INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

A Comparative Study of Image Retrieval Algorithm in Medical Imaging

Yang Muhammad Putra Abdullah [a], Suraya Abu Bakar [a,*], Wan Nural Jawahir Hj Wan Yussof [b], Raseeda Hamzah [c], Rahayu A Hamid [d], Deni Satria [e]

[a] *Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Pahang, Malaysia*
[b] *Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, Terengganu, Malaysia*
[c] *College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Melaka Branch, Jasin,Campus, Melaka, Malaysia*
[d] *Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Johor, Malaysia*
[e] *Department of Information Technology, Politeknik Negeri Padang, Padang, Indonesia*
Corresponding author: [*]surayaab@umpsa.edu.my

***Abstract*—In recent times, digital environments have become more complex, and the need for secure, efficient, and reliable identification systems is growing in demand. Consequently, image retrieval has emerged as a critical area focusing on artificial intelligence and machine learning applications. Medical image retrieval has become increasingly crucial in today's healthcare field, as it involves accurate diagnostics, treatment planning, and advanced medical research. As the quantity of medical imaging data grows rapidly, the ability to efficiently and accurately retrieve relevant images from extensive datasets becomes critical. Advanced retrieval systems, such as content-based image retrieval, are imperative for managing complex data, ensuring that healthcare professionals can access the most relevant information to improve patient outcomes and advance medical knowledge. This paper compares three algorithms: Scale Invariant Feature Transform, Speeded Robust Features, and Convolutional Neural Networks in the context of two medical image datasets, ImageCLEF and Unifesp. The findings highlight the trade-offs between precision and recall for each algorithm, providing invaluable insights into selecting the most suitable algorithm for specific tasks. The study evaluates the algorithms based on precision and recall, two critical performance metrics in image retrieval.**

***Keywords*—Image retrieval; SIFT; SURF; CNN; CBIR.**

Manuscript received 5 Dec. 2023; revised 9 Apr. 2024; accepted 14 Sep. 2024. Date of publication 30 Nov. 2024.

I. INTRODUCTION

In today's digital world, image retrieval plays an essential role in managing information in a wide range of applications, especially in healthcare [1], [2] and security [3]. Image retrieval involves searching, identifying, and retrieving images from large datasets based on specific criteria, often driven by user queries or automated systems. As the volume of digital images expands exponentially, the demand for efficient and accurate image retrieval systems has grown correspondingly. Traditionally, image retrieval methods have been categorized into text-based and content-based approaches. The text-based approach dates back to the late 1970s and involves manually annotating images with text descriptors. A database management system (DBMS) then utilizes these descriptors to retrieve relevant images. However, traditional DBMS approaches that rely heavily on metadata and text annotations are being progressively replaced by more advanced methods that use content-based image retrieval (CBIR). CBIR emerged in the early 1980s to tackle the issues with text-based annotation, particularly the inaccurate nature of human perception.

In CBIR, images are indexed using both features: visual features such as colon, texture, and shapes, and text-based features, including keywords and annotations. Most identification systems apply shape or color features because of the better accuracy performance, especially in areas of fingerprint [4], [5], and object recognition [6], [7]. Image features are typically classified into global and local features. Global features encompass intensity histograms and pixel distribution, representing overall image characteristics. In contrast, local image features are more specific to image properties of local image regions, including lines, corners, and edges. While global features may struggle with certain

application requirements, such as cluttered scenes or varying object sizes, local features offer robustness against challenges like occlusion and changes in scale or rotation. This benefit was proved by using Scale Invariant Features Transform (SIFT), one of the local image features techniques, which was successfully applied to improve image registration performance [8], [9]. The combination of SIFT and Speeded Up Robust Features (SURF) also gives better performance in face recognition [10]-[12].

In recent years, Convolutional Neural Networks (CNN) have emerged as a powerful tool in the field of image retrieval, particularly in medical imaging, where they play a critical role in clinical decision-making [13] and research [14]. CNNs have revolutionized image analysis by automatically learning hierarchical features from data, enabling more accurate and efficient image retrieval. Unlike traditional methods that rely on manually extracted features, CNNs can learn complex patterns and representations directly from the images, leading to superior performance in medical image classification, segmentation, and retrieval tasks. The application of CNNs in medical image retrieval has demonstrated significant improvements in precision and recall, mainly when dealing with large and complex datasets. As medical datasets grow in complexity and size, developing robust image retrieval systems that leverage these advanced techniques becomes increasingly essential for improving patient outcomes and advancing medical research.

The ability to efficiently retrieve relevant medical images from large databases enables more accurate diagnoses, enhances treatment planning, and supports educational initiatives. Advanced techniques, such as CBIR [15], [16], leverage the unique features of medical images to facilitate precise searches across diverse modalities and conditions. As medical datasets grow in complexity and size, developing robust image retrieval systems becomes increasingly essential for improving patient outcomes and advancing medical research. This paper compares three algorithms—SIFT, SURF, and Convolutional Neural Networks (CNN)—using two medical image datasets: ImageCLEF and Unifesp.

In the medical image retrieval system, several studies have proved the effectiveness of some methods in improving the average precision and recall of medical image retrieval, particularly in scenarios involving large image datasets. Notable among these methods are the SIFT [17], SURF [18], [19] and CNN [20], [21].

A. Scale Invariant Feature Transform (SIFT)

In 2004, SIFT was introduced as a robust method for detecting and describing local image features [22]. The principal SIFT methods consist of four critical phases: scale-space extrema detection, keypoint localization, orientation assignment, and keypoint descriptor. Fig. 1 illustrates the computation of the keypoint descriptor of SIFT, which divides a 16x16 window into a 4x4 grid of cells.

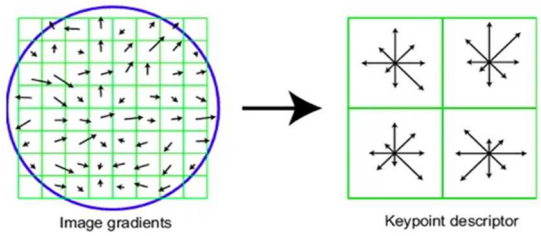


Fig. 1 SIFT key point descriptor [22]

SIFT is highlighted for its robustness in extracting distinctive features from images, even under varying conditions such as scale, rotation, and illumination. Implementing SIFT [17] in CBIR demonstrates how it can enhance the accuracy of retrieving relevant medical images from large datasets. They compare SIFT-based retrieval methods with traditional approaches, the study illustrates the superiority of SIFT in handling the complex features present in medical images. The research findings suggest that integrating SIFT into medical image retrieval systems can substantially improve retrieval performance, thus providing a valuable tool for healthcare professionals in diagnosing and planning treatments.

B. Speeded Up Robust Features (SURF)

SURF is a robust feature detection and description algorithm that excels in identifying and matching image features under various transformations, such as scale, rotation, and illumination changes. This makes SURF particularly well-suited for medical image retrieval, where precise and reliable identification of anatomical structures is critical. The advantages of using SURF have been discussed in [18], highlighting its computational efficiency and robustness. The ability to rapidly compute features without compromising accuracy is emphasized as a key factor in enhancing the performance of image retrieval systems. Reference [18] demonstrates how SURF can be integrated into content-based image retrieval (CBIR) frameworks, significantly improving the accuracy and speed of retrieving relevant medical images from large datasets.

The authors in [19] claim that their proposed technique improves retrieval accuracy when the SURF algorithm is applied for detection, description, extracting reference images, and matching feature points in the image. The proposed method was experimentally evaluated on lung images using SURF features, and the results reportedly showed better outcomes. The authors conclude that their proposed method contributes to efficient medical image retrieval.

TABLE I
COMPARISON OF RELATED WORKS FOR PRECISION-RECALL PERFORMANCE

Methods	Dataset	Precision (%)	Recall (%)
SIFT [17]	304 CT images-body parts X-ray	91	-
SURF [19]	50 lung images from Research Institute Coimbatore	92	54
ANN [23]	Breast images of 161 patients with the number of 322 images.	96	-

The performance comparison of precision and recall for different image retrieval methods applied to medical datasets is shown in Table 1. SIFT shows a precision of 91% on CT images, while SURF achieves 92% precision and 54% recall on lung images. ANN outperforms with 96% precision on breast images. These results are crucial for medical diagnostics, where image retrieval accuracy can significantly impact the identification and treatment of conditions. An improved Artificial Neural Network (ANN) model, optimized with Particle Swarm Optimization (PSO), which significantly enhances the accuracy and retrieval speed of CBIR systems, was present in [23]. The details include extracting image features such as texture and shape, using k-means clustering for feature clustering, and applying a Particle Swarm Optimization Artificial Neural Network (PSO-ANN) classifier to retrieve images related to a query image.

II. Materials and Method

The proposed flow and design for medical image retrieval aim to enhance the efficiency and accuracy of retrieving relevant images from medical datasets. By integrating a few algorithms with a structured retrieval framework, the implementation emphasizes content-based image retrieval (CBIR) techniques, leveraging robust feature extraction and matching methods to ensure precise identification of relevant images. This study uses comparative analysis as a comprehensive research design throughout the studies. Fig. 2 depicts the flow of the proposed implementation process.

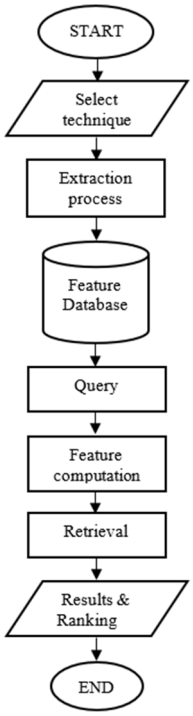


Fig. 2 Proposed implementation process

The flow begins with selecting a technique, followed by an extraction process. After the extraction phase, the features are stored in the feature database. Then, queries and feature computation processes followed by the retrieval of relevant images. The results were ranked before it was displayed. The evaluation result is analyzed and compared to conclude the best technique by using all three methods and two datasets.

All images obtained the features through the selected technique extracted from the extraction process. Subsequently, these extracted features are stored in the feature database. After that, a query was selected from a list of images from a medical image database. Features of the selected methods will be computed from the query images, followed by computing the distance measurement. Images with similar properties will be sorted and ranked accordingly before they are displayed.

III. Results and Discussions

An experimental investigation was undertaken employing three methods (SIFT, SURF, and CNN) to further investigate the competencies in retrieving medical images, with a particular focus on their precision and recall performance. In these experiments, the methods were implemented using the MATLAB programming environment. This implementation involves the use of two medical image datasets: ImageCLEF and Unifesp.

A. Medical Image Dataset

The experiments are conducted on 100 X-ray images of different body parts acquired from the CLEF 2009 dataset. Initiated in the year 2003 [24] under the aegis of the Cross-Language Evaluation Forum (CLEF), ImageCLEF was established to facilitate the assessment of several key areas. The dataset comprises 12,941 images featuring various parts of X-ray images. For testing purposes, 100 X-ray images have been selected and distributed across 5 categories: hand, hip bone, ribs, breast, and leg bone. Each category contains 20 images as shown in Fig. 3.

Fig. 3 An example of ImageCLEF dataset category

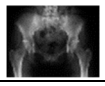
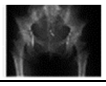
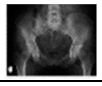
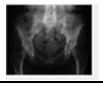
A different dataset, Unifesp [25], is used for comparison with the preceding dataset. The dataset contains 2481 X-ray images from 20 body parts. For the validation, a set of examples such as hand, skull, ribs, backbone and leg bone are selected. 100 images of various parts are selected with a total of 5 categories, and each category contains 20 images, as shown in Fig 4.

Fig. 4 An example of Unifesp dataset category

B. Retrieval Ranking

In the retrieval process, the descriptor of keypoint features for the query image by the chosen methods is compared against the features descriptor archived in the database. This comparison is important in ranking each indexed image according to its distance from the query. Only the top 10 retrieved images are presented in the retrieval ranking results table. Table 2 shows a sample of a hip bone image from ImageCLEF as the query with the outcomes of retrieval ranking of three different methods.

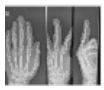
TABLE II
RETRIEVAL RANKING RESULTS OF IMAGECLEF DATASET

Methods	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8	Rank 9	Rank 10
SIFT										
SURF										
CNN										

Each method is evaluated based on its ability to accurately retrieve images of a hip bone from a database, with the results displayed in a ranked order from most to least relevant which is rank 1 until rank 10. The SIFT method demonstrates good retrieval accuracy, with the first four ranks retrieving images that are in the same category to the query but failed to retrieve the similar category in ranks 5 to 7. In rank 8 until rank 9, it retrieves the similar image category again. However, in rank 10, it again retrieved a different category. In contrast, the SURF methods maintain a high level of retrieval accuracy up to the fourth rank, but its performance slightly decreases when retrieving hand bone images at rank 5, followed by a similar category again in ranks 6 and 7. It failed to retrieve similar categories at rank 8, however, at ranks 9 and 10 it successfully retrieved images from the similar category once again. CNN demonstrated superior performance with consistent retrieving hip bone images up to the seventh rank. This indicates a robust feature recognition capability that closely aligns with the query image. However, at rank 8, the retrieval accuracy decreases as evidenced by the retrieval of a rib image. Nevertheless, the retrieval successfully retrieves the similar image category back at ranks 9 and 10. This consistency suggests that the CNN method may offer a more reliable approach for medical image retrieval tasks, particularly when the precision of the search results is paramount.

Table 3 presents a sample hand bone image from the Unifesp dataset used as the query image. The retrieval ranking table indicates that the SIFT algorithm, known for its ability to handle scale and rotation changes, perfectly retrieves images from the same category up to rank 9. However, it fails to maintain this consistency at rank 10 by retrieving an image of a leg bone instead. From the retrieval ranking results, it can be observed that SURF and CNN give the best retrieval ranking where most of similar category images are able to be retrieved in the first top 10 ranking. SURF, known for its computational efficiency, exhibits a good performance in this retrieval process. Nonetheless, the retrieval ranking table displays only the top ten results. The overall performance of each method is comprehensively presented using precision and recall graphs which give a clearer picture of the accuracy and effectiveness of each method.

TABLE III
RETRIEVAL RANKING RESULTS OF UNIFESP DATASET

Methods	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8	Rank 9	Rank 10
SIFT										
SURF										
CNN										

The evaluation results comparing the precision rates of the top twenty retrieval rankings are summarized in Table 4. Referring to the table, for the ImageCLEF dataset with hip bone as the query, the SIFT achieves a precision rate of 60%. For the Unifesp dataset with hand bone as the query, its precision significantly improves to 90%. SURF outperforms SIFT with a precision rate of 80% and achieves a perfect precision rate of 100% on the Unifesp dataset. CNN performs equally to SIFT on the ImageCLEF dataset achieving a 60% precision rate. However, on the Unifesp dataset, CNN excels, achieving a perfect precision rate of 100%. The comparison

highlights the impact of different algorithms on different datasets.

TABLE IV
EVALUATION RESULTS COMPARISON OF PRECISION RATES

Methods	Precision Rate (%)	
	ImageCLEF (hip bone)	Unifesp (hand bone)
SIFT	60	90
SURF	80	100
CNN	60	100

C. Precision and Recall Graph

In the domains of image retrieval and classification, precision and recall are essential metrics that offer insights into the performance of algorithms. Precision, defined as the ratio of true positives to the sum of true and false positives, reflects the algorithm's ability to return only relevant instances. Recall, or the ratio of true positives to the sum of true positives and false negatives, measures the algorithm's capacity to identify all relevant instances. The precision and recall rates are defined in Equation 1 and Equation 2, respectively.

$$Precision\ rate = \frac{number\ of\ relevant\ image\ selected}{total\ number\ of\ retrieved\ images} \quad (1)$$

$$Recall\ rate = \frac{number\ of\ relevant\ images\ selected}{total\ number\ of\ similar\ images\ in\ the\ database} \quad (2)$$

A comparative analysis of precision and recall metrics for three algorithms, SIFT, SURF, and CNN, was conducted by using the medical image dataset. This analysis provides a detailed evaluation of each method's accuracy and effectiveness in medical image retrieval. The blue line represents the SIFT algorithm, while the red line depicts the SURF algorithm, and the yellow line marks the CNN. In the graph, the x-axis represents recall, and the y-axis represents precision, both ranging from 0.0 to 1.0.

1) Results for ImageCLEF Dataset: Fig. 5 depicts a precision and recall graph of SIFT, SURF, and CNN by using the ImageCLEF dataset. From the graph, it can be observed that the SURF algorithm demonstrates the best retrieval results compared to SIFT and CNN. It proved that SURF capabilities more robust feature extraction to provide keypoint descriptors, thus leading to a higher probability of image matching and retrieval. The SIFT algorithm shows a more gradual decline, indicating a more balanced approach between precision and recall. At the beginning of retrieval, the CNN approach appears to maintain a higher retrieval, but it slightly decreased along with the decreasing number of retrieved images.

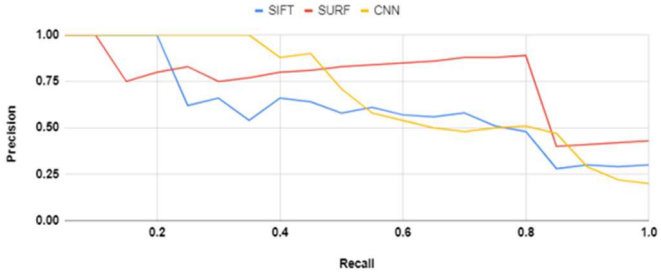


Fig. 5 The precision and recall graph performance using ImageCLEF

The graph analysis provides an understanding of the performance of SIFT, SURF, and CNN algorithms within the context of the ImageCLEF dataset. It highlights the importance of considering both precision and recall in evaluating and selecting algorithms for image retrieval systems. This study underscores the importance of algorithm selection based on specific image retrieval system requirements, balancing computational cost against performance metrics.

2) Results for Unifesp Dataset: Fig. 6 illustrates the precision and recall graph of the performance for three algorithms using Unifesp dataset. The graph shown in Fig. 6 indicates that SIFT gives the best retrieval results compared to SURF and CNN. Although SURF maintains good retrieval at the beginning, the performance decreases slightly until the end of retrieval. The graph indicates that the CNN algorithm initially demonstrates robust retrieval at the beginning. However, its performance drops in the middle of retrieval and continues to decline until the end of the retrieval process. This suggests that while CNN is initially accurate in its predictions, it becomes less precise as it tries to cover more positive instances. In contrast, SIFT shows an increasing recall, indicating a more stable performance across different levels of recall. The graph thus serves as an empirical foundation for decision-making in the selection of algorithms for image classification tasks, particularly within the context of the Unifesp dataset. It highlights the importance of evaluating both precision and recall in determining the efficacy of an algorithm, and it underscores the need for a balanced approach that considers the specific requirements and constraints of the application.

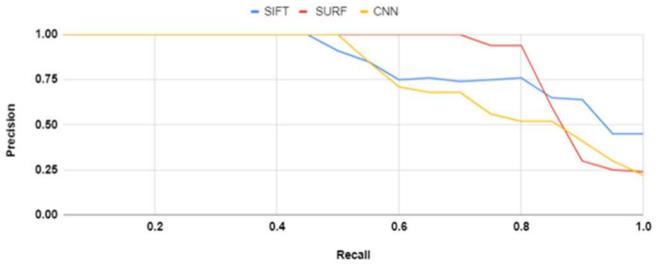


Fig. 6 The precision and recall graph performance using Unifesp

IV. CONCLUSION

This paper has thoroughly investigated and successfully explored using SIFT, SURF, and CNN methods in medical image retrieval by comparing their performance with that of different medical image datasets. The comparative analysis provides a helpful guide for researchers and practitioners, showing how each method performs and helping in choosing the best one for the unique challenges of medical image retrieval. The ongoing development of these systems promises to enhance data management efficiency and expand the potential applications of image retrieval. The results highlight the trade-offs between these metrics for each algorithm, providing insights for selecting the most appropriate algorithm for specific tasks. The results discuss an estimate of the quality work that SIFT, SURF, and CNN can perform for medical image retrieval. The analysis of the data is where their statistical methods for comparing performance are presented, which use measures including

precision and recall. In conclusion, the comparative analysis of SIFT, SURF, and CNN emphasizes the importance of selecting the methods and the need for robust image retrieval systems to manage growing medical datasets effectively and improve patient outcomes.

ACKNOWLEDGMENT

The authors thank Universiti Malaysia Pahang Al-Sultan Abdullah for providing the facilities and support for this research.

REFERENCES

[1] H. Chugh, S. Gupta, M. Garg, D. Gupta, S. Juneja, H. Turabieh, Y. Na and Z. K. Bitsue, "Image Retrieval Using Different Distance Methods and Color Difference Histogram Descriptor for Human Healthcare", Journal of Healthcare Engineering, 2022.

[2] A. Gain, "Optimization of CNN for Content-Based Image Retrieval in Healthcare", Chapman and Hall/CRC, 2024.

[3] W. Lu, A. L. Varna, A. Swaminathan and M. Wu, "Secure image retrieval through feature protection", 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 2009

[4] H. M. Ali and M. I. Islam, "Preliminary Identification of Fingerprint based on Shape Features", International Journal of Computer Applications, vol. 120, no. 15, 2015

[5] E. J. Baker, S. A. Alazawi, N. T. Ahmed, M. A. Ismail, R. Hassan, S. A. Halim and T. Sutikno, "User identification system for inked fingerprint pattern based on central moments", Indonesian Journal of Electrical Engineering and Computer Science, vol. 24, no. 2, 2021

[6] A. Diplaros, T. Gevers and I. Patras, "Color-Shape Context for Object Recognition", Journal of Photometric Methods in Computer Vision, vol. 1, no. 11, 2002.

[7] F. S. Khan, J. Van de Weijer and M. Vanrell "Modulating Shape Features by Color Attention for Object Recognition", International Journal of Computer Vision, vol. 98. pp. 49-64, 2011.

[8] L. Jinxia and Q. Yuehong, "Application of SIFT feature extraction algorithm on the image registration", IEEE 2011 10th International Conference on Electronic Measurement & Instruments, 2011.

[9] P. M. Panchal, S. R. Panchal and S. K. Shah, "A Comparison of SIFT and SURF", International Journal of Innovative Research in Computer and Communication Engineering, vol. 1, 2013

[10] A. Vinay, D. Hebbar, V. S. Shekhar and K. N. B. Murthy, "Two Novel Detector-Descriptor Based Approaches for Face Recognition Using SIFT and SURF", Procedia Computer Science, vol. 70, 2015.

[11] L. Lenc and P. Kral, "A combined SIFT/SURF descriptor for automatic face recognition", Conference on Machine Vision (ICMV), 2013

[12] S. Gupta, K. Thakur and M. Kumar, "2D-human face recognition using SIFT and SURF descriptors of face's feature regions", The Visual Computer, 2021.

[13] S. R. Qamar, D. Evans, B. Gibney, G. E. Redmond, M. U. Nasir, K. Wong and S. Nicolaou, "Emergent comprehensive imaging of the major trauma patient: a new paradigm for improved clinical decision-making", Canadian Association of Radiologists Journal, vol. 72, 2021.

[14] C. G. Sotomayor, M. Mendoza, V. Castaneda, H. Farias, G. Molina, G. Pereira, S. Hartel, M. Solar and M. Araya, "Content-Based Medical Image Retrieval and Intelligent Interactive Visual Browser for Medical Education, Research and Care", vol. 11, 2021

[15] M. Kashif, G. Raja and F. Shaukat, "An Efficient Content-Based Image Retrieval System for the Diagnosis of Lung Diseases", Journal of Digital Imaging, vol. 33, pp. 971-987, 2020.

[16] V. Majhi and S. Paul, "Application of Content-Based Image Retrieval in Medical Image Acquisition", Challenges and Applications for Implementing Machine Learning in Computer Vision, 2020.

[17] L. J. Zhi, S. M. Zhang, D. Z. Zhao, H. Zhao and S. Lin, "Medical image retrieval using SIFT feature", 2009 2nd International congress on image and signal processing, 2009.

[18] E. P. Yudha, N. Suciati and C. Fatichah, "Preprocessing Analysis on Medical Image Retrieval Using One-to-one Matching of SURF Keypoints", 2021 5th International Conference on Informatics and Computational Science (ICICoS), 2021.

[19] S. Govindaraju and G. P. Ramesh Kumar, "A Novel Content Based Medical Image Retrieval using SURF Features", Indian Journal of Science and Technology, vol. 9, 2016.

[20] H. Hu, W. Zheng, X. Zhang, X. Zhang, J. Liu, W. Hu, H. Duan and J. Si, "Content-based gastric image retrieval using convolutional neural networks", International Journal of Imaging Systems and Technology, 2020.

[21] D. Barac, T. Manojlovic, M. Napravnik, F. Hrzic, M. M. Saracevic, D. Miletic and I. Stajduhar, "Content-Based Medical Image Retrieval for Medical Radiology Images", Lecture Notes in Computer Science, vol. 14845, 2024.

[22] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", International Journal on Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.

[23] R. B. Kumar and P. Marikkannu, "An Efficient Content Based Image Retrieval using an Optimized Neural Network for Medical Application," Multimedia Tools Applicatios, vol. 79, no. 31–32, pp. 22277–22292, 2020.

[24] Cross Language Evaluation Forum (CLEF), (2003), [Online]. Available: https://www.imageclef.org/

[25] Unifesp, [online]. Available: https://www.kaggle.com/datasets/felipekitamura/unifesp-xray-bodypart-classification