S), in addition to presenting the Confusion Matrix (CM). Accuracy assesses how well a model correctly retrieved similar patches to the query [129]. Precision measures the accuracy of positive predictions, which is vital when false positives are costly. The F1S combines precision and recall into a single metric [130]. In this paper, to evaluate the proposed FedCBMIR, each of the images in the test set was considered a query. Across the entire training and validation set, the model is meant to detect similar patches.

It is worth considering what "accuracy" means in the context of a CBMIR. The accuracy of CBMIR depends on what we are looking for and what is displayed by the search engine. In order to determine the performance of the experiments top K score at retrieving images of the same histologic features engaged in the prior research. The evaluation method will consider a correct answer from the model whenever it finds at least one correct image within the K set [24]. In this paper, we set $K = 5$, which evaluates the performance of our model to correctly present at least

TABLE 12: Provides a comparison in the test set between the performance of CBMIR and FedCBMIR in the EXP 1 as a result of aggregating CAM17 and BreaKHis 40× with 2 communication rounds. Hours and seconds, respectively, are used to measure the periods of training and searching.

Data	**Model**	**Accuracy**	**Precision**	**F1S**	**Training time**	**Searching time**
CAM17 (TY)	CBMIR	0.96	0.96	0.96	8.7 h	0.28 S
	FedCBMIR (Fedavg)	**0.981**	**0.970**	**0.981**	**6.21** h	0.29 S
	FedCBMIR (FedAdagrad)	0.98	0.97	0.98	7.92 h	0.30 S
BreaKHis 40× (UvA)	CBMIR	0.93	0.94	0.95	9.33 h	0.018 S
	FedCBMIR (Fedavg)	**0.978**	**0.969**	**0.984**	6.59 h	0.024 S
	FedCBMIR (FedAdagrad)	0.94	0.92	0.96	**6.11** h	0.04 S

one correct result in the top K retrieved images.

$$ACC@K = \frac{1}{N}\sum_{i}^{N}\varepsilon(\alpha_i, TOP(ans[:K])) \tag{4.2}$$

In this equation, N denotes the number of query patches, and α_i represents the label of the i-th query patch. The function $TOP(ans_i[:K])$ retrieves the top k most similar results for the query and outputs 1 if any of these results match with the query, and 0 otherwise. In other words, if $TOP(ans_i[:K])$ belongs to the set of labels of the i-th query, denoted by α_i, the function $\varepsilon()$ returns 1.

4.4.2 Results of EXP 1

For this particular experiment, BreaKHis 40× and CAM17 data sets were aggregated to train the model. As a result, each client (UvA and TY) could develop a well-trained model to retrieve their respective images. The underlying assumption made in this experiment is that neither client had an agreement in place for sharing or accessing each other's images. Table 12 provides a comprehensive view of the model. As it is mentioned above, CAM17 was provided by five hospitals. To do this evaluation, the CAM17 images from Hospital 5 were isolated from the images in the other four hospitals that were utilized for the training and validating task. Each image from Hospital 5 serves as a query in the testing assignment, and the platform's function is to seek patches with a similar pattern from the other four hospitals. Table 12 illustrates that the accuracy of local training of CAM17 without aggregating with BreaKHis is less than the FedCBMIR with aggregated data. This table indicates that FedCBMIR using the Fedavg approach, achieved better results than FedCBMIR using FedAdagrad. As a result, Fedavg was selected as the aggregation technique for the subsequent experiments.

In terms of time and accuracy, local training of the CBMIR model on BreaKHis40× and CAM17 requires 9.33 and 8.7 training hours, resulting in an accuracy of 93% and 96% in the test set, respectively. However, FedCBMIR was trained more efficiently and achieved a higher accuracy level of 98.1% in retrieving similar patches in CAM17, and 97.8% accuracy for UvA, with a reduction of 2.49 and 2.74 hours in training time, respectively. In order to have two distinct models on both data sets separately, 18.04 hours is needed while FedCBMIR trains two generalized models on both data sets in 6.59 hours ($Max(6.21h, 6.59h) = 6.59h$). This means FedCBMIR provides more generalized models for clients in 11.44 hours faster.

Training time and accuracy are essential factors for DL scientists in building an optimal model, whereas accuracy and searching time are crucial for pathologists in retrieving similar patches. The table shows that TY client can obtain a second opinion with labels and similar patches in only 0.28 seconds per image. Upon examining the "Training time" and "Searching time" columns, it becomes evident that the utilization of FL has no noticeable impact on the searching time, while it substantially influences in reducing the training time.

Figure 25 represents three random queries in the test set of CAM17 with their top 5 retrieved images among the training and validation sets. Figure 26 represents the comparison of image search results with two CMs in the test set of CAM17 as a result of local training (CBMIR) and FedCBMIR.

4.4.3 Results of EXP 2

In EXP 2, the performance evaluation of the proposed framework was conducted using two distinct scenarios. The first scenario, ***Sen1***, assumed that the clients did not have access to images from other clients, and it was only allowed to share the model weights during the training phase. This scenario was designed to test the performance of the framework when the participating clients faced technical limitations in sharing large amounts of medical imaging data. In this scenario, each client had to train their model on their local data, and the models' weights were shared with other clients. Then, the weights were combined and trained using the entire data set from all participating clients. Finally, the model was evaluated on each client's local test set.

Sen1 mirrors the situation where clients can only obtain patches that are similar to their Q at the same magnification. Because there is no explicit agreement among the institutions, the model is obliged to search for similar cases in a few cases at that particular magnification.

Table 13 summarizes the results of the proposed FedCBMIR on the BreaKHis data set at all four magnifications. This table shows the accuracy and precision of

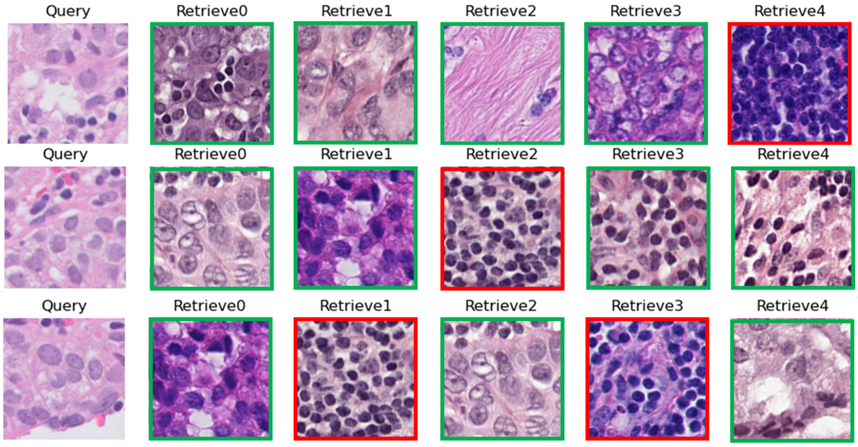
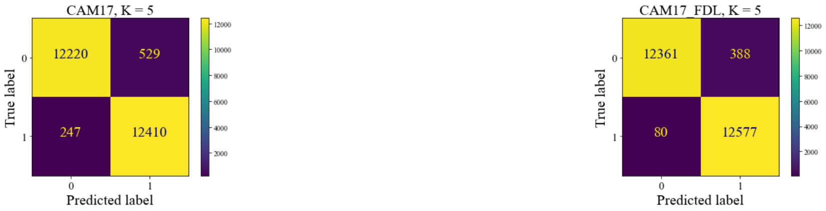


FIGURE 25: Three random queries from the hospital 5 of CAM17 (test set). Corresponding to each query, the top 5 images are shown from four other hospitals with the most similar patterns to the query. The green and red lines around the retrieved images explain the correct and wrong retrieved images.



(A) (B)

FIGURE 26: **(a)** shows the results of local training on CAM17 in the TY server. **(b)** is the result of the searching task in CAM17 by applying the well-train FedCBMIR model from the first experiment.

the retrieved images at each magnification, achieved by each client after training their models for 300 epochs within their server and without using FL. The highest accuracy of 95% for the retrieved images at 40× magnification was achieved by the client at UGR in 9.37 hours, while client 3 spent 8.59 hours to achieve a minimum accuracy of 89% and precision of 87%, which is the lowest among all the clients.

TABLE 13: Obtained results of CBMIR on 40×, 100×, 200×, and 400× at $K = 5$. We measure ACC, Precision, and F1S in the test set of each client at their corresponding magnification. The FedCBMIR was trained with the FedAvg strategy with 5 communication rounds in EXP 1. Time is reported in hour.

Client	Model	Training time	Accuracy	Precision	F1S
1	CBMIR	9.37 h	0.95	0.93	0.96
	FedCBMIR	6.82 h	**0.97**	**0.96**	**0.98**
2	CBMIR	5.45 h	0.90	0.88	0.94
	FedCBMIR	5.78 h	**0.94**	**0.92**	**0.96**
3	CBMIR	8.59 h	0.89	0.87	0.93
	FedCBMIR	6.65 h	**0.92**	**0.89**	**0.94**
4	CBMIR	8.95 h	0.92	0.89	0.94
	FedCBMIR	6.83 h	**0.96**	**0.94**	**0.97**

As demonstrated in Table 13, using the proposed approach, the four models were trained in a federated setting, which took $(Max(6.82h, 5.78h, 6.65h, 6.83h) = 6.83h)$ hours to complete the training process, is much faster than training one by one that took 32.36 hours in total, thereby reducing the total training time around 25.53 hours. This reduction in training time is particularly significant for large data sets and can facilitate more rapid and accurate diagnoses and treatments of cancers.

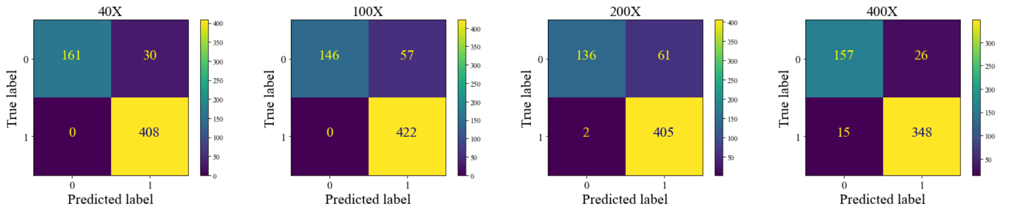
The performance evaluation of the proposed framework in the test set was compared with local training CBMIR and FedCBMIR, as shown in Figure 27. Each CM is associated with a specific magnification and reports the top 5 accuracy using ***Sen1*** in its search stage. The results of the ***Sen1*** in the test set are presented in Figure 27, where each client receives the top 5 images on average in 13.84 seconds. Table 14 presents a comprehensive comparison of various state-of-the-art CBMIR methods on the BreaKHis data set at 40× magnification. It is evident that FedCBMIR achieved the highest performance in both experiments conducted in this paper (EXP 1 and EXP 2). Notably, in EXP 1, where the model was trained by sharing weights with the CAM17 client, FedCBMIR exhibited superior performance. Previous studies by [63] and [131] utilized a hash method on BreaKHis 40× images and reported results in 16-bit, 32-bit, and 64-bit formats. Since the best-reported performance was achieved with 64-bit, we compare our results solely with this format, excluding 16-bit and 32-bit comparisons.

Quantifying mitosis count is a crucial criterion in breast cancer diagnosis [132]. The availability of advanced technology, such as high-resolution scanners, is not always guaranteed in every part of the world. Figure 28 demonstrates that as the magnification level increases, a smaller area of the tissue is displayed, and more relevant information becomes visible.

Figure 29 shows a comparison between the proposed method and its obtained results under EXP 1 and EXP 2, *Sen1* condition. The methods mentioned in the figure were applied to a binary breast tissue microscopic image data set built in [55] and [63]. In this paper [131], 20 retrieved images were taken into consideration in evaluating their proposed method. In Figure 29, since the authors in [131] did not name their two methods, we named them Method1 and Method2, then compared their results with the top 20 retrieved images with our results at the top 5 images. As can be understood from the bar charts 29, FedCBMIR in both experiments could overpass the other methods and they have higher precision in the retrieval performance. It is important to mention that Method1 and Method2 in [131] are supervised and reached 87% and 89.5% precision by proposing a supervised hashing method with multiple features. While FedCBMIR with an unsupervised FE obtained precision equal to 97% and 96% for both EXP 1 and EXP 2, *Sen1*.

TABLE 14: A comparison is presented between the Average Precision values of state-of-the-art papers and the results obtained in the reported experiments of this paper (EXP 1 and EXP 2, *Sen1*.)

Methods	**CBMIR, EXP 1**	**FedCBMIR, EXP 1**	**FedCBMIR, EXP 2, *Sen1***	**Method[133]**	**MCCH [69]**	**KSH, 64bits [63]**	**JKSH, 64bits [131]**
Precision	0.93	**0.97**	0.96	0.95	0.94	0.91	0.87



(a). 40× **(b). 100×** **(c). 200×** **(d). 400×**

The proposed ***Sen2*** approach can serve as an important tool for pathologists in developing nations to overcome the limitations of their scanners by enabling them to access tissue images at higher magnifications. FedCBMIR can facilitate cross-border collaborations, where pathologists from different regions can share their knowledge and expertise by analyzing similar patches at higher magnifications. In contrast to CBMIR, FedCBMIR in ***Sen2*** allows pathologists to retrieve similar cases at all four magnifications, not just from the same magnification as their query (Q). However, sharing images with a single server is not feasible due to storage and privacy concerns. To address this issue, the proposed FedCBMIR can retrieve similar patches at the same and higher magnifications.

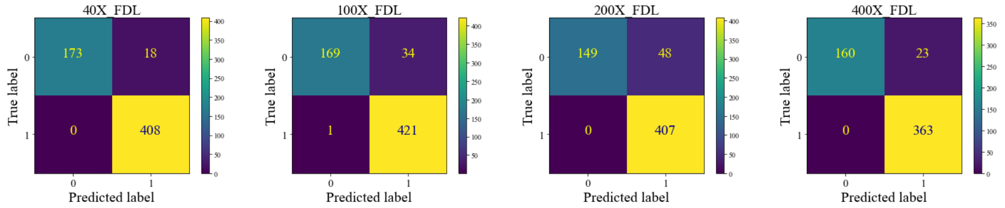
Table 15 proves that the proposed FedCBMIR is highly robust to receive a query at a specific magnification and retrieve the top 5 similar patches at all four magnifications. Each client fed the test set at the corresponding magnification and received the top 5 retrieved patches at all four magnifications.

TABLE 15: The ACC, Precision, and F1S for the second scenario of the EXP 2, *Sen2* with $K = 5$.

Client	**Accuracy**	**Precision**	**F1S**
1	0.94	0.92	0.95
2	0.95	0.93	0.96
3	0.95	0.93	0.96
4	0.95	0.92	0.96

The results of feeding the model with five random queries at 40× magnification by following (***Sen2***) are presented in Figure 30. The 40× is selected because it is the lowest magnification in the data set, and it is easier to measure the number of mitoses in images with higher magnifications. By feeding the model with images at 40×, pathologists can receive top 5 similar images at 40×, 100×, 200×, and 400×, which can significantly reduce the time and effort required to obtain a second opinion. The proposed approach has the potential to improve the speed and accuracy of cancer diagnosis and treatment. As such, it can serve as a user-friendly platform for pathologists to address their concerns more. Furthermore, it has the potential to be a valuable tool for telepathology in the future.

One of the challenges in collecting WSIs for use in DL models is the variability in color distribution due to differences in the staining material used across different hospitals and over time [134]. This variability can have a significant impact on the accuracy and reliability of DL models. However, an important finding from the results shown in Figure 30 is that the proposed approach, ***Sen2***, is not affected by differences in color distribution resulting from the staining process at different hospitals. This is a noteworthy result, as it indicates that the proposed approach can effectively overcome one of the major challenges associated with collecting and



(e). 40× **(f). 100×** **(g). 200×** **(h). 400×**

FIGURE 27: **(a)-(d)** show the CMs as a result of local training and searching at the same magnification. **e-h** are the CMs of FL models. The reported results are with top *K* retrieved images. In all CMs, "0" and "1" indicate "**Benign** and "**Malignant**", respectively. "**True labels**" and "**Predicted labels**" correspond to the query and the retrieved labels, accordingly.

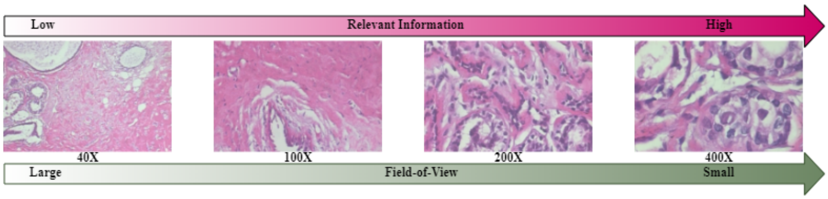


FIGURE 28: BreaKHis images at four different magnification levels (40×, 100×, 200×, and 400×). The higher magnification offers increased access to relevant information with a reduced field of view.

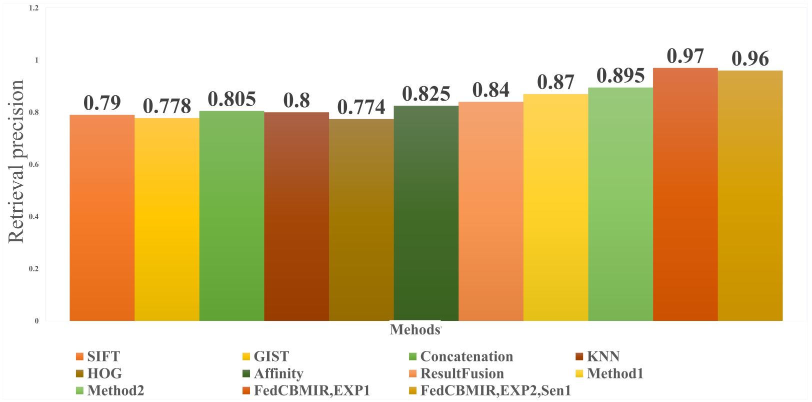


FIGURE 29: An indirect comparison between the results of FedCBMIR in both experiments and some recent methods at different amounts of *K*.

utilizing WSIs in telepathology. By eliminating the impact of color distribution variability, ***Sen2*** provides a more robust and reliable platform for pathologists to obtain accurate and consistent diagnoses, regardless of the specific staining materials used at different centers.

The proposed approach can contribute significantly to improving the accuracy and speed of disease diagnosis, particularly in regions where access to advanced technology is limited. In this way, ***Sen2*** has the potential to bridge the gap in healthcare and provide a more equitable and accessible healthcare system for all.

All the experimental results in both experiments and scenarios have verified that the proposed FedCBMIR has covered both concerns of DL scientists and pathologists with a fast-trained and accurate CBMIR, which is more generalized.

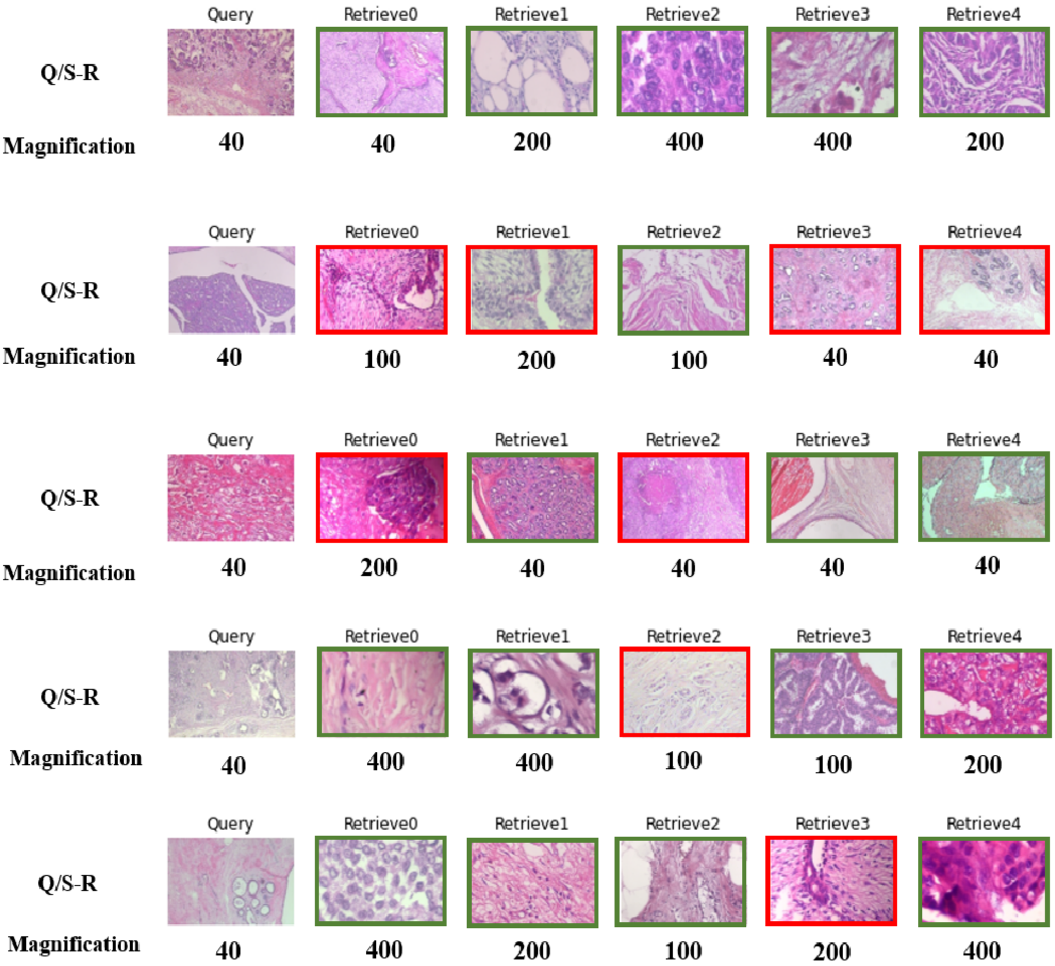


FIGURE 30: Five lines of random histopathological WSIs with their magnifications. The first column is the query, and the following five columns show the retrieved images. This figure brings a proper overview on ***Sen2***. The retrieved image with the same and different labels as the query is indicated by the green and red borders, accordingly.

4.5 Conclusion

The present study proposes a FedCBMIR approach that addresses two significant challenges in digital pathology faced by pathologists and engineers. By retrieving the top 5 similar images in a short amount of time, the proposed method reduces the workload of pathologists and decreases the time and cost associated with developing a high-performing DL-based method.

To evaluate the proposed approach, two experiments (EXP 1 and EXP 2) were conducted while EXP 2 contains two scenarios. EXP 1 aimed to provide a generalized model with Camelyon17 (CAM17) and BreaKHis 40× for clients that do not have enough images to train a model effectively. FedCBMIR in EXP 1 provides precision of 97.0% and 96.9% with training an unsupervised feature extractor within 11.44 hours faster.

EXP 2 comprised two scenarios: *Sen1*, where image institutions are not in agreement for sharing images, and *Sen2*, where images are delivered in different magnifications for institutions that lack the equipment to scan tissues at higher magnifications. The proposed method reached 98%, 96%, 94%, and 97% F1S for each client in *Sen1*. In *Sen2*, the BreaKHis data set was distributed across four institutions, resulting in accuracy rates of 97%, 94%, 92%, and 96% for pathologists at magnifications of 40×, 100×, 200×, and 400×, respectively. The average retrieval time was 13.84 seconds, and the well-trained models required 25.53 fewer hours to train four generalized models.

On one hand, WWCBMIR provides a chance to have a more accurate diagnosis for less-developed countries.

On the other hand, FedCBMIR can be a valuable tool for new graduate pathologists in their training and professional practice, offering benefits such as improved education, decision-making, research, and time efficiency.

Overall, this work offers a promising tool for hospitals to enhance diagnostic accuracy, medical education, and reduce the workload of pathologists by decreasing training time and increasing accuracy in compared to CBMIR methods. FedCBMIR aids in recognizing rare cases by connecting hospitals from the whole of the world. Although FedCBMIR tackles the challenges of data privacy, limited clinical context, and algorithm accuracy, the ongoing issues, such as dependence on image quality and security concerns, are still challenging for both hospitals and AI experts. Therefore, both hospitals and engineers must weigh the advantages and drawbacks while considering WWFedCBMIR as a tool.

4.6 Future work

To further enhance the performance of FedCBMIR for breast cancer diagnosis, it may be worthwhile to explore the use of additional data sets. This could include larger data sets with a greater number of labeled images, as well as data sets that encompass a wider range of malignancy levels and tumor subtypes. The incorporation of these data sets into the FL process, it may be possible to improve the accuracy and robustness of a CBMIR.

In addition to expanding the data sets used in Federated CBMIR, it may also be valuable to incorporate other types of clinical data into the system. Patient demographic information and clinical history could provide additional context and help to further refine the diagnostic process. Exploring the integration of these types of data could be a promising avenue for future research.

CRediT authorship contribution statement

Zahra Tabatabaei: Conceptualization, Methodology, Analyzing, Writing, Reviewing & editing, Formal analysis, visualization.
Yuandou Wang: Conceptualization, Methodology, Investigation, Formal analysis, Writing - Original Draft in sections related to federated learning, Review.
Adrián Colomer: Supervision
Javier Oliver: Review
Zhiming Zhao: Review
Valery Naranjo: Supervision

Acknowledgment

This study is funded by European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 860627 (CLARIFY Project). The work of Adrián Colomer has been supported by the ValgrAI – Valencian Graduate School and Research Network for Artificial Intelligence & Generalitat Valenciana and Universitat Politècnica de València (PAID-PD-22).

We would like to express our gratitude to all who have contributed to the completion of this research paper. We thank Laëtitia Launet for her valuable assistance in configuring the GPU at the Polytechnic University of Valencia (UPV). We also acknowledge the University of Granada (UGR), especially Prof. Rafael Molina and Prof. Javier Mateos, for granting us access to their high-performance computing resources, including the NVIDIA GeForce RTX 3090 GPU, which greatly accelerated our deep learning computations and enabled us to achieve state-of-the-art results in our research.

Chapter 5

Siamese Content-based Search Engine

This Chapter corresponds to the author's version of the following published paper: Tabatabaei, Zahra, et al. "Siamese Content-based Search Engine for a More Transparent Skin and Breast Cancer Diagnosis through Histological Imaging." Computers in Biology and Medicine journal, under review, 2024.

Siamese Content-based Search Engine for a More Transparent Skin and Breast Cancer Diagnosis through Histological Imaging

Zahra Tabatabaei [1,2†,*], Adrián Colomer [2], Javier Oliver Moll [1], and Valery Naranjo [2]

[1] Dept. of Artificial Intelligence, Tyris Tech S.L., Valencia, Spain.
[2] Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano, HUMAN-tech, Universitat Politècnica de València, Spain.

Abstract

Computer Aid Diagnosis (CAD) has developed digital pathology with Deep Learning (DL)-based tools to assist pathologists in decision-making. Content-Based Histopathological Image Retrieval (CBHIR) is a novel tool to seek highly correlated patches in terms of similarity in histopathological features. In this work, we proposed two CBHIR approaches on breast (Breast-twins) and skin cancer (Skin-twins) data sets for robust and accurate patch-level retrieval, integrating a custom-built Siamese network as a feature extractor. The proposed Siamese network is able to generalize for unseen images by focusing on the similar histopathological features of the input pairs. The proposed CBHIR approaches are evaluated on the Breast (public) and Skin (private) data sets with top K accuracy. Finding the optimum amount of K is challenging, but also, as much as K increases, the dissimilarity between the query and the returned images increases which might mislead the pathologists. To the best of the author's belief, this paper is tackling this issue for the first time on histopathological images by evaluating the top first retrieved images. The Breast-twins model achieves 70% of the F1score at the top first, which exceeds the other state-of-the-art methods at a higher amount of K such as 5 and 400. Skin-twins overpasses the recently proposed Convolutional Auto Encoder (CAE) by 67%, increasing the precision. Besides, the Skin-twins model tackles the challenges of Spitzoid Tumors of Uncertain Malignant Potential (STUMP) to assist pathologists with retrieving top K images and their corresponding labels. So, this approach can offer a more explainable CAD tool to pathologists in terms of transparency, trustworthiness, or reliability among other characteristics.

5.1 Introduction

Skin cancer is one out of three diagnosed cancers worldwide, according to the World Health Organization (WHO) [23]. Basal and squamous cell carcinoma are the two most frequently occurring types of skin cancer, while the most perilous type is malignant [135]. Classifying and distinguishing between benign (melanocytic nevus) and malignant (melanoma) can be reliably feasible in common melanocytic tumors [136]. One of the diagnostic challenges for pathologists in spitzoid tumors is Spitzoid Tumors of Uncertain Malignant Potential (STUMP) because their prognostic implications are unknown [137]. In this case, in order to make an accurate cancer diagnosis, pathologists often need to transfer the glass of biopsy to other centers to consult with their peers about the grade of the tissue. This workflow is time-consuming and expensive, and some accidents might occur with the glass of the biopsy [27].

Another world's most prevalent cancer type according to WHO is breast cancer with 685, 000 deaths globally in 2020 [23]. Breast cancer is caused by abnormal breast cells that grow out of control and form tumors. Cancer cells can spread to nearby lymph nodes or other organs. Treatment for breast cancer depends on the type and sub-type of cancer and how it has spread outside of the breast [138]. There are some treatments for breast cancer including, surgery to remove the breast tumor, radiation therapy to reduce recurrence risk in the breast and surrounding tissues, and medications to kill cancer cells and prevent spread, including hormonal therapies, chemotherapy, or targeted biological therapies. Doctors might combine some treatments to make sure that the possibility of the cancer coming back is minimal. But the over-diagnosis or over-treatment, and variability in interpretation are some key challenges in breast cancer diagnosis [139].

Computer Aid Diagnosis (CAD) offers some efficient Deep Learning (DL) tools that can assist pathologists and address human errors in cancer diagnosis and treatment [140]. The advance of DL methods opened the door to reducing the workload for pathologists and enhanced patient care [141]. Digitization of the tissue slides as high-resolution Whole Slide Images (WSIs), emerges computer vision tools as a support for pathologists in their daily tasks. Some of these tools, such as classification [142], segmentation [78], Content-Based Histopathological Image Retrieval (CBHIR) [143], etc., provide a second opinion for pathologists to have a more accurate cancer diagnosis [28]. Image-based tools of CAD are classified into medical image classification and Content-Based Medical Image Retrieval (CBMIR) [144]. The main objective of the classifiers is to categorize the images, while CBMIR aims to rank and show similar images in addition to their labels [145]. CBMIR, particularly CBHIR tools, demand representative features to retrieve similar histological patches with the same histopathological patterns effectively. To reach the most meaningful features of the patches, a proper Feature Extractor (FE) is needed.

Finding an optimum FE for the goal under study is challenging. There are plenty of FE, including SIFT [146], color histogram-based features [147], Gabor features [148], and GIST features [56] as some hand-craft features. In the other works [149, 131, 150, 151], DL-based FEs, including CNN, pre-trained models, Auto Encoder (AE), etc., were applied to extract the deep features. However, most of DL-based techniques face some challenges, mostly related to the digitized histopathological images. For instance, the annotated data set is one of the critical requirements for developing a high-performance DL- model. However, annotating histopathological images is time-consuming and costly, hindering the performance of DL-based tools. To cope with this challenge, some of the recent research on WSIs has focused on self/unsupervised techniques, which need fewer annotated images [152].

As a means to tackle this lack of enough annotated images, Zahra et al. proposed an unsupervised ResCAE to perform a search engine in prostate cancer. This paper reported 85%, 78% of accuracy at top 7 and 5, respectively [24]. Histopathology Siamese Deep Hashing (HSDH) is proposed in [153] for histopathology retrieval and achieved 97%, 98%, and 99% Mean Average Precision (MAP) for 32, 64, and 128 bits. The authors reported the results at the top 100, 150,..., and 400 for BreaKHis as a binary data set. RetCCL in [154] is a clustering-guided contrastive learning approach for WSI retrieval. RetCCL has improved 24% at the patch-level retrieval on the TissueNet data set in terms of mMV@5 in comparison with ImageNet pre-trained features. Authors in [88] proposed a CBHIR framework for a data set containing WSIs and size-scalable query Regions Of Interest (ROI). The reported retrieval results at the top 20 returned images, reached 96% of precision on the Motic database. SMILY [70] was proposed by Google AI Healthcare and it used an automatic high-level feature extraction to provide the feature vectors for the search engine. SMILY retrieved images with the top 5 similar patches of prostate cancer with 73% accuracy.

In the recent papers on CBHIR [27, 24, 153, 88, 70, 25, 13], the performance of the model was reported by top K accuracy. This evaluation technique assumes that if just one of the retrieved images out of K is correct, the model performed correctly [154]. Choosing the correct amount of K in this evaluation technique is challenging. In some papers such as [153], the amount of K was chosen in $[100-400]$, which is a high amount, especially for a binary data set. In [88], the reported accuracy is at the top 20 images. In other papers [155, 156, 157, 158], mostly the amount of K was chosen as $5, 7, 10$. Retrieving top K images brings some benefits for pathologists; for instance, they can analyze more cases to see similar histopathological features. The issue with top K accuracy in evaluating the performance of CBHIR is that, in some cases, the model could retrieve one correct similar patch out of 5, 7, or even 400 retrieved images. Although this can ensure pathologists that at least one image out of K is correct, it cannot provide a highly accurate second opinion for them. However, in the top K accuracy technique, even if the k_{th} retrieved out of K, was the only similar image to the query, the accuracy of the model is high. This issue is highlighted more in the papers with a high amount of K.

In a departure from conventional methods, this paper proposes a CBHIR by applying a custom-built Siamese network as an FE on skin and breast cancer data sets, which are the prevalent cancer types. Since these types of data sets can exhibit substantial variability, the Siamese network is a promising option for histopathological images. This network is adapted to capture the intricate histopathological patterns in the patches that are crucial for grading and diagnosis. The Siamese network is specifically designed to excel in similarity-based tasks and this feature is particularly well-suited for CBHIR applications. The proposed Siamese network employs a contrastive loss function to emphasize the relationships between data points in the embedding space. This behavior can enhance the ability of the proposed Siamese network to identify similarities between images and improve the distinction between dissimilar ones. The extracted features by the proposed Siamese network is highly transferable and they can be employed effectively in the search engine of CBHIR without extensive retraining. This transferability is valuable when dealing with histopathological images.

To the best of the author's knowledge, no previous studies were conducted based on the Siamese network in a CBHIR model to assist pathologists in making a more accurate diagnosis of spitzoid cancer. Also, this is the first time that a CBHIR model has been dedicated to grading the STUMP cases in a spitzoid melanocytic lesions database. Furthermore, in this paper, the proposed CBHIR model can have high performance with the top first retrieved patches, which is the first paper that could reach this power at the top first, to the author's best belief. The performance of the proposed CBHIR model confirms that the model is generalized well to unseen data.

In summary, this paper makes the following main contributions:

1. We propose two Siamese networks for breast cancer (public data set) and skin cancer (private data set) to show the generalization of the proposed CBHIR technique. These proposed Siamese networks are robust to imbalanced data sets;

2. Siamese network is employed to address the shortcomings of histopathological images, including small inter-class variations and large intra-class variances;

3. The CBHIR results on both data sets were reported at the top first retrieved images to demonstrate the model's efficacy in retrieving relevant patches;

4. The proposed CBHIR approach provides a second opinion to pathologists to tackle the challenges in grading STUMP by providing deep insights into the complexities. We show the Gradient-weighted Class Activation Mapping (Grad-CAM) figures as explainable Skin-twins to provide interpretability to the uncifrable STUMP cases;

5. The performance of the proposed CBHIR technique in retrieving the images with the same cancer type was evaluated in comparison with some state-of-the-art classifiers;

6. Based on the experimental results on two histopathological data sets, the proposed CBHIR framework, thanks to the Siamese network, outperforms other image retrieval methods.

5.2 Methodology

Figure 31 illustrates the proposed CBHIR system in a general overview. First, medical centers have to scan the biopsies with WSI scanners [159]. Second, the Siamese network needs to train to extract the features of the images. Third, the extracted features of the database have to be indexed and saved. Later, a search process needs to find similar patches to the query among the images in the database. Finally, it is time to visualize the top similar patches to pathologists for further analysis. This is how the CBHIR system can make a bridge between the current cancer diagnosis and digital pathology.

5.2.1 Siamese network

The Siamese network is characterized by its twin neural network structure. Siamese networks are designed to compare the similarity between pairs of inputs. This network can be effective even with limited images in the train set, and it learns the representative features by focusing on the relative relationships between the pairs of input. Siamese networks bring advantages to the CBHIR approach, such as robustness to class imbalance and learning from semantic similarity. Figure 32 illustrates the pros and cons of applying the Siamese network to histopathological images. Few-shot learning in the Siamese network empowers the model to be able to recognize patterns and measure the similarity between patches with very limited labeled data. Besides all the benefits of this network, it is important to mention that

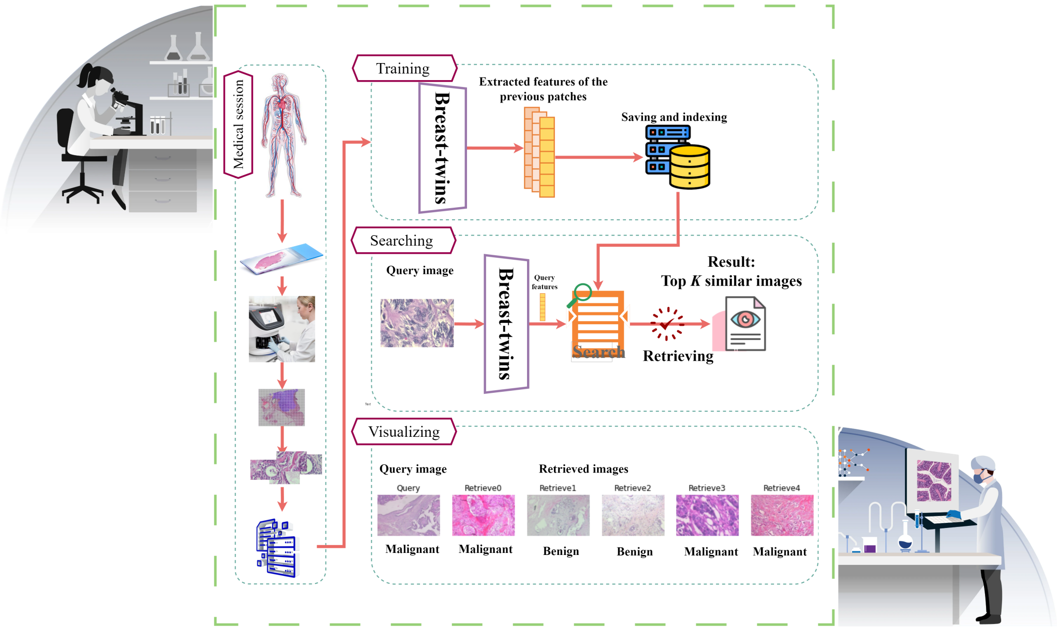


FIGURE 31: A general overview of the proposed CBHIR to show how the proposed CBHIR can bridge traditional cancer diagnosis workflow to digital pathology.

its performance might be sensitive to the hyperparameter settings and it requires extensive hyperparameter tuning. The Siamese network can be complex to design and train compared to the traditional image retrieval techniques.

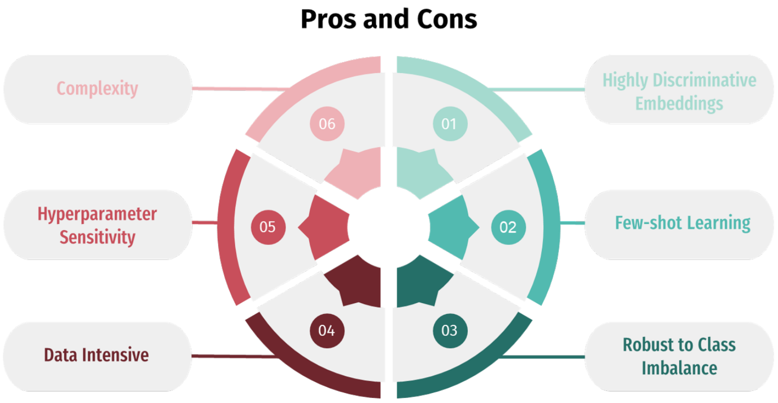

Pros and Cons
Complexity
06
01
Highly Discriminative Embeddings
Hyperparameter Sensitivity
05
02
Few-shot Learning
Data Intensive
04
03
Robust to Class Imbalance

FIGURE 32: Advantages and challenges of implementing the Siamese network in a CBHIR. Points 1 to 3 mention the difficulties in using Siamese. Points 4 to 6 are about the positive points of using the Siamese network.

Figure 33[1] shows an overview of the proposed Siamese network on the BreaKHis data set. As can be seen, there are two identical networks, which are named sister networks. Each of these sisters contains the same architecture comprising a deep convolutional network, an encoding layer, and a distance metric. The main objective

[1] The figure is plotted by breast histopathological images. The same pipeline was applied to the skin data set with a small difference in the internal structure of the Siamese network. The structure of these convolutional networks is slightly different for skin and breast data sets regarding filter size due to the difference between the size of images in each data set. More details are mentioned in section 5.3.2

of these sisters is to minimize the contrastive loss in order to minimize the distance between images with the same histological features.

The Siamese network receives pairs of images as input. Each sister network takes one instance from the input pair and passes it through the convolutional layers of the proposed network for feature extraction. While training, these sisters shared the weights regarding encoding the input images to ensure that both networks learn to extract similar features of the input image. During training, in order to measure if each pair is similar or dissimilar, the label of the data is needed. Once the Siamese network learns to extract the representative features for pairs of input, it can be used to measure the similarity between pairs by computing the distance between the extracted features generated by the sister network for each patch instance.

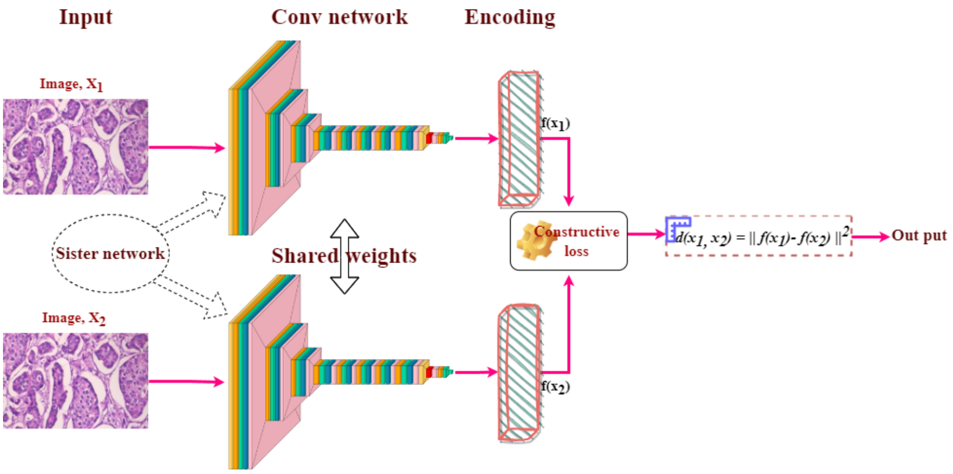


FIGURE 33: An overview of the proposed network. This is plotted with BreaKHis images. Colors of orange, red, pink, green, and teal represent Convolutional layers, MaxPooling2D, Dropout, and Flatten, respectively.

5.2.2 Contrastive loss

The benefit of contrastive loss compared with the other loss functions is enabling the network to converge faster and reducing overfitting. This occurs since the network learns to be more generalized by focusing on the key features and determining the similarity.

The contrastive loss takes pairs of samples as anchor and neighbor or anchor and distant [160]. In the case of anchor and neighbor, they are pulled towards each other. In the other case, the distance between the anchor and the distant needs to be increased. The optimum embedding feature space is extracted while the anchor-distant distances get larger than the anchor-neighbor distances by a margin of m. This can enhance the embedding space and improve the discriminative capacity. The utilized contrastive loss defined for this work is as follows:

$$L_c = \sum_{i=1}^{b} [(1-y)||f(x_1^i) - f(x_2^i)||_2^2 + y[-||f(x_1^i) - f(x_2^i)||_2^2 + m]_+] \qquad (5.1)$$

The contrastive loss (L_c) (1) should be minimized, which helps the network focus on capturing invariant features [161]. y is zero or one when the pair $\{x_1^i, x_2^i\}$ is anchor-neighbor and anchor-distant, respectively. We denote the m as the margin and $f(x)$ as the output of the network (i.e., embedding). Let b be the mini-batch size, and $||.||_2$ the L_2 norm [162]. Figure 34 illustrates the workflow of the contrastive loss by receiving anchor-neighbor and anchor-distant. Three points in the colors red, green, and blue correspond to distant, neighbor, and anchor samples. Initially, the green point is located far from the anchor sample. As the model undergoes training, it acquires the ability to discriminate between samples. Ultimately, the contrastive loss function propels the red point to a distant location while maintaining the green point close to the anchor sample.

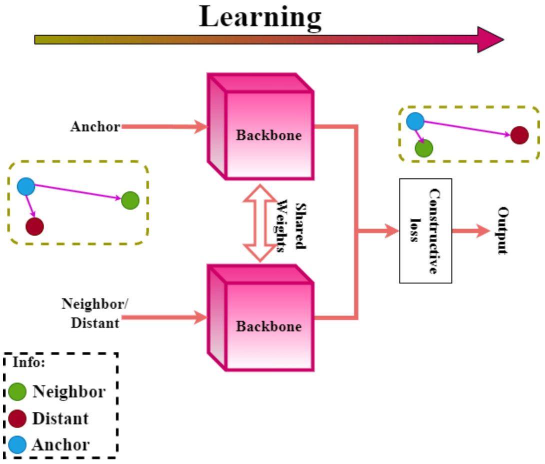


FIGURE 34: An overview of how contrastive loss receives the anchor-distant and the anchor-neighbor. The blue, red, and green points are anchor, distant, and neighbor samples.

5.2.3 Search engine

The trained Siamese network provides an FE that can extract the meaningful features of the data set. By reaching the well-trained FEs, it is time to employ them in a CBHIR framework. Figure 35[2] shows an overview of the proposed CBHIR. The trained Siamese network is used as an FE to extract two meaningful features of each input image. Each image i in the data set is saved with a feature vector F_i containing the two most representative features of that. All these feature vectors are saved in the feature storage.

When the CBHIR model receives a new query, the FE extracts the most meaningful features of the query (F_Q). As the next step, a distance function is needed to measure the distance of F_Q with all saved F_is in the feature storage (F_D). Among different types of distance functions, Euclidean distance was chosen for this paper as it is one of the most used metrics in order to calculate the distances between two images in CBHIR [163]. Euclidean distance is straightforward to understand geometrically and is computationally efficient. The other reason that the Euclidean function was selected is that it works well in high-dimensional vector spaces.

In the searching part of the retrieval step, the images are ranked with the shortest distance at the top, based on the Euclidean distance. This means the images with the most similarities are the top images. Then, top K retrieved images are returned to pathologists for further investigations. So, it is expected to have the most similar images at the first top retrieved and the least similar image at the K_{th} $(K > 1)$

[2] The figure shows the histopathological breast images. The same pipeline was applied for the skin images.

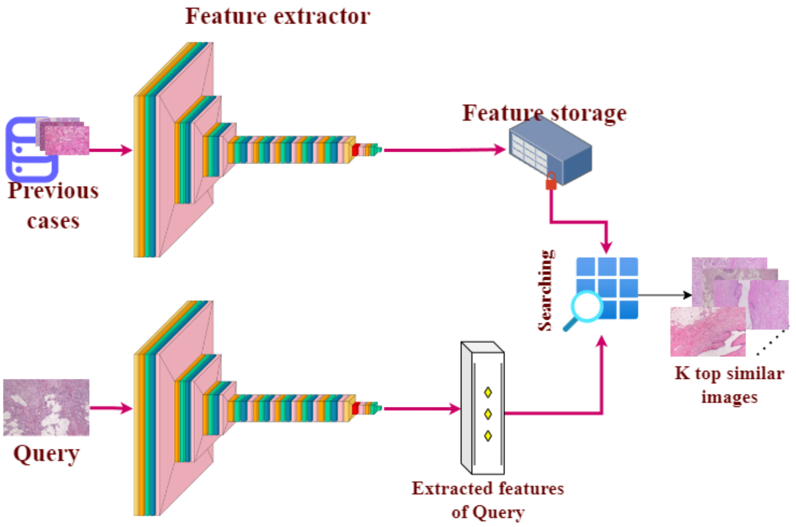


FIGURE 35: An overview of the proposed CBHIR by using the well-trained Siamese network as the FE.

retrieved. As much as the amount of K is less, the number of images that are taken under evaluation is less. So, the model that reaches higher accuracy in less number of retrieving images is a more reliable tool. In this paper top, 1, 3, and 5 retrieved images on both data sets are shown.

5.3 Material and Experimental setup

5.3.1 Material

Two binary data sets of breast cancer and skin cancer were taken under study in this paper. Both data sets were stained with Hematoxylin and Eosin (H & E) [164].

Breast cancer data set

BreaKHis data set contains 7909 microscopic images of breast tumors at four magnification levels. This paper applied the experiments to images at 400× magnification of this data set in the size of $224 \times 224 \times 3$ [25]. In grading breast cancer, the number of mitoses is a critical criterion. In order to measure it, 400× magnification was selected, which is the highest level of magnification of BreaKHis. This binary data set comprises 1232 malignant and 588 benign at 400× magnification [27]. In the training strategy, 70% of the data set is randomly chosen for training, while the rest is considered for testing.

Spitzoid melanocytic data set

Spitzoid melanocytic is one of the most challenging skin tumors for pathologists because of its ambiguous histological features [165]. The data set was collected from 79 different patients containing 84 biopsies. Table 16 describes the spitzoid melanocytic lesions database [166]. The data set is collected and annotated by the pathology laboratory from the University Clinic Hospital of Valencia. The data set contains uncommon skin tumors and 42 STUMP images.

Following [166], the images were segmented from ROIs. The segmented images are non-overlapped, in the size of $512 \times 512 \times 3$. Following [141], in our experiments,

TABLE 16: Description of the Skin data set.

Data type	Malignant	Benign
WSIs	34	50
Patients	30	49
Annotated	16	11
Unannotated	29	41

in order to prevent data leakage, the data set was later split into 70% training and 30% testing. In this paper, in the training, validation, and testing steps, the STUMP images were excluded. Later, in Section 5.4.3. they are considered a second test set, and the trained CBHIR is applied to them to retrieve the most similar patches to them. In this experiment, we can elucidate if the STUMP query provides more malignant or benign retrievals. The idea is to use the CBHIR as a transparent diagnostic tool for this kind of cases which often pose significant difficulties for pathologists.

5.3.2 Experimental setting

The proposed convolutional network in this sister network comprises two stages of convolutional layers. More details about the Siamese structure and the hyperparameters for the breast (Breast-twins) and skin (Skin-twins) data set are mentioned in the following subsections and Figure 36.

Tuning hyperparameters in Siamese networks is challenging for different purposes, and it relates to the type of the data set. So, it is important to note that Breast-twins and Skin-twins models were trained with a learning rate schedule from Keras with an initial learning rate of 0.01, decay steps of 10000, and decay rate of 0.9. Stochastic Gradient Descent (SGD) optimized the contrastive loss while training. For both data sets, many trials and errors found the amount of margin (m), and 0.9 was found as the optimum amount for this task.

First, the network is initialized with a contrastive loss and SGD as the optimizer. Then, the first image of the image pairs is processed through the network. Subsequently, the second image of the image pair is fed into the network. Then, the loss is determined by the outputs from the first and the second images. To optimize the model, the gradient is computed, and the model weights are updated using the Stochastic Gradient Descent (SGD) algorithm.

The experiments in this paper were run on GPU with the *NVIDIA GeForce RTX 3090*.

Breast-twins in details

Breast-twins start with four convolutional layers ($[32, 64, 128, 256]$) with the filter size of 3 by 3 while stride is $[1, 2, 2, 2]$. In order to extract deeper features, the network goes deeper with more convolutional layers with the stride of 1 and a kernel size of 3 by 3. The number of convolutional filters in each layer of the residual block is $[64, 32, 1, 256]$ to reduce the spatial dimensions while increasing the number of filters and compressing the information. In this case, decreasing the filter size helped control the number of parameters while still capturing relevant features. Ultimately, a convolutional layer and a Global Max Pooling2D (GMP) are applied to reduce the spatial dimensions and retain important features.

In this experiment, the Breast-twins network was trained over 300 epochs with a batch size of 16.

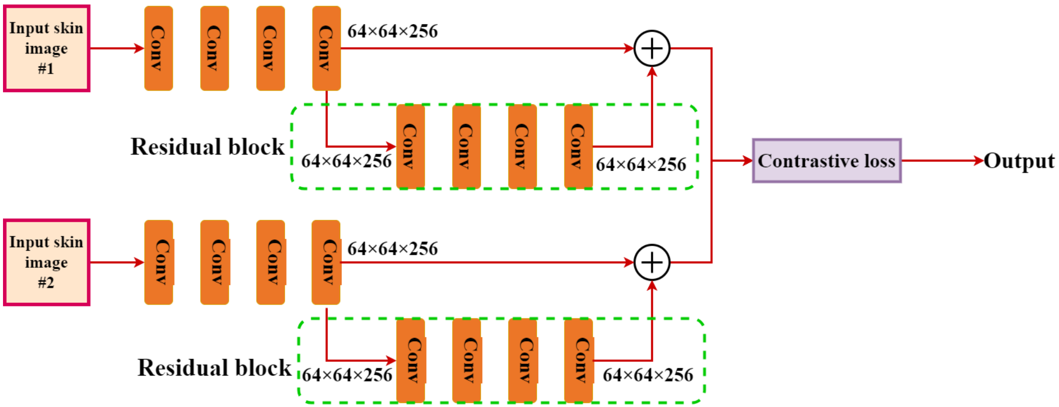


FIGURE 36: An overview of the proposed method on skin images in perspective of convolutional layers. Each of the Convolutional layers in the figure is named *Conv*.

Skin-twins in details

In the case of the skin data set, since the size of the patches is bigger than BreaKHis, different filter sizes are needed. So, as can be seen in Figure 36, first, four convolutional layers with the filter size of 3 by 3 and the layers of $[32, 64, 128, 256]$ and stride of $[1, 2, 2, 2]$ compress the input. In this part of the network, we aim to gradually reduce the spatial dimensions by increasing the filter sizes. Then, in order to get more meaningful features of the input, we added a new series of convolutional layers of $[32, 16, 1, 256]$ with stride fixed to 1 and filter size of 3×3 in the residual block of the proposed architecture. This makes the network deep enough but not excessively complex. The Skin-twins network was trained over 300 epochs with a batch size of 16. Figure 36 provides an overview of the proposed structure from the convolutional point of view.

In the training part of the Skin-twins, the skin data set is considered a binary data set and excludes the STUMP class of the data set to use as the test set for the experiments in section 5.4.3.

5.4 Results and Discussion

This section reports the results of Breast-twins and Skin-twins with tables of the results and visual evaluations to provide a comprehensive comparison.

5.4.1 Results of Breast-twins

Table 17 provides a comparison between some state-of-the-art papers that reported the results with different amounts of K. In this table, the amount of K for each study is mentioned since to the best of the author's knowledge, there are not any papers reporting the results at the first top retrieval on a breast data set. As can be seen in Table 17, the Breast-twins network is able to correctly retrieve the first top image with a 59% of accuracy.

Paying attention to the results, it can be understood that with $K = 5$, Breast-twins outperforms MCCH [133]. Breast-twins models has only a 0.06 difference in terms of accuracy with HSDH [153], which sets the amount of $K = 400$, which is a high amount for a binary data set. On the other hand, although FedCBMIR in [27] was

TABLE 17: Comparison between the cutting-edge methods and Breast-twins model. This comparison is on BreaKHis at 400× magnification.

Method	K	Accuracy	Precision	F1score
Breast-twins	1	0.59	0.67	0.70
Breast-twins	3	0.83	0.81	0.88
Breast-twins	5	0.92	0.90	**0.94**
MCCH [133]	5	-	0.89	-
FedCBMIR [27]	5	0.96	**0.94**	-
HSDH [153]	400	**0.99**	-	-

trained on all four magnifications and then tested on BreaKHis at 400×, it has only 0.04 higher accuracy than the Breast-twins model which only trained on BreaKHis at 400× magnification. According to Table 17, the performance of Breast-twins at top 1, 3, and 5 shows its reliability.

Table 18 compares the performance metrics, including accuracy, recall, precision, and F1score, between the proposed Breast-twins model with the previously proposed CAE [25]. The results showcased in this table demonstrate that although the CAE could provide a comparable performance in terms of accuracy, precision, and F1score at the top 3, there is a notable disparity in its performance regarding the first top retrieval. This table can clarify the potential risks of a CBHIR technique with high accuracy only at top K while $K > 1$.

TABLE 18: Comparison with other state-of-the-art methods

Method	K	Accuracy	Recall	Precision	F1score
CAE	1	0.48	0.54	0.64	0.56
Breast-twins	1	**0.58**	**0.73**	**0.67**	**0.70**
CAE	3	0.83	0.95	**0.82**	0.88
Breast-twins	3	**0.83**	**0.97**	0.81	**0.88**

Finding a sufficient amount of K is challenging for DL experts. For instance, in the case of [153], the results were reported at the top 400, which delivered 99% accuracy for a binary data set. According to the definition of accuracy in CBHIR, increasing the amount of K yields higher accuracy. But a high amount of K, means retrieving images with longer distances and less similarity. Therefore, a CBHIR with high accuracy at a high amount of K cannot provide a confident tool. The impacts of the amount of K can be seen in Figure 37 which shows three Confusion Matrix (CM) at top 1, 3, and 5 for breast cancer.

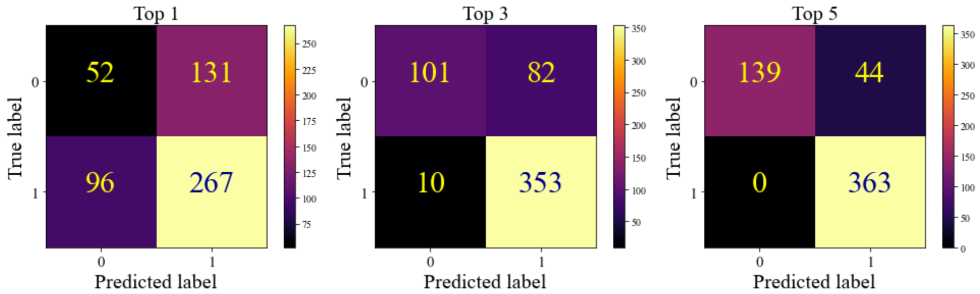
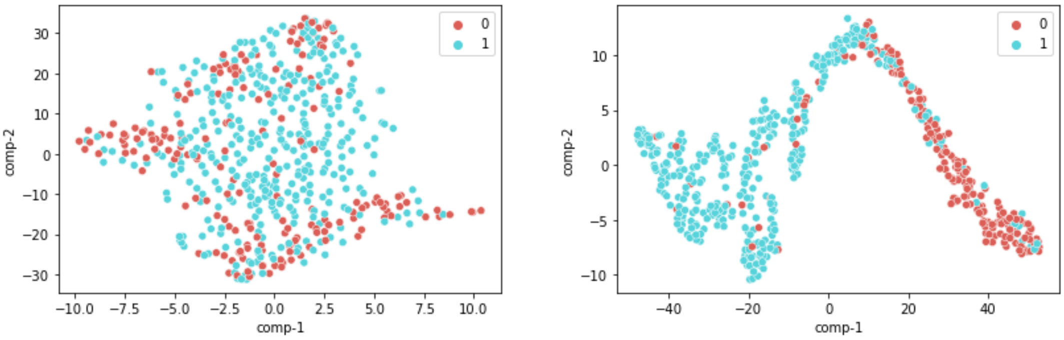


FIGURE 37: Three CMs at top 1, 3, and 5 for Breast-twins CBHIR model.

Figure 38 plots the T-distributed Stochastic Neighbor Embedding (TSNE) of the extracted features of the breast data set. TSNE is a data visualization technique that is used for reducing the dimensionality of high-dimensional data while preserving the similarities between data points. Figure 38 compares the extracted embedding space as a result of FedCAE [27] and the proposed Breast-twins. FedCAE was trained on all four magnifications of the BreaKHis data set and then applied to BreaKHis at 400× magnification. As can be seen in Figure 38, although FedCAE is more generalized due to the different magnifications, the embedding space resulting from Breast-twins is more discriminative.



(A) Embedding space extracted by FedCBMIR (B) Embedding space extracted by Breast-twins

FIGURE 38: TSNE plot of the extracted feature space by applying two different FE to the breast data set. **Zero** label in red shows the embedding space of benign tissues and the **One** labeled points in blue correspond to the malignant cases.

In breast cancer diagnosis, pathologists need to take into account the histopathological patterns of the tissue to be able to grade the tissue. Well-differentiated patterns, low nuclear pleomorphism, and mitotic activities are some of the important criteria that pathologists analyze in their query to write a diagnosis report. These grading patterns assist pathologists in determining the best treatment and predicting the likelihood of the tumor spreading. So, the CBHIR system answers these demands from pathologists by providing similar patterns for the pathologists.

Figure 39 shows eight random patches of BreaKHis data set with their Grad-CAM plots [167]. The presented Grad-CAM visualizations were generated by fusing the GMP layer of the proposed Breast-twins network to emphasize the final feature representation of the model.

These Grad-CAM plots provide insight into the DL-based Siamese networks to clarify why the model made a particular decision. This means that it can identify if the model focused on the correct histopathological features or retrieved the images based on the artifact. Figure 39 clarifies the important regions of the tissue that the Breast-twins network detected as an important part of the patches for the final decision of the search engine in the CBHIR model. This figure proves the interpretability of the proposed model.

Figure 40 visualizes the output of the proposed CBHIR by using Breast-twins for three random queries. The retrieved images were compared with the query based on their labels. Images with a different label than the query are marked in red. It is seen that the proposed Breast-twins network has the ability to return the patches with similar histopathological features with the correct class labels of these patches. Results indicate that the Breast-twins network is highly efficient in retrieving similar patches.

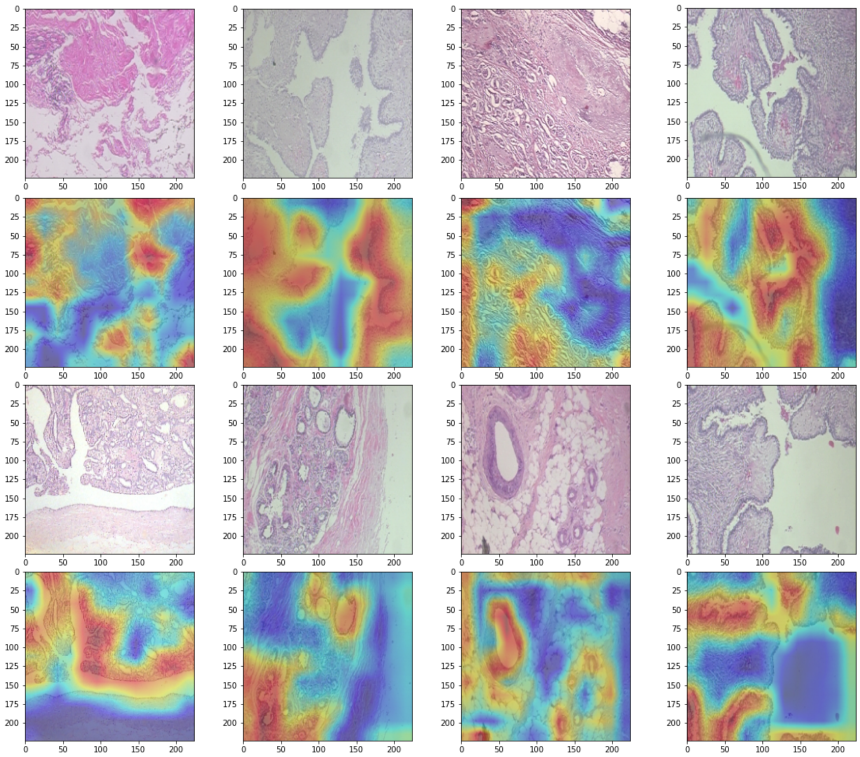


FIGURE 39: Eight random patches of breast cancer data set. The odd rows show the original images. The even rows are the Grad-CAMs mapped on top of the original images.

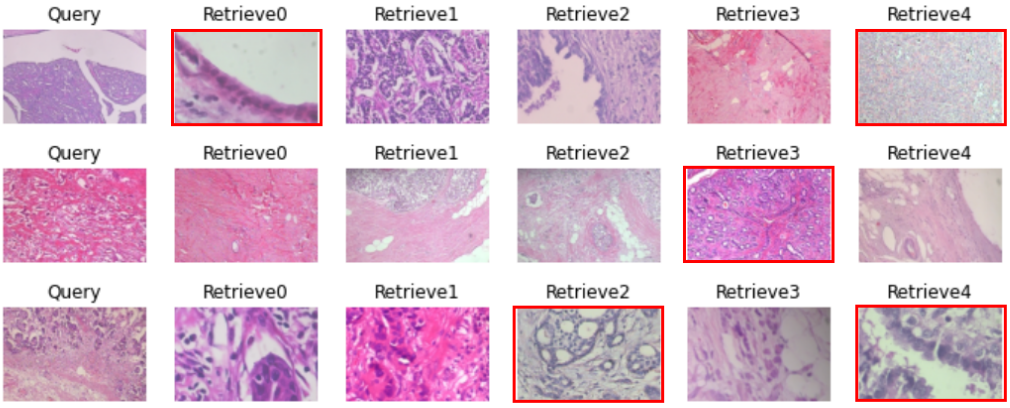


FIGURE 40: Three random breast queries with their 5 top retrieved images. The miss-retrieved images are marked in red.

5.4.2 Results of Skin-twins

Table 19 illustrates the performance of Skin-twins at top 3 and 5. Skin-twins can find similar patches to the query from the spitzoid cancer data set, accurately with 64% and 94% of F1score for the classes of *Benign* and *Malignant*, respectively.

TABLE 19: Results of Skin-twins in searching between "*Malignant*" and "*Benign*" at top 3 and 5. K = 3, 5)

Method	K	Precision	Recall	F1score		F1s-avg	Accuracy
Skin-twins	3	0.88	0.99	0.64	0.94	0.94	0.89
	5	0.92	1	0.80	0.96	0.96	0.93

Table 20 compares the proposed model with our previously proposed CBMIR model based on CAE [25] at the top first retrieved images. As can be understood from the reported results, Skin-twins framework reaches 67% higher precision at the top first retrieved images, which is high development in CBHIR systems.

TABLE 20: A comparison between the results of Skin-twins and state-of-the-art studies at the first top retrieval K = 1).

Method	Precision	Recall	F1score	Accuracy
CAE	0.13	0.14	0.135	**0.72**
Skin-twins	**0.80**	**0.82**	**0.81**	0.69

Figure 41 shows the CM of the proposed method on the skin data set at top K = 1, 3, and 5. In these CMs, *Benign* and *Malignant* are labeled as zero and one. The figure illustrates that by increasing the amount of K, the amounts of True positive and True negative are increasing. Figure 42 shows the TSNE plot of the extracted features of the skin data set. Each point in this plot corresponds to a data point, and the proximity of points indicates similarity in the underlying feature space. The points are color-coded to represent the benign and malignant classes in red and blue, respectively. This figure provides insight information of the inherent patterns and relationships of the embedding space.

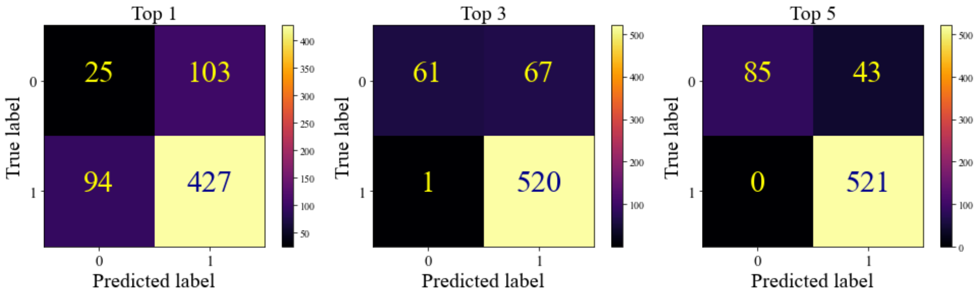


FIGURE 41: Three CMs at top 1, 3, and 5 for Skin-twins CBHIR model.

The Grad-CAM plots of the Skin-twins are shown in Figure 43. In skin cancer detection, Cellularity, Nuclear Features, Cellular Morphology, Architecture, Mitotic activity, Stroma and Connective Tissue, etc., are some of the histological features that pathologists consider to grade a skin cancer [168]. These are the histological patterns that the CBHIR should focus on to find similar patches. Figure 43 shows the success of the model in detecting the cancerous and abnormal regions of the tissue. For instance, in some cases, the main focus is cellularity and assessing the density of the cells within the tissue. High cellularity may indicate increased cell proliferation or

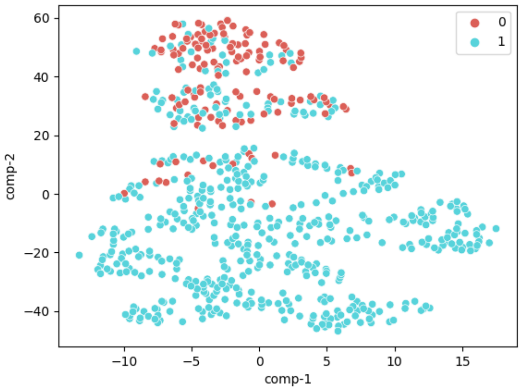


FIGURE 42: TSNE plot of the extracted features of skin data set by Skin-twins network. Features corresponding to benign tissues are plotted in red and labeled as **Zero**. The blue points are related to the malignant tissues which are labeled as **One**.

inflammation. So, the Grad-CAM plots generate a visually interpretable explanation for the retrieved images, which explains why the Skin-twins model retrieved the images for the pathologists.

Figure 44 depicts three query examples of the skin data set with their retrieved images. The red line around the retrieved patches clarifies that the retrieved image does not have the same label as the query. As can be seen in this figure, just one out of five retrieved patches was not at the same cancer grade as the query.

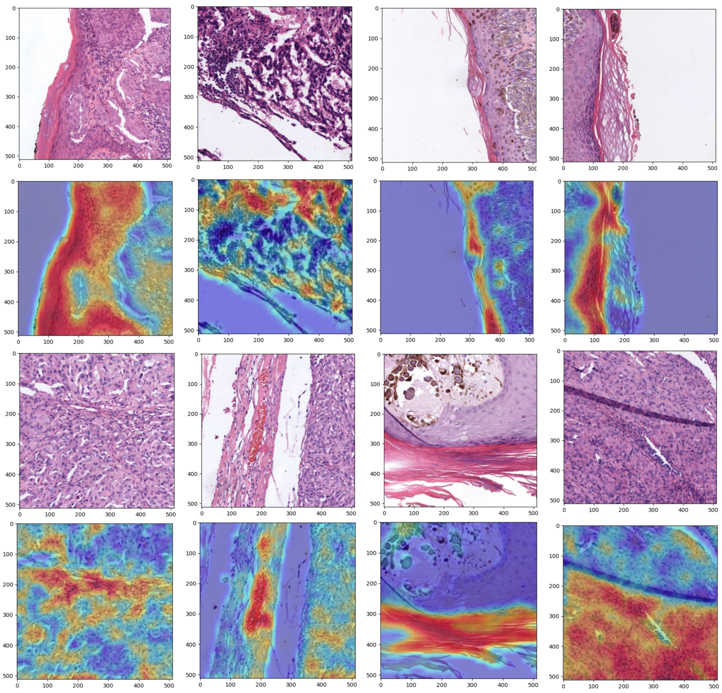


FIGURE 43: Eight random patches of skin cancer data set. The odd rows show the original images. The even rows are the Grad-CAMs mapped on top of the original images.

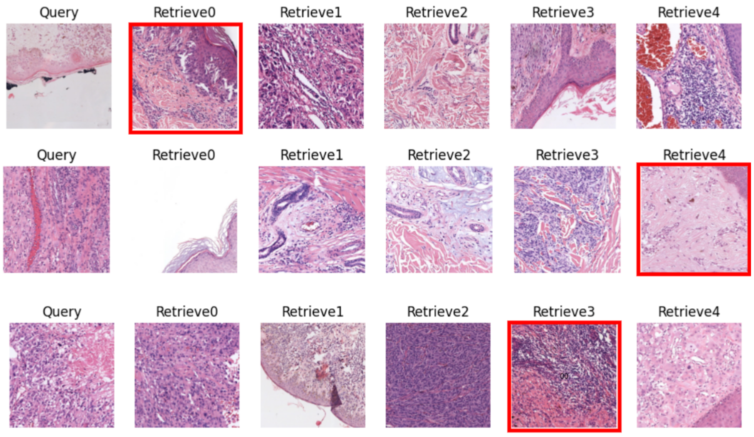


FIGURE 44: Three random images with their top 5 retrieved images. The returned images with a different label are marked in red.

5.4.3 STUMP searching

Spitzoid lesions present a diagnostic conundrum due to their intricate histology, creating challenges in establishing clear parameters between benign nevi and potentially malignant melanomas. In particular, STUMP cases require careful assessment to determine their true nature. In this paper, the model was isolated from STUMP images while training and testing steps for the above experiments. Figure 45 explains that this experiment mimics the situation that pathologists face with a STUMP tissue and need a second opinion to assist them in writing the diagnosis report. To do so, the Skin-twins model, which is trained on the binary data set including *Benign* and *Malignant*, is fed by the STUMP queries. The Skin-twins model displays top K similar patches to pathologists. The top K return patches with their labels providing some clues and hints regarding histopathological features and patterns for the pathologists.

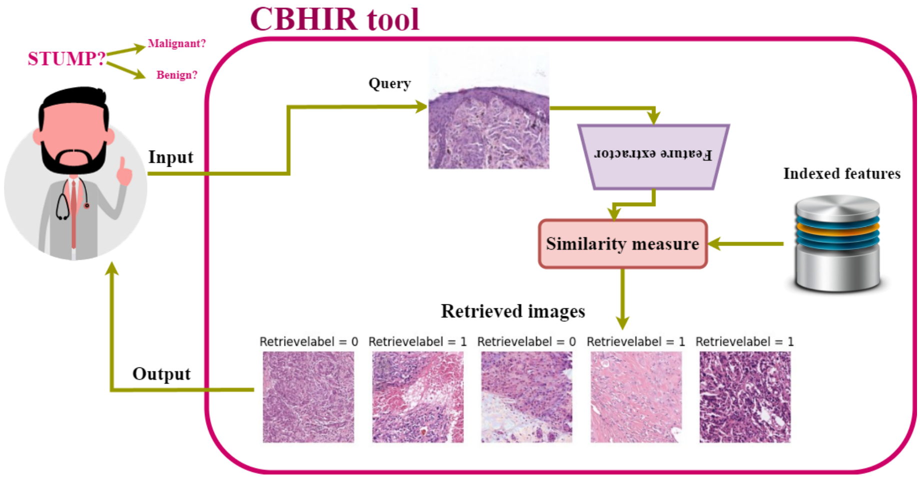


FIGURE 45: A real-world scenario that a pathologist faces with a STUMP case and uploads the query as input to the proposed CBHIR approach. First, the model feeds the query to the proposed FE and extracts the query features. Second, a similarity measure applies to the query features and indexed features of the database. Then, it is time to rank and visualize the top 5 similar patches to pathologists as the output of the proposed CBHIR.

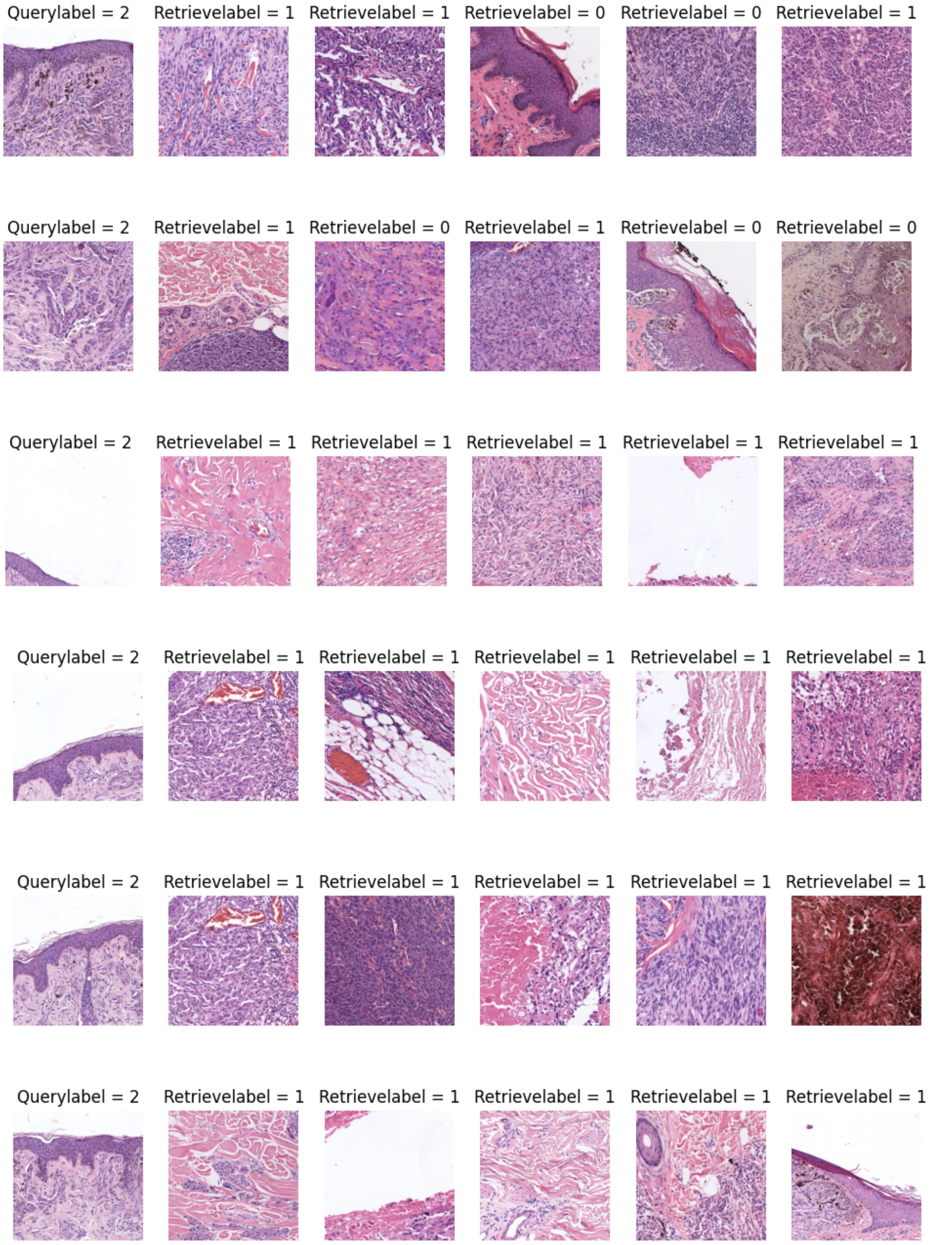


FIGURE 46: Six random examples of STUMP queries and their retrieved images. The STUMP queries were labeled = 2, Malignant = 1, and Benign = 0.

In Figure 46, for each query STUMP (labeled = 2), five top similar patches and their labels are retrieved from the data set, including *Benign* (labeled = 0) and *Malignant* (labeled = 1). In this case, pathologists can analyze the patterns in the retrieved tissues and compare them with the histopathological features of their STUMP query. The final decision depends on the pathologists' point of view and their knowledge. For instance, in the case of line 3 in Figure 46, it can be concluded that the STUMP tissue is Malignant since all the retrieved images are labeled = 1.

In Figure 47 there are Grad-CAM plots of three random STUMP queries with their top 5 retrieved images. The histopathological patterns and the regions of interest at the patch level that the Skin-twins network paid attention to retrieve the patches related to the STUMP queries are highlighted. As can be seen in Figure 46 and Figure 47, pathologists can diagnose and grade the tissue more confidently by receiving the top 5 similar patches and their corresponding grades in addition to the highlighted histopathological patterns and the tumor cells.

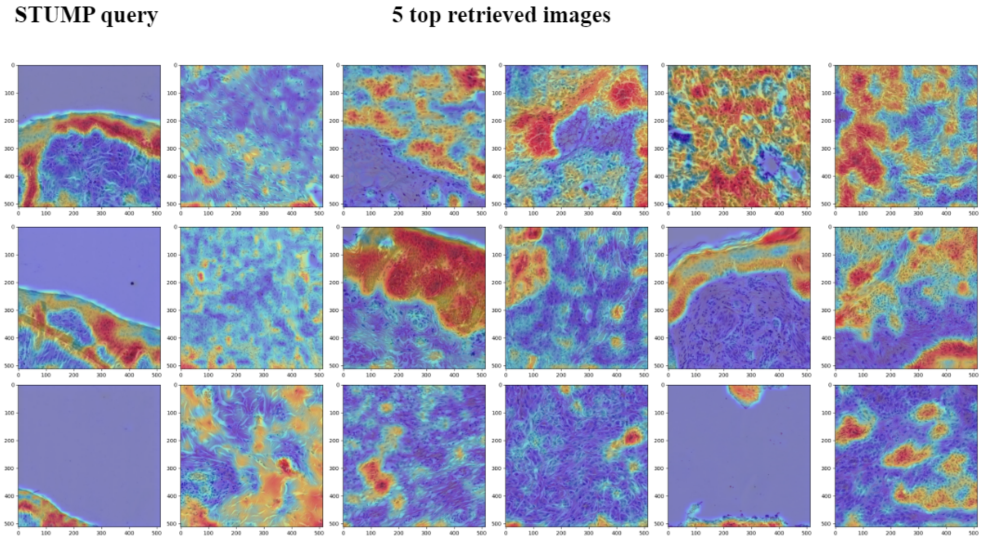


FIGURE 47: Three random STUMP patches with their top 5 retrieved image. The Grad-CAMs show the main patterns that the CBHIR took into account to find the similarity between the query and the retrieved patches.

5.4.4 Comparison with classification approaches

Classification delivers the label of the input image to provide a second opinion on the query tissue for the pathologists. While, CBHIR not only provides the label, but also retrieves similar patches. The outputs of CBHIR assist pathologists in finding the relevant histopathological features in the return images.

The only similarity between a CBHIR and a classifier is that both report the labels. A classifier has done this as a black box, while in CBHIR, pathologists have this opportunity to analyze the retrieved patches by their knowledge to understand why the model retrieves images. So, a classification framework by its nature, directly reports the label. However, in CBHIR the label is obtained based on the label of the retrievals and the final decision of the pathologists by analyzing the patterns of the query and the returned images from a CBHIR model.

In order to evaluate the performance of the CBHIR, the retrieved images have to be in the same cancer type as the query. Table 21 presents a comparative analysis between the Skin-twins CBHIR model and the classifiers reported in [166] which have been recently published.

In this table, the results of Skin-twins are reported in the top 3 retrieved images. The obtained results of Skin-twins CBHIR could exhibit higher performance compared with all the mentioned methods in terms of F1score. Furthermore, assume that K was set as 5 as the most used amount of K in the previous studies [27, 24, 153, 88, 70, 25, 13], the F1score and accuracy are 96% and 93%, respectively, much higher than the results of classifiers.

Table 22 provides the comparison between the Breast-twins CBHIR model at the top 5 and the recent classifiers on the BreaKHis data set. Breast-twins CBHIR could reach a higher accuracy than the recent classifiers, although, in some studies, it is slightly under-performing a classifier.

TABLE 21: A comparison between classifiers and the Skin-twins CBHIR, $K = 3$.

Method	**VGG16** [166]	**ResNet-34** [166]	**ResNet-50** [166]	**ResNet-34** ($\overset{\alpha}{S}$) [166]	**Skin-twins**
F1score	0.70	0.84	0.76	0.74	**0.94**
Accuracy	0.72	0.85	0.77	0.68	**0.90**

TABLE 22: A comparison between the recently proposed classifiers and the Breast-twins CBHIR, $K = 5$.

Method	**PFTAS + SVM** [64]	**IDSNet** [49]	**FE -VGGNET16 -SVM(POLY)** [169]	**FCN-Bi -LSTM** [170]	**AE + Siamese Network** [45]	**Breast -twins**
Accuracy	0.823	0.845	0.934	0.942	**0.967**	0.92

For instance, in [45], the authors trained the model twice. Once the model was trained as an Auto Encoder (AE), the model was re-trained as a Siamese classifier. This double-trained AE + Siamese network could deliver 0.04 amount more accuracy than the proposed Breast-twins CBHIR model. These two tables (Table 22 and Table 21) prove that the proposed CBHIR models could only retrieve images within the same cancer type thanks to the Breast/Skin-twins.

5.5 Conclusions and future lines

In this paper, we have proposed a novel Content-Based Histopathological Image Retrieval (CBHIR) approach on breast and skin cancer data sets based on a costume-built Siamese network. These methods are named Breast-twins and Skin-twins in this paper since it is a pairwise framework. In order to train them, a contrastive loss function, which is a distance-based loss, has been applied.

The proposed Siamese network is used to extract the discriminative features of the patches to feed to the search engine. In the retrieval step, the trained model compares the query image with all the images in the data set to find and rank similar patches based on Euclidean distance. In this paper, we evaluated the performance of the approach on both data sets at the top 1, 3, and 5. To the best of the author's knowledge, this is the first time that a CBHIR approach reached a high performance at the top first retrieved patches on breast and skin cancer. This benefit of the proposed approach makes it more reliable for pathologists to consider it in their daily workflow.

On one hand, the proposed Breast-twins network works with the highest magnification available in the BreaKHis data set. It is designed to work on images at 400× magnification since it can assist pathologists in measuring the amount of mitosis as an important criterion in breast cancer grading. We compared the performance of the Breast-twins network with some state-of-the-art CBHIR techniques. As a result of the comparisons, the proposed CBHIR could overpass the other techniques with a significant difference at the top first retrieved patches. Moreover, by comparing the obtained results with cutting-edge classifiers, we illustrated the high performance of the model in retrieving the patches in the same cancer type.

On the other hand, we implemented the recently published CBHIR framework based on Convolution Auto Encoder (CAE) into the skin data set to provide a comparison between the Skin-twins CBHIR approach and the CAE CBHIR algorithm. According to the comparison, the proposed Skin-twins CBHIR model overpassed the CAE CBHIR model with the difference of 67% of precision, 68% of Recall, and 67.5% of F1Score higher amount at the top first retrieved images.

In this study, as far as the author is aware, for the first time, a CBHIR technique was proposed to tackle the challenges of grading Spitzoid Tumors of Uncertain Malignant Potential (STUMP) queries by retrieving top K similar patches. Since STUMP tissues are highly challenging for pathologists, by retrieving top K similar patches, pathologists can analyze the histopathological patterns based on their knowledge and grade the query STUMP more accurately and faster. As far as the importance of histopathological patterns in cancer diagnosis, we reported the Grad-CAM figures of the STUMP query and the retrieved patches to highlight the region of interest for the Skin-twins network for finding similarity.

Based on the discussions in this paper and the high performance of the proposed CBHIR model in breast and skin cancer data sets with promising properties in histopathological diagnosis support across different cancer types, the CBHIR models can yield values in the daily workflow of hospitals in cancer diagnosis.

In further investigations, Federated Learning (FL) can enhance the performance of the CBHIR framework by training the model with richer data sets coming from different centers. This enhancement can assist pathologists in different medical centers to have more accurate cancer diagnoses with a generalized technique. In addition, expanding the breast data set with different magnifications can provide a multi-magnification CBHIR model that can retrieve patches from different levels of magnifications. Consequently, pathologists can reach similar histopathological images at several levels of magnification.

Another future line can be dedicated to exploring histology foundation models within CBHIR systems. CBHIR models encompass both image and text data to enhance the retrieval of histology images based on their content.

5.6 Author Contributions

Zahra Tabatabaei: Conceptualization, Methodology, Analyzing, Writing- Original draft, Reviewing & editing, Formal analysis, visualization.

Adrián Colomer: Supervision, Conceptualization, Methodology, and Review.

Javier Oliver Moll: Supervision, Conceptualization, Methodology.

Valery Naranjo: Supervision, Conceptualization, Methodology.

5.7 Acknowledgment

This study is funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 860627 (CLARIFY Project [3]). The work of Adrián Colomer has been supported by Ayuda a Primeros Proyectos de Investigación (PAID-06-22), Vicerrectorado de Investigación de la Universitat Politècnica de València (UPV).

We would like to express our gratitude to Dr. Gonzalo Safont Armero and Daniel Fernández Gómez for their valuable assistance in this research.

[3] http://www.clarify-project.eu/

Chapter 6

Conclusion

This Chapter corresponds to the findings from each paper to the final aim of the thesis. It concludes the main research goals of the thesis for each proposal learning framework.

6.1 Global remarks

In this thesis, we have designed, developed, and validated different DL-based methods for CBHIR to support cancer diagnosis and improve expert efficiency to aid decision-making. We proposed some cutting-edge DL-based techniques through different strategies and scenarios to assist pathologists in cancer diagnosis and grading. These proposed techniques undergo different types of learning paradigms including unsupervised, federated, and contrastive. Furthermore, in the paradigm of unsupervised learning, we investigated the impacts of color normalization on CBHIR methodologies.

This thesis mainly focused on histopathological images by analyzing patches of WSIs. However, lack of the annotated data sets is one of the main challenges in training DL-based models to employ as an FE in the search engine part of a CBHIR framework. The unsupervised learning strategy of this thesis is mainly focused on tackling this issue. So, the proposed techniques are robust against this issue with high performance in finding images with the same cancer type as the query. Also, this thesis provides in-depth statistical analyses of the impacts of three different color normalization techniques on the final results of a CBHIR framework under the umbrella of unsupervised learning. In addition to the unsupervised learning strategy, this thesis also includes federated learning-based methods to deal with the lack of labeled data, limited storage capacity, medical regulations, etc. In further investigations, we deploy a custom-built Siamese network as a contrastive learning-based technique to extract the features based on the distance function.

In short from a broad perspective, all these investigations are dedicated to extracting the most significant features of the patches to find the similarities among cases more accurately and faster. To achieve this objective, we not only explored various DL-based structures but also scrutinized the pre-processing step to enhance the efficiency of the trained FE in capturing features relevant to our primary goal. These features were fed into the search engine to find the similarities.

In each chapter, in-depth statistical analyses have been performed to evaluate the performance of the proposed CBHIR framework. Besides, an extended visual evaluation was conducted to report the performance of the proposed CBHIR frameworks. Visual evaluation not only provides a comprehensive understanding of the proposed methods in each chapter, but it also increases the reliability of the methods for pathologists. To do so, attention mechanisms and Grad-CAM maps have been employed, enabling the visualization of relevant elements and their contributions to retrieve relevant patches. This can provide clear insights into the decision-making process and instill trust among the users. So, understanding the model's behavior becomes more accessible, allowing for the identification of biases, limitations, and potential areas for improvement. These aspects underscore the importance of transparency and explainability in CBHIR, ensuring responsible and trustworthy deployment across diverse methods.

6.2 Specific remarks

In Chapter 2, we have presented a CBHIR framework on breast and prostate cancer as the two most prevalent cancer types. In this chapter, by discarding the decoder part of a custom-built CAE, an unsupervised FE was applied to the histological patches of prostate and breast cancer to extract their deep and meaningful features. Then, these features were used to measure the similarities of the patches to retrieve top K images. Although the proposed framework was unsupervised, it could overpass the other state-of-the-art methods. The performance evaluation of this technique for both data sets was conducted statically and visually. Also, an indirect evaluation was conducted by comparing the obtained results with some state-of-the-art classifiers. The main aim of this evaluation was to scrutinize whether the retrieved images are in the same cancer type as the query. The proposed framework reached precision rates of 91% on BreaKHis and 70% of precision on SICAPv2. Furthermore, an external evaluation was conducted on the prostate cancer data set. The used FE was trained on the SICAPv2 data set and it was tested on the Arvaniti data set to prove the generalization of the proposed technique. The external evaluation achieved an accuracy of 81% in the top 5.

In Chapter 3, color normalization as a pre-processing step has been applied to the proposed CAE. The investigation focused on studying patches extracted from the five centers of the CAM17 dataset, which serves as a breast cancer data set. These patches were normalized by BKSVD, Vah, and Mac. The reported results highlight a crucial observation: the effectiveness of a color normalization technique is not only evident in diminishing intra-center variance but also significantly contributes to the enhanced overall performance of a CBHIR framework. It is noteworthy to mention that utilizing a non-sufficient CN technique not only cannot improve the final results but also can decrease the performance of the main model. Employing BKSVD as the state-of-the-art color normalization technique yields 18% higher accuracy than working with the original color space of the patches.

In Chapter 4, we have mimicked a minimized scenario of worldwide CBMIR which connects different medical centers and pathologists from the whole of the world. In this case, we evaluated our proposed technique in three scenarios with two breast cancer data sets. First, we trained the unsupervised CAE with the distributed data of BreaKHis 400× and CAM17. The model was trained 11.44 hours faster than local training and it yielded 98.4% and 98.1% F1 on BreaKHis and CAM 17, respectively, which overpasses the local training results. Then, we extended this context by training the model on the BreaKHis data set within four distributed nodes. In contrast to the local training, the obtained results were higher than local training and the models were trained 25.53 hours faster. In this scenario, training four distinct models locally takes 32.36 hours in total while it only took 6.83 hours to train all four together thanks to FL. Training time is an important criterion in the training process but of course, accuracy is also important. The reported results, visually and statistically, have proved that the proposed FedCBMIR is more generalized at four different magnifications of breast cancer.

The last scenario in this chapter was conducted on the assumption that pathologists not only need similar patches but also need them at different magnification levels. The proposed framework answered this demand by retrieving similar patches to the query at 4 different magnifications of the BreaKHis data set.

In Chapter 5, a custom-built Siamese network was applied to the breast and skin cancer data sets. This is the first time that a Siamese network was conducted as an FE in a CBHIR framework on skin cancer, to the best of our beliefs. In this Chapter, the proposed technique could overpass CAE as the previously proposed techniques at the top first retrieval with 67% higher precision. In this thesis, retrieval evaluation at the top first on breast and skin cancer data sets was conducted for the first time, to the best of the author's knowledge. The obtained result proves the power of the proposed technique in retrieving similar patches at the top first images. Besides, in this Chapter, we dedicated some experiments on the STUMP cases of skin cancer as one of the challenging cancer types for pathologists. Therefore, pathologists can conclude the grade of the STUMP query based on the grade of the retrieved patches. It was the first time that a study implemented such experiments on STUMP queries, as far as we are aware. The proposed framework was evaluated by conducting an extended comparison with the state-of-the-art papers in terms of Grad-CAM maps, TSNE plots, visual evaluation, and statistical results.

In summary, the DL-based methods detailed in this thesis have proven to be a valuable tool to assist pathologists in cancer diagnosis whether for prostate, breast, or skin cancer. Nevertheless, this thesis has covered just a very small portion of the full potential of DL-based methods for digital pathology.

6.3 Future work

In this thesis, several methodologies have been taken into account and explored to mitigate the dependency of DL-based methods on annotated data sets. The methodologies in this thesis have several goals including a retrieval framework with a high level of transparency, trustworthiness, and reliability for pathologists. However, there are many research possibilities for further investigations in the explainability domain. For instance, in future studies, blockchain and security systems can be integrated with the concepts of CBHIR and FedCBMIR to increase the security and safety of the framework. Moreover, text-based search can also integrate with CBHIR to provide the prescribed medications from the previous cases. Integrating blockchain, FL, and cross-modal search can move the CBHIR to an upper level of explainability and offer a more comprehensive understanding of model decisions for pathologists. In other words, these advanced methods can open the black box for pathologists and yield deeper insights from the developed models. The other future line can be dedicated to cross-modal retrieval via foundation models. WSIs contain dense information and even individual image patches can hold unique, complex patterns according to the characterization of the tissue. This feature of WSIs which represent thousands of sub-types of diseases can not be concluded in a single label that might fail to capture the complexity of the field. An interconnected representation such as natural language descriptions can go beyond a singular categorical label. Another unexplored avenue can be filling the gap to tackle the absence of spatial grounding in histopathology image caption in the visual instruction data set which creates question/answer pairs for individual image patches.

Finally, we would like to highlight some future possibilities regarding real-world applications. Future research directions include the exploration of advanced unsupervised FE to train the models without requiring annotated data sets. The other future possibility is the integration of multi-modal data, such as clinical information. This can offer a more comprehensive understanding of pathology. Additionally, collaboration with clinical partners offers opportunities for practical integration, potentially impacting patient stratification and treatment planning.

Overall, these research topics and leveraging them into the concept of CBHIR offer a better automatic cancer diagnosis.

Merits

Journal papers

- **Tabatabaei, Z**., Wang, Y., Colomer, A., Oliver Moll, J., Zhao, Z., & Naranjo, V. (2023). WWFedCBMIR: World-Wide Federated Content-Based Medical Image Retrieval. Bioengineering, 10(10), 1144.
- **Tabatabaei, Z**., Colomer, A., Oliver Moll, J., & Naranjo, V. (2023). Toward More Transparent and Accurate Cancer Diagnosis With an Unsupervised CAE Approach. IEEE Access, vol. 11, pp. 143387-143401, 2023.
- **Tabatabaei, Z.**, Pérez Bueno, F., Colomer, A., Moll, JO, Molina, R., & Naranjo, V. (2024). A Comparative Analysis of the Effects of Color Normalization Techniques." Applied Sciences 14.5 (2024): 2063.
- **Tabatabaei, Z**., Colomer, A., Oliver Moll, J., & Naranjo, V. (2024). Siamese Content-based Search Engine for a More Transparent Skin and Breast Cancer Diagnosis through Histological Imaging. Under review by Computers in Biology and Medicine. 2024
- Mosquera-Zamudio, A., Launet, L., **Tabatabaei, Z.**, Parra-Medina, R., Colomer, A., Oliver Moll, J., ... & Naranjo, V. (2022). Deep Learning for Skin Melanocytic Tumors in Whole-Slide Images: A Systematic Review. Cancers, 15(1), 42.

International Conference papers

- **Tabatabaei, Z.**, Colomer, A., Engan, K., Oliver, J., & Naranjo, V. (2022, June). Residual block convolutional auto encoder in content-based medical image retrieval. In 2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP) (pp. 1-5). IEEE.
- **Tabatabaei, Z.**, Colomer, A., Engan, K., Oliver, J., & Naranjo, V. (2023, September). Self-supervised learning of a tailored Convolutional Auto Encoder for histopathological prostate grading. In 2023 31st European Signal Processing Conference (EUSIPCO) (pp. 980-984). IEEE.

Bibliography

[1] Andreas Kaplan and Michael Haenlein. “Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence”. In: *Business horizons* 62.1 (2019), pp. 15–25.

[2] Jie Xu, Kanmin Xue, and Kang Zhang. “Current status and future trends of clinical diagnoses via image-based deep learning”. In: *Theranostics* 9.25 (2019), p. 7556.

[3] Julio Silva-Rodríguez et al. “Going deeper through the Gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection”. In: *Computer Methods and Programs in Biomedicine* 195 (2020), p. 105637.

[4] Arthur T Skarin. *Atlas of Diagnostic Oncology E-Book*. Elsevier Health Sciences, 2015.

[5] Ashnil Kumar et al. “Content-based medical image retrieval: a survey of applications to multidimensional and multimodality data”. In: *Journal of digital imaging* 26 (2013), pp. 1025–1039.

[6] Şaban Öztürk. “Stacked auto-encoder based tagging with deep features for content-based medical image retrieval”. In: *Expert Systems with Applications* 161 (2020), p. 113693.

[7] M Srinivas et al. “Content based medical image retrieval using dictionary learning”. In: *Neurocomputing* 168 (2015), pp. 880–895.

[8] Farbod Khoraminia et al. “Artificial Intelligence in Digital Pathology for Bladder Cancer: Hype or Hope? A Systematic Review”. In: *Cancers* 15.18 (2023), p. 4518.

[9] Hamid Tizhoosh. *Searching is intelligence*. 2018. URL: https://thepathologist.com/diagnostics/searching-is-intelligence.

[10] Pooria Mazaheri et al. “Ranking Loss and Sequestering Learning for Reducing Image Search Bias in Histopathology”. In: *Available at SSRN 4216426* ().

[11] Maral Rasoolijaberi et al. “Multi-Magnification Image Search in Digital Pathology”. In: *IEEE Journal of Biomedical and Health Informatics* (2022).

[12] Shivam Kalra et al. “Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence”. In: *NPJ digital medicine* 3.1 (2020), pp. 1–15.

[13] Shivam Kalra et al. “Yottixel–an image search engine for large archives of histopathology whole slide images”. In: *Medical Image Analysis* 65 (2020), p. 101757.

[14] Morteza Babaie et al. “Classification and retrieval of digital pathology scans: A new dataset”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2017, pp. 8–16.

[15] World Health Organization et al. “Suicide worldwide in 2019: global health estimates”. In: (2021).

[16] World Health Organization et al. "The effect of occupational exposure to solar ultraviolet radiation on malignant skin melanoma and non-melanoma skin cancer: a systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury". In: (2021).

[17] Frank Pega et al. "Global, regional and national burdens of non-melanoma skin cancer attributable to occupational exposure to solar ultraviolet radiation for 183 countries, 2000–2019: a systematic analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury". In: *Environment International* (2023), p. 108226.

[18] N Loprieno. "International Agency for Research on Cancer (IARC) monographs on the evaluation of carcinogenic risk of chemicals to man:" relevance of data on mutagenicity"". In: *Mutation research* 31.3 (1975), p. 210.

[19] Marie Bø-Sande et al. "A Dual Convolutional Neural Network Pipeline for Melanoma Diagnostics and Prognostics". In: *Northern Lights Deep Learning Conference 2024*. 2023.

[20] Stefan Studer et al. "Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology". In: *Machine learning and knowledge extraction* 3.2 (2021), pp. 392–413.

[21] Kai Rakovic et al. "The use of digital pathology and artificial intelligence in histopathological diagnostic assessment of prostate cancer: a survey of prostate cancer UK supporters". In: *Diagnostics* 12.5 (2022), p. 1225.

[22] Asmaa Ibrahim et al. "Artificial intelligence in digital breast pathology: techniques and applications". In: *The Breast* 49 (2020), pp. 267–273.

[23] Zoe Apalla et al. "Epidemiological trends in skin cancer". In: *Dermatology practical & conceptual* 7.2 (2017), p. 1.

[24] Zahra Tabatabaei et al. "Residual block convolutional auto encoder in content-based medical image retrieval". In: *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE. 2022, pp. 1–5.

[25] Zahra Tabatabaei et al. "Toward More Transparent and Accurate Cancer Diagnosis With an Unsupervised CAE Approach". In: *IEEE Access* 11 (2023), pp. 143387–143401. DOI: 10.1109/ACCESS.2023.3343845.

[26] Zahra Tabatabaei et al. "Advancing Content-Based Histopathological Image Retrieval Pre-Processing: A Comparative Analysis of the Effects of Color Normalization Techniques". In: *Applied Sciences* 14.5 (2024), p. 2063.

[27] Zahra Tabatabaei et al. "WWFedCBMIR: World-Wide Federated Content-Based Medical Image Retrieval". In: *Bioengineering* 10.10 (2023), p. 1144.

[28] Zahra Tabatabaei et al. "Self-supervised learning of a tailored Convolutional Auto Encoder for histopathological prostate grading". In: *2023 31st European Signal Processing Conference (EUSIPCO)*. 2023, pp. 980–984. DOI: 10.23919/EUSIPCO58844.2023.10289741.

[29] Zahra Tabatabaei et al. "Siamese Content-based Search Engine for a More Transparent Skin and Breast Cancer Diagnosis through Histological Imaging". In: *arXiv preprint arXiv:2401.08272* (2024).

[30] Fernando Pérez-Bueno et al. "Bayesian K-SVD for H and E blind color deconvolution. Applications to stain normalization, data augmentation and cancer classification". In: *Computerized Medical Imaging and Graphics* 97 (2022), p. 102048.

[31] Marion Piñeros et al. “Scaling up the surveillance of childhood cancer: a global roadmap”. In: *JNCI: Journal of the National Cancer Institute* 113.1 (2021), pp. 9–15.

[32] Hardeep Singh, Ashley ND Meyer, and Eric J Thomas. “The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations”. In: *BMJ quality & safety* 23.9 (2014), pp. 727–731.

[33] World Health Organization. *Global action plan on physical activity 2018-2030: more active people for a healthier world*. World Health Organization, 2019.

[34] Hamid Reza Tizhoosh and Liron Pantanowitz. “Artificial intelligence and digital pathology: challenges and opportunities”. In: *Journal of pathology informatics* 9.1 (2018), p. 38.

[35] Saul Fuster et al. “Invasive Cancerous Area Detection in Non-Muscle Invasive Bladder Cancer Whole Slide Images”. In: *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE. 2022, pp. 1–5.

[36] Neel Kanwal et al. “Quantifying the effect of color processing on blood and damaged tissue detection in Whole Slide Images”. In: *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE. 2022, pp. 1–5.

[37] Iqra Kiran et al. “DenseRes-Unet: Segmentation of overlapped/clustered nuclei from multi organ histopathology images”. In: *Computers in Biology and Medicine* 143 (2022), p. 105267.

[38] Hongping Hu et al. “Breast cancer histopathological images recognition based on two-stage nuclei segmentation strategy”. In: *Plos one* 17.4 (2022), e0266973.

[39] Jun Xu et al. “Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images”. In: *IEEE transactions on medical imaging* 35.1 (2015), pp. 119–130.

[40] Wenyuan Li et al. “Path R-CNN for prostate cancer diagnosis and gleason grading of histological images”. In: *IEEE transactions on medical imaging* 38.4 (2018), pp. 945–954.

[41] Geert Litjens et al. “A survey on deep learning in medical image analysis”. In: *Medical image analysis* 42 (2017), pp. 60–88.

[42] Wenyuan Li et al. “High resolution histopathology image generation and segmentation through adversarial training”. In: *Medical Image Analysis* 75 (2022), p. 102251.

[43] Francesco Prinzi et al. “A Yolo-Based Model for Breast Cancer Detection in Mammograms”. In: *Cognitive Computation* (2023), pp. 1–14.

[44] Jiayun Li et al. “A multi-resolution model for histopathology image classification and localization with multiple instance learning”. In: *Computers in biology and medicine* 131 (2021), p. 104253.

[45] Min Liu et al. “Breast Histopathological Image Classification Method Based on Autoencoder and Siamese Framework”. In: *Information* 13.3 (2022), p. 107.

[46] Nouman Ahmad, Sohail Asghar, and Saira Andleeb Gillani. “Transfer learning-assisted multi-resolution breast cancer histopathological images classification”. In: *The Visual Computer* 38.8 (2022), pp. 2751–2770.

[47] Rui Man, Ping Yang, and Bowen Xu. "Classification of Breast Cancer Histopathological Images Using Discriminative Patches Screened by Generative Adversarial Networks". In: *IEEE Access* 8 (2020), pp. 155362–155377. DOI: 10.1109/ACCESS.2020.3019327.

[48] Md Zahangir Alom et al. "Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network". In: *Journal of digital imaging* 32.4 (2019), pp. 605–617.

[49] Xia Li et al. "Classification of breast cancer histopathological images using interleaved DenseNet with SENet (IDSNet)". In: *PloS one* 15.5 (2020), e0232127.

[50] Mohammad Reza Abbasniya et al. "Classification of Breast Tumours Based on Histopathology Images Using Deep Features and Ensemble of Gradient Boosting Methods". In: *arXiv preprint arXiv:2209.01380* (2022).

[51] Fatemeh Taheri, Kambiz Rahbar, and Pedram Salimi. "Effective features in content-based image retrieval from a combination of low-level features and deep Boltzmann machine". In: *Multimedia Tools and Applications* (2022), pp. 1–24.

[52] Agus Eko Minarno et al. "CNN Based Autoencoder Application in Breast Cancer Image Retrieval". In: *2021 International Seminar on Intelligent Technology and Its Applications (ISITIA)*. 2021, pp. 29–34. DOI: 10.1109/ISITIA52817.2021.9502205.

[53] Forrest Iandola et al. "Densenet: Implementing efficient convnet descriptor pyramids". In: *arXiv preprint arXiv:1404.1869* (2014).

[54] Parsa Ashrafi Fashi et al. "A self-supervised contrastive learning approach for whole slide image representation in digital pathology". In: *Journal of Pathology Informatics* (2022), p. 100133.

[55] Lei Zheng et al. "Design and analysis of a content-based pathology image retrieval system". In: *IEEE transactions on information technology in biomedicine* 7.4 (2003), pp. 249–255.

[56] Xiaoshuang Shi et al. "Supervised graph hashing for histopathology image retrieval and classification". In: *Medical image analysis* 42 (2017), pp. 117–128.

[57] Liron Pantanowitz et al. "A digital pathology solution to resolve the tissue floater conundrum". In: *Archives of pathology & laboratory medicine* 145.3 (2021), pp. 359–364.

[58] Juan C Caicedo, Fabio A González, and Eduardo Romero. "Content-based histopathology image retrieval using a kernel-based semantic annotation framework". In: *Journal of biomedical informatics* 44.4 (2011), pp. 519–528.

[59] Xin Qi et al. "Content-based histopathology image retrieval using CometCloud". In: *BMC bioinformatics* 15.1 (2014), pp. 1–17.

[60] Akshay Sridhar, Scott Doyle, and Anant Madabhushi. "Content-based image retrieval of digitized histopathology in boosted spectrally embedded spaces". In: *Journal of pathology informatics* 6.1 (2015), p. 41.

[61] Jin Tae Kwak et al. "Automated prostate tissue referencing for cancer detection and diagnosis". In: *BMC bioinformatics* 17.1 (2016), pp. 1–12.

[62] Shujin Zhu et al. "Multiple disjoint dictionaries for representation of histopathology images". In: *Journal of Visual Communication and Image Representation* 55 (2018), pp. 243–252.

[63] Xiaofan Zhang et al. "Towards large-scale histopathological image analysis: Hashing-based image retrieval". In: *IEEE Transactions on Medical Imaging* 34.2 (2014), pp. 496–506.

[64] Fabio A Spanhol et al. "A dataset for breast cancer histopathological image classification". In: *Ieee transactions on biomedical engineering* 63.7 (2015), pp. 1455–1462.

[65] Eirini Arvaniti et al. "Automated Gleason grading of prostate cancer tissue microarrays via deep learning". In: *Scientific reports* 8.1 (2018), pp. 1–11.

[66] Wouter Bulten et al. "Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge". In: *Nature medicine* 28.1 (2022), pp. 154–163.

[67] Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. "Training very deep networks". In: *Advances in neural information processing systems* 28 (2015).

[68] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[69] Yun Gu and Jie Yang. "Multi-level magnification correlation hashing for scalable histopathological image retrieval". In: *Neurocomputing* 351 (2019), pp. 134–145.

[70] Narayan Hegde et al. "Similar image search for histopathology: SMILY". In: *NPJ digital medicine* 2.1 (2019), p. 56.

[71] Neville Mehta, Alomari Raja'S, and Vipin Chaudhary. "Content based subimage retrieval system for high resolution pathology images using salient interest points". In: *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2009, pp. 3719–3722.

[72] Shuying Liu and Weihong Deng. "Very deep convolutional neural network based image classification using small training sample size". In: *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. 2015, pp. 730–734. DOI: 10.1109/ACPR.2015.7486599.

[73] Khalil Al-Hussaeni et al. "CNN-Based Pill Image Recognition for Retrieval Systems". In: *Applied Sciences* 13.8 (2023), p. 5050.

[74] Saman Khalil et al. "Enhancing Ductal Carcinoma Classification Using Transfer Learning with 3D U-Net Models in Breast Cancer Imaging". In: *Applied Sciences* 13.7 (2023), p. 4255.

[75] Hamid Reza Shahdoosti and Adel Mehrabi. "Multimodal image fusion using sparse representation classification in tetrolet domain". In: *Digital Signal Processing* 79 (2018), pp. 9–22.

[76] Fadwa Alrowais et al. "Enhanced Pelican Optimization Algorithm with Deep Learning-Driven Mitotic Nuclei Classification on Breast Histopathology Images". In: *Biomimetics* 8.7 (2023), p. 538.

[77] Anika Strittmatter, Anna Caroli, and Frank G Zöllner. "A Multistage Rigid-Affine-Deformable Network for Three-Dimensional Multimodal Medical Image Registration". In: *Applied Sciences* 13.24 (2023), p. 13298.

[78] Hamid Reza Shahdoosti and Zahra Tabatabaei. "MRI and PET/SPECT image fusion at feature level using ant colony based segmentation". In: *Biomedical Signal Processing and Control* 47 (2019), pp. 63–74.

[79] Fuhui Long, Hongjiang Zhang, and David Dagan Feng. "Fundamentals of content-based image retrieval". In: *Multimedia Information Retrieval and Management: Technological Fundamentals and Applications*. Springer, 2003, pp. 1–26.

[80] Arnold WM Smeulders et al. "Content-based image retrieval at the end of the early years". In: *IEEE Transactions on pattern analysis and machine intelligence* 22.12 (2000), pp. 1349–1380.

[81] Francesco Bianconi, Jakob N Kather, and Constantino Carlos Reyes-Aldasoro. "Experimental assessment of color deconvolution and color normalization for automated classification of histology images stained with hematoxylin and eosin". In: *Cancers* 12.11 (2020), p. 3337.

[82] Thaína A Azevedo Tosta et al. "Computational normalization of H&E-stained histological images: Progress, challenges and future potential". In: *Artificial intelligence in medicine* 95 (2019), pp. 118–132.

[83] Peter Bandi et al. "From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge". In: *IEEE transactions on medical imaging* 38.2 (2018), pp. 550–560.

[84] Massimo Salvi et al. "The impact of pre-and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis". In: *Computers in Biology and Medicine* 128 (2021), p. 104129.

[85] Surbhi Vijh, Mukesh Saraswat, and Sumit Kumar. "A new complete color normalization method for H&E stained histopatholgical images". In: *Applied Intelligence* (2021), pp. 1–14.

[86] Santanu Roy et al. "A study about color normalization methods for histopathology images". In: *Micron* 114 (2018), pp. 42–61.

[87] Bogdan Ionescu et al. "ImageCLEF 2023 Highlight: Multimedia Retrieval in Medical, Social Media and Content Recommendation Applications". In: *European Conference on Information Retrieval*. Springer. 2023, pp. 557–567.

[88] Yushan Zheng et al. "Size-scalable content-based histopathological image retrieval from database that consists of WSIs". In: *IEEE journal of biomedical and health informatics* 22.4 (2017), pp. 1278–1287.

[89] Jorge A Vanegas, John Arevalo, and Fabio A González. "Unsupervised feature learning for content-based histopathology image retrieval". In: *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE. 2014, pp. 1–6.

[90] Komal Nain Sukhia et al. "Content-based histopathological image retrieval using multi-scale and multichannel decoder based LTP". In: *Biomedical Signal Processing and Control* 54 (2019), p. 101582.

[91] Abtin Riasatian et al. "Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides". In: *Medical Image Analysis* 70 (2021), p. 102032.

[92] M Tarek Shaban et al. "Staingan: Stain style transfer for digital histological images". In: *2019 Ieee 16th international symposium on biomedical imaging (Isbi 2019)*. IEEE. 2019, pp. 953–956.

[93] Marc Macenko et al. "A method for normalizing histology slides for quantitative analysis". In: *2009 IEEE international symposium on biomedical imaging: from nano to macro*. IEEE. 2009, pp. 1107–1110.

[94] Abhishek Vahadane et al. "Structure-preserving color normalization and sparse stain separation for histological images". In: *IEEE transactions on medical imaging* 35.8 (2016), pp. 1962–1971.

[95] Aïcha BenTaieb and Ghassan Hamarneh. "Adversarial stain transfer for histopathology image analysis". In: *IEEE transactions on medical imaging* 37.3 (2017), pp. 792–802.

[96] Farhad Ghazvinian Zanjani et al. "Stain normalization of histopathology images using generative adversarial networks". In: *2018 IEEE 15th International symposium on biomedical imaging (ISBI 2018)*. IEEE. 2018, pp. 573–577.

[97] David Tellez et al. "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology". In: *Medical image analysis* 58 (2019), p. 101544.

[98] Min Chen et al. "Deep feature learning for medical image analysis with convolutional autoencoder neural network". In: *IEEE Transactions on Big Data* 7.4 (2017), pp. 750–758.

[99] Euijoon Ahn et al. "Unsupervised domain adaptation to classify medical images using zero-bias convolutional auto-encoders and context-based feature augmentation". In: *IEEE transactions on medical imaging* 39.7 (2020), pp. 2385–2394.

[100] Mohammad I Daoud et al. "Content-based image retrieval for breast ultrasound images using convolutional autoencoders: A feasibility study". In: *2019 3rd International Conference on Bio-engineering for Smart Technologies (BioSMART)*. IEEE. 2019, pp. 1–4.

[101] Seonyeong Park et al. "Autoencoder-inspired convolutional network-based super-resolution method in MRI". In: *IEEE Journal of Translational Engineering in Health and Medicine* 9 (2021), pp. 1–13.

[102] Mohammad I. Daoud et al. "Content-based Image Retrieval for Breast Ultrasound Images using Convolutional Autoencoders: A Feasibility Study". In: *2019 3rd International Conference on Bio-engineering for Smart Technologies (BioSMART)*. 2019, pp. 1–4. DOI: 10.1109/BIOSMART.2019.8734190.

[103] Yushan Zheng et al. "Adaptive color deconvolution for histological WSI normalization". In: *Computer methods and programs in biomedicine* 170 (2019), pp. 107–120.

[104] Zhangjie Cao et al. "Hashnet: Deep learning to hash by continuation". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 5608–5617.

[105] Melina Arnold et al. "Current and future burden of breast cancer: Global statistics for 2020 and 2040". In: *The Breast* 66 (2022), pp. 15–23.

[106] Tengfei Zhao et al. "RGSB-UNet: Hybrid Deep Learning Framework for Tumour Segmentation in Digital Pathology Images". In: *Bioengineering* 10.8 (2023), p. 957.

[107] R Rashmi, Keerthana Prasad, and Chethana Babu K Udupa. "Breast histopathological image analysis using image processing techniques for diagnostic purposes: A methodological review". In: *Journal of Medical Systems* 46 (2022), pp. 1–24.

[108] Pamela Morelli et al. "Analysis of errors in histology by root cause analysis: a pilot study". In: *Journal of preventive medicine and hygiene* 54.2 (2013), p. 90.

[109] World Health Organization et al. “Laboratory quality standards and their implementation”. In: (2011).

[110] Min Young Kim et al. “Diagnostic accuracy of breast cancer in core needle biopsy using a standardized reporting system”. In: *Journal of Clinical Pathology* 65.9 (2012), pp. 790–794.

[111] Vipul Baxi et al. “Digital pathology and artificial intelligence in translational medicine and clinical practice”. In: *Modern Pathology* 35.1 (2022), pp. 23–32.

[112] Abir Baâzaoui, Marwa Abderrahim, and Walid Barhoumi. “Dynamic distance learning for joint assessment of visual and semantic similarities within the framework of medical image retrieval”. In: *Computers in Biology and Medicine* 122 (2020), p. 103833.

[113] Brendan McMahan et al. “Communication-efficient learning of deep networks from decentralized data”. In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282.

[114] Ming Y Lu et al. “Federated learning for computational pathology on gigapixel whole slide images”. In: *Medical image analysis* 76 (2022), p. 102298.

[115] Micah J Sheller et al. “Federated learning in medicine: facilitating multi institutional collaborations without sharing patient data”. In: *Scientific reports* 10.1 (2020), pp. 1–12.

[116] Wenqing Wang et al. “Two-stage content based image retrieval using sparse representation and feature fusion”. In: *Multimedia Tools and Applications* 81.12 (2022), pp. 16621–16644.

[117] Neville Mehta, Raja’ S. Alomari, and Vipin Chaudhary. “Content based subimage retrieval system for high resolution pathology images using salient interest points”. In: *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2009, pp. 3719–3722. DOI: 10.1109/IEMBS.2009.5334811.

[118] David G Lowe. “Object recognition from local scale-invariant features”. In: *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. Ieee. 1999, pp. 1150–1157.

[119] Keith Bonawitz et al. “Towards federated learning at scale: System design”. In: *Proceedings of machine learning and systems* 1 (2019), pp. 374–388.

[120] Sashank Reddi et al. “Adaptive federated optimization”. In: *arXiv preprint arXiv:2003.00295* (2020).

[121] Alexander Ziller et al. “Pysyft: A library for easy federated learning”. In: *Federated Learning Systems: Towards Next-Generation AI* (2021), pp. 111–139.

[122] Daniel J Beutel et al. “Flower: A friendly federated learning research framework”. In: *arXiv preprint arXiv:2007.14390* (2020).

[123] Laëtitia Launet et al. “Federating Medical Deep Learning Models from Private Jupyter Notebooks to Distributed Institutions”. In: *Applied Sciences* 13.2 (2023), p. 919.

[124] Daniel Truhn et al. “Encrypted federated learning for secure decentralized collaboration in cancer image analysis”. In: *medRxiv* (2022).

[125] Firas Khader et al. “Medical Diagnosis with Large Scale Multimodal Transformers–Leveraging Diverse Data for More Accurate Diagnosis”. In: *arXiv preprint arXiv:2212.09162* (2022).

[126] Lei Zhang et al. "FLSIR: Secure Image Retrieval Based on Federated Learning and Additive Secret Sharing". In: *IEEE Access* 10 (2022), pp. 64028–64042.

[127] Andrew H Fischer et al. "Hematoxylin and eosin staining of tissue and cell sections". In: *Cold spring harbor protocols* 2008.5 (2008), pdb–prot4986.

[128] David Tellez et al. "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology". In: *Medical image analysis* 58 (2019), p. 101544.

[129] M Esmel ElAlami. "A new matching strategy for content based image retrieval system". In: *Applied Soft Computing* 14 (2014), pp. 407–418.

[130] T Rajasenbagam, S Jeyanthi, and J Arun Pandian. "Detection of pneumonia infection in lungs from chest X-ray images using deep convolutional neural network and content-based image retrieval techniques". In: *Journal of Ambient Intelligence and Humanized Computing* (2021), pp. 1–8.

[131] Menglin Jiang et al. "Scalable histopathological image analysis via supervised hashing with multiple features". In: *Medical image analysis* 34 (2016), pp. 3–12.

[132] Haibo Wang et al. "Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features". In: *Journal of Medical Imaging* 1.3 (2014), pp. 034003–034003.

[133] Agus Eko Minarno et al. "Cnn based autoencoder application in breast cancer image retrieval". In: *2021 International Seminar on Intelligent Technology and Its Applications (ISITIA)*. IEEE. 2021, pp. 29–34.

[134] Neel Kanwal et al. "The devil is in the details: Whole slide image acquisition and processing for artifacts detection, color variation, and data augmentation: A review". In: *IEEE Access* 10 (2022), pp. 58821–58844.

[135] Jörg Reichrath et al. "Epidemiology of skin cancer". In: *Sunlight, vitamin D and skin cancer* (2014), pp. 120–140.

[136] Maria S Soengas and Scott W Lowe. "Apoptosis and melanoma chemoresistance". In: *Oncogene* 22.20 (2003), pp. 3138–3151.

[137] Thomas Wiesner et al. "Genomic aberrations in spitzoid melanocytic tumours and their implications for diagnosis, prognosis and therapy". In: *Pathology* 48.2 (2016), pp. 113–131.

[138] Bless Lord Y Agbley et al. "Federated Fusion of Magnified Histopathological Images for Breast Tumor Classification in the Internet of Medical Things". In: *IEEE Journal of Biomedical and Health Informatics* (2023).

[139] Chao-Hui Huang et al. "Time-efficient sparse analysis of histopathological whole slide images". In: *Computerized medical imaging and graphics* 35.7-8 (2011), pp. 579–591.

[140] David Ahmedt-Aristizabal et al. "A survey on graph-based deep learning for computational histopathology". In: *Computerized Medical Imaging and Graphics* 95 (2022), p. 102027.

[141] Neel Kanwal et al. "Vision Transformers for Small Histological Datasets Learned Through Knowledge Distillation". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2023, pp. 167–179.

[142] Hamid Reza Shahdoosti and Zahra Tabatabaei. "Object-based feature extraction for hyperspectral data using firefly algorithm". In: *International Journal of Machine Learning and Cybernetics* 11 (2020), pp. 1277–1291.

[143] Muhammad Owais et al. "Effective diagnosis and treatment through content-based medical image retrieval (CBMIR) by using artificial intelligence". In: *Journal of clinical medicine* 8.4 (2019), p. 462.

[144] Howard Lee and Yi-Ping Phoebe Chen. "Image based computer aided diagnosis system for cancer detection". In: *Expert Systems with Applications* 42.12 (2015), pp. 5356–5365. ISSN: 0957-4174. DOI: `https://doi.org/10.1016/j.eswa.2015.02.005`.

[145] Xiaoqing Li, Jiansheng Yang, and Jinwen Ma. "Recent developments of content-based image retrieval (CBIR)". In: *Neurocomputing* 452 (2021), pp. 675–689. ISSN: 0925-2312. DOI: `https://doi.org/10.1016/j.neucom.2020.07.139`.

[146] XinQuan Zhang et al. "A study of the diamond tool wear suppression mechanism in vibration-assisted machining of steel". In: *Journal of Materials Processing Technology* 214.2 (2014), pp. 496–506.

[147] Yibing Ma et al. "Generating region proposals for histopathological whole slide image retrieval". In: *Computer methods and programs in biomedicine* 159 (2018), pp. 1–10.

[148] Yibing Ma et al. "Breast histopathological image retrieval based on latent dirichlet allocation". In: *IEEE journal of biomedical and health informatics* 21.4 (2016), pp. 1114–1123.

[149] Elaheh Mahraban Nejad et al. "Transferred semantic scores for scalable retrieval of histopathological breast cancer images". In: *International Journal of Multimedia Information Retrieval* 7 (2018), pp. 241–249.

[150] Meenakshi Garg and Gaurav Dhiman. "A novel content-based image retrieval approach for classification using GLCM features and texture fused LBP variants". In: *Neural Computing and Applications* 33 (2021), pp. 1311–1328.

[151] Satya Rajendra Singh et al. "Joint triplet autoencoder for histopathological colon cancer nuclei retrieval". In: *Multimedia Tools and Applications* (2023), pp. 1–20.

[152] Gabriele Campanella et al. "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images". In: *Nature medicine* 25.8 (2019), pp. 1301–1309.

[153] Seyed Mohammad Alizadeh, Mohammad Sadegh Helfroush, and Henning Müller. "A novel Siamese deep hashing model for histopathology image retrieval". In: *Expert Systems with Applications* 225 (2023), p. 120169.

[154] Xiyue Wang et al. "RetCCL: clustering-guided contrastive learning for whole-slide image retrieval". In: *Medical image analysis* 83 (2023), p. 102645.

[155] Sobhan Hemati et al. "Learning binary and sparse permutation-invariant representations for fast and memory efficient whole slide image search". In: *Computers in Biology and Medicine* 162 (2023), p. 107026.

[156] Nasim Kayhan and Shervan Fekri-Ershad. "Content based image retrieval based on weighted fusion of texture and color features derived from modified local binary patterns and local neighborhood difference patterns". In: *Multimedia Tools and Applications* 80.21-23 (2021), pp. 32763–32790.

[157] Mukul Majhi and Arup Kumar Pal. "An image retrieval scheme based on block level hybrid dct-svd fused features". In: *Multimedia Tools and Applications* 80 (2021), pp. 7271–7312.

[158] LaiHang Yu et al. “Weber’s law based multi-level convolution correlation features for image retrieval”. In: *Multimedia Tools and Applications* 80 (2021), pp. 19157–19177.

[159] Daniel Racoceanu and Frédérique Capron. “Towards semantic-driven high-content image analysis: An operational instantiation for mitosis detection in digital histopathology”. In: *Computerized Medical Imaging and Graphics* 42 (2015), pp. 2–15.

[160] Marc Fischer et al. “Self-supervised contrastive learning with random walks for medical image segmentation with limited annotations”. In: *Computerized Medical Imaging and Graphics* 104 (2023), p. 102174.

[161] Benyamin Ghojogh et al. “Fisher Discriminant Triplet and Contrastive Losses for Training Siamese Networks”. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. 2020, pp. 1–7. DOI: 10.1109/IJCNN48605.2020.9206833.

[162] R. Hadsell, S. Chopra, and Y. LeCun. “Dimensionality Reduction by Learning an Invariant Mapping”. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*. Vol. 2. 2006, pp. 1735–1742. DOI: 10.1109/CVPR.2006.100.

[163] Ashnil Kumar et al. “Adapting content-based image retrieval techniques for the semantic annotation of medical images”. In: *Computerized Medical Imaging and Graphics* 49 (2016), pp. 37–45.

[164] Md Ziaul Hoque et al. “Retinex model based stain normalization technique for whole slide image analysis”. In: *Computerized Medical Imaging and Graphics* 90 (2021), p. 101901.

[165] Saurabh Lodha et al. “Discordance in the histopathologic diagnosis of difficult melanocytic neoplasms in the clinical setting”. In: *Journal of cutaneous pathology* 35.4 (2008), pp. 349–352.

[166] Laëtitia Launet et al. “A Self-Training Weakly-Supervised Framework for Pathologist-Like Histopathological Image Analysis”. In: *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2022, pp. 3401–3405.

[167] Ryan Zhang et al. “HistoKT: Cross Knowledge Transfer in Computational Pathology”. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 1276–1280.

[168] Andrés Mosquera-Zamudio et al. “Deep Learning for Skin Melanocytic Tumors in Whole-Slide Images: A Systematic Review”. In: *Cancers* 15.1 (2022), p. 42.

[169] Abhinav Kumar et al. “Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer”. In: *Information Sciences* 508 (2020), pp. 405–421.

[170] Ümit Budak et al. “Computer-aided diagnosis system combining FCN and Bi-LSTM model for efficient breast cancer detection from histopathological images”. In: *Applied Soft Computing* 85 (2019), p. 105765.