I) Introduction:

We are highly familiar with terms like AI, Machine Learning, and Deep Learning, but another key element—a familiar yet powerful tool supporting AI—is Computer Vision. Simply put, Computer Vision mimics human abilities by recognizing features of objects from images, texts, videos, etc., to classify and identify whether the input is the intended object. Additionally, Computer Vision can segment specific regions within an image and provide detailed information as required. For example, when you need a security check at the airport, you provide your ID to the security staff. They scan the photo on your passport and compare it with your actual face captured by the scanner at the counter. The staff simply verifies the percentage of similarity shown by the system and visually inspects the match. This process ensures both subjective and objective accuracy. This illustrates that Computer Vision is already embedded in our daily lives. However, the most advanced applications of Computer Vision today are found in healthcare. It is used to detect diseases, locate affected cells, and enable non-invasive treatments to avoid impacting other parts of the body, ultimately aiming to help patients recover as quickly as possible. This analysis explores how Computer Vision and Machine Learning diagnose medical conditions through imaging, examines current limitations, and discusses future solutions.

II) What is Computer vision ?

Computer Vision is an application developed from Artificial Intelligence (AI), in which its function resembles human vision. When static images or dynamic content like videos are scanned by Computer Vision, the AI system analyzes, identifies features, and provides results such as classification or object recognition [1]. This helps minimize subjective evaluations and enhances the accuracy of predictions by being able to analyze a vast amount of data efficiently.

III) Comparison of SIFT, SURF, and CNN – Methods for Image Analysis and Feature Recognition in Computer Vision

SIFT and SURF [2,3] are traditional approaches developed in the early 20th century. Both are manual feature extraction methods, meaning engineers must manually identify features in images such as keypoints, orientation, brightness, etc. They then design a process and apply that formula to subsequent images. This requires knowledge of mathematics and computer vision theory. The primary task of SIFT and SURF is to identify invariant features of an image. In other words, after identifying key features, they calculate metrics that allow the image to be rotated,

brightened, tilted, etc., without losing the distinctiveness of the object. This ensures that if a similar image appears in the future, even with variations in tilt, lighting, etc., the model can still classify and recognize the object using the pre-defined formula. However, the significant challenge with these methods is their lack of flexibility in handling highly variable images. They can struggle with noisy data, glare, and other distortions, limiting their performance in diverse scenarios.

As science continues to advance, so does the demand for processing large volumes of data while minimizing labor costs. The year 2012 marked the rise of CNN [4] (Convolutional Neural Network), a type of neural network specifically designed to automatically learn features directly from image data. It uses convolutional layers to filter and detect features. At each layer, the network refines its understanding of the image using filters such as 3x3 or 4x4 pixels, allowing it to capture features ranging from low-level (edges, textures) to high-level (shapes, patterns) without human intervention. Raw images are fed directly into the CNN, which then autonomously learns to extract relevant features for tasks such as classification, recognition, or segmentation. As a result, CNNs are now widely applied across various industries, including delivery robots, self-driving cars, facial recognition at airports, and, notably, robotic-assisted surgeries in healthcare.

Machine Learning and Deep Learning Models Used in Computer Vision. As discussed earlier regarding manual feature extraction methods like SIFT and SURF, once the feature recognition process is complete, the extracted data is fed into Machine Learning models such as Support Vector Machines (SVM) [5], K-Nearest Neighbors (KNN), and Random Forest. These models use the extracted features to classify and recognize objects. However, these models face challenges when dealing with complex data and require multiple data preprocessing steps. One key limitation is that Machine Learning models rely on supervised learning, meaning the input images must be labeled, and each label must correspond to the manually extracted features. For these models to perform effectively, a large amount of labeled data is needed, which is not always feasible in practice, especially for complex tasks or datasets that lack diversity. Additionally, traditional Machine Learning methods cannot learn directly from raw data, requiring intricate preprocessing steps to optimize model performance. In contrast, Deep Learning models such as Convolutional Neural Networks (CNNs) can automatically learn and extract features from images without requiring manual preprocessing. CNNs [6] have the ability to learn from large datasets and adjust parameters automatically to improve accuracy without the need for detailed labeling of each object in the image. This significantly

saves time and effort while delivering superior performance in complex tasks, such as facial recognition under changing lighting conditions or object detection in noisy environments. Therefore, while traditional Machine Learning models like SVM and KNN are still useful for simpler, well-labeled datasets, Deep Learning models like CNNs dominate in handling complex Computer Vision problems that demand high accuracy.

IV) Computer Vision tasks

The two main tasks of Computer Vision are object localization and object recognition. It determines the location of an object using a bounding box and recognizes the object by classifying it according to predefined labels. One of the popular techniques to accomplish these two tasks is YOLO (You Only Look Once) [7], which has been known since 2012 and has become a widely used technique due to its high applicability.

A practical example of applying YOLO is detecting skin lesions through images and predicting whether they are malignant moles or skin cancer. First, the program will be provided with a supervised learning dataset called ISIC, which contains labeled images of skin lesions, each marked with a bounding box around the affected skin area. Then, each input image is transformed into a vector in the form of:

$$\begin{pmatrix} P_c \\ b_x \\ b_y \\ w \\ h \\ C_1 \\ C_2 \end{pmatrix}$$ Such that:

b_x : Determine the center x of the bounding box on the x-axis.

b_y: Determine the center y of the bounding box on the y-axis.

w is width of the bounding box

h is height of the boudning box

c_1, c_2: the binary classification of the object (malignant/benign) respectively.

Next, after training with the vectors provided, YOLO [8] will divide the image into grids, with each grid corresponding to a bounding box. YOLO will then retain only the bounding boxes with the highest confidence and eliminate duplicate boxes.

Afterward, when the model is given a test dataset with completely new images, YOLO will return the result as a vector with P_c = 0 if there is no object or P_c = 1 if there is an object in the grid, C_1 = 0 if it is benign and C_1 = 1 if malignant.

On the other hand, accurately determining the location and recognizing the object with higher detail is achieved using the Fully Convolutional Neural Network (FCN) method. While the goal of YOLO is to locate and classify objects in an image by predicting bounding boxes, FCN will precisely segment the edges of a skin lesion. In other words, the lesions will be accurately colorized according to the shape of the lesion, helping doctors easily determine the size and shape of the damage.

V) Data:

6.1. Medical Dataset:

Today, with the support of large volumes of images and videos, a major issue of data scarcity has been addressed, especially in the healthcare field (which will be discussed in the challenges section). Datasets such as the ADNI dataset [9], which contains over 300 MRI images of brain structure and functional brain activities, are extremely important in analyzing structural changes in the brain of Alzheimer's patients – a form of dementia commonly seen in the elderly. Another example is the MURA dataset [10], which is publicly available and contains over 40,000 X-ray images of musculoskeletal conditions from 14,863 studies, representing 11,184 patients. In predicting lung cancer from the ACS dataset at ASAN Hospital in South Korea, there are 9,792 chest X-ray images, and a study using the ACS dataset applied the YOLO model to detect abnormalities in lung images with an accuracy of over 93% [11]. Additionally, there are numerous studies on cancer analysis and chest image analysis. Most of these analyses have been collected from various imaging methods, including CT, MRI, ultrasound, and X-ray.

6.2. Data challenges:

It is well understood that Computer Vision operates based on input images or videos. A significant advantage of modern deep learning models is their ability to autonomously learn and identify features, even from raw, unprocessed images. However, the accuracy of these models is largely contingent on the diversity of the input data, necessitating a substantial volume of data to enable the model to learn in a generalized and robust manner.

6.3. Solutions:

The first solution to address the issue of limited medical data, due to regulations such as HIPAA in the U.S. or GDPR in Europe that prohibit sharing patient information outside of hospitals, involves federated learning. In this approach, hospitals build their own AI models using patient records and medical data available at their facilities. They then send the model parameters—not the raw data—to a central server. The server aggregates these parameters from multiple hospitals to create a shared global model. This ensures that raw data never leaves the hospital premises. Once the global model is created, it is sent back to the hospitals for further development and application. The second solution involves leveraging pre-trained models for deep learning tasks. For instance, instead of training a model from scratch to recognize specific features like skin lesions, a pre-trained model (such as one trained on general image datasets) can be fine-tuned for the specific task. By making slight adjustments at the fully connected layer (fine-tuning), the model can learn and improve its accuracy for the new task, such as classifying skin lesions, without starting from zero. This significantly reduces both time and memory requirements. This process is referred to as Transfer Learning [12]. The third solution is Data Augmentation, which involves artificially expanding the dataset by applying transformations such as rotation, zooming in and out, cropping, or adding noise to existing images. This enhances the diversity of the dataset without requiring additional real-world data collection.

VI) Application:

Pathologists play a critical role in cancer detection and treatment, examining tumor cells under a microscope. However, visual analysis by the human eye introduces subjectivity, often resulting in varying diagnostic outcomes. Since the advent of ultra-high-resolution tissue scanners combined with advancements in computer vision, the efficiency and accuracy of routine diagnostic tasks have significantly improved. These technologies enable the discovery of new disease markers and therapeutic indicators from morphological structures that are imperceptible to the naked eye, highlighting one of the key applications of computer vision in healthcare.

Another application of Computer Vision (CV) and Deep Learning (DL) lies in surgery and endoscopy, where they enhance procedural efficiency and treatment outcomes. CV can assist surgeons by providing real-time contextual information during procedures. For instance, an AI system can recognize surgical tools or analyze the steps of a procedure, thereby helping surgeons avoid errors or manage

complex situations more effectively. Additionally, Computer Vision can evaluate and score surgeons' skills based on standardized criteria. One example is GOALS (Global Operative Assessment of Laparoscopic Skills), a framework for assessing laparoscopic surgical skills using the following metrics:

+ Precision in handling instruments.

+ Speed and efficiency.

+ Control of movements and execution time.

By analyzing the movement of surgical instruments and the behavior of surgeons, AI can offer detailed feedback to improve skills or assess competency during training.

In practical applications, during the COVID-19 pandemic, thermography devices were deployed in public spaces to detect individuals with elevated body temperatures (fever), enabling early screening and disease detection [14]. To improve accuracy in body temperature measurement, researchers found that the temperature emitted from the inner corner of the eye is more reliable than that from the forehead. Computer Vision techniques are utilized to analyze images of the inner eye area, providing precise evaluations of body temperature.

VII) Challenges for ai in the healthcare industry:

Despite continuous improvements and the adoption of new methods, Computer Vision, and particularly AI, still face significant challenges. In the healthcare field, trust in AI-generated results is a crucial factor for acceptance. The foundation of clinical trust primarily stems from rigorous prospective trials, where AI algorithms are validated in real clinical environments. Additionally, building patient trust—particularly concerning privacy concerns—is essential. An important focus lies in establishing next-generation regulations that align with advancements in privacy-preserving technologies. To extract insights from such sensitive data, further development of security techniques, such as federated learning and federated analytics, is necessary.

VIII) Solutions:

To address challenges in the field of computer vision (CV), particularly in medical image analysis, various solutions have been proposed. First, data quality can be enhanced through data augmentation tools such as light adjustment, noise reduction, or the use of Generative Adversarial Networks (GANs) to generate new images from

existing data. Deep learning algorithms also need to be developed to work effectively with small datasets through transfer learning or few-shot learning. Additionally, data annotation can be automated using AI systems, combining semi-supervised learning and learning from partially labeled data to reduce manual workload. To tackle class imbalance issues, techniques like oversampling or undersampling are employed, alongside weighted loss functions and transfer learning.

Furthermore, the development of interpretable models and the application of post-hoc explanation algorithms like SHAP or LIME [15] enhance trust among physicians and researchers. To improve generalization capabilities, multi-task learning, continual learning, and training on diverse datasets are implemented, enabling models to adapt to various conditions. Lastly, interdisciplinary collaboration between AI and healthcare experts not only ensures that models are appropriately designed but also helps standardize workflows, fostering the comprehensive and sustainable advancement of AI applications in healthcare.

IX) References:

[1] Gonzalez RC, Woods RE. Digital image processing (3rd Edition). USA: Prentice-Hall, Inc.; 2006

[2] Lowe DG. Distinctive image features from scale-invariant keypoints. Inter J Comput Vis 2004 Nov; 60:91–110

[3] Bay H, Tuytelaars T, Gool LV. SURF: Speeded up robust features. In: Computer Vision – ECCV 2006. Springer Berlin Heidelberg; 2006. pp. 404–17

[4] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016.

[5] Hearst M, Dumais S, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Their Appl 1998;13:18-28

[6] Schwab E, Gooßen A, Deshpande H, Saalbach A. Localization of critical findings in chest X-ray without local annotations using multi-instance learning; 2020

[7] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. pp. 779–88

[8] Elyan E, Vuttipittayamongkol P, Johnston P, Martin K, McPherson K, Moreno-García CF, Jayne C, Sarker MMK. Computer vision and machine learning for

medical image analysis: recent advances, challenges, and way forward. Art Int Surg 2022. Pp. 29-30.

[9] Elyan E, Vuttipittayamongkol P, Johnston P, Martin K, McPherson K, Moreno-García CF, Jayne C, Sarker MMK. Computer vision and machine learning for medical image analysis: recent advances, challenges, and way forward. Art Int Surg 2022. Pp. 33.

[10] Rajpurkar P, Irvin J, Bagul A, et al. MURA: Large dataset for abnormality detection in musculoskeletal radiographs; 2018.

[11] Liu X, Dong S, An M, Bai L, Luan J. Quantitative assessment of facial paralysis using infrared thermal imaging. In: 2015 8th International Conference on Biomedical Engineering and Informatics (BMEI). IEEE; 2015. pp. 106–10

[12] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010;22:1345-59

[13] Vassiliou, M. C. et al. A global assessment tool for evaluation of intraoperative

laparoscopic skills. Am. J. Surg. 190, 107–113 (2005)

[14] Ismael AM, Şengür A. Deep learning approaches for COVID-19 detection based on chest X-ray images. Expert Syst Appl 2021;164:114054

[15] Aldughayfiq, B.; Ashfaq, F.;Jhanjhi, N.Z.; Humayun, M. Explainable AI for Retinoblastoma Diagnosis: Interpreting Deep Learning Models with LIME and SHAP. Diagnostics 2023, 13, 1932