I.J. Image, Graphics and Signal Processing, **2025, 6, 38-59**
Published Online on December 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2025.06.03

Application of Large Language Models for Data-Driven Analytics in Oncology: Insights and Evidence Generation from Real-World Imaging Data

Shobhit Shrotriya*
Department of Computer Science, CHRIST (Deemed to be University), Bangalore - 560029 and Accenture, India
E-mail: shobhit.shrotriya@accenture.com/shobhit.shrotriya@res.christuniversity.in
ORCID iD: https://orcid.org/0000-0002-4430-8440
*Corresponding Author

Nizar Banu P. K.
Department of Computer Science, CHRIST (Deemed to be University), Bangalore Central Campus, Hosur Road, Near Dairy Circle, Bangalore – 560029, India
E-mail: nizar.banu@christuniversity.in
ORCID iD: https://orcid.org/0000-0001-7213-4276

Avi Kulkarni
ThoughtSphere, 99 S Almaden Blvd, San Jose, CA 95113, United States
E-mail: avi.kulkarni@thoughtsphere.com
ORCID iD: https://orcid.org/0009-0009-3818-7660

Vinod G. Kumar
Accenture, Prestige Technopolis, 8/1 Dr M. H. Marigowda Road, Adugodi Rd., Bangalore-560029, India
E-mail: vinod.g.kumar@accenture.com
ORCID iD: https://orcid.org/0009-0008-8140-204X

Received: 05 January, 2025; Revised: 24 April, 2025; Accepted: 20 July, 2025; Published: 08 December, 2025

Abstract: Breast cancer is one of the most common and serious types of cancer. It can affect people of all ages and genders around the world. The increasing incidence of breast cancer, coupled with its complexity, has placed a significant burden on healthcare systems and patients alike. Traditional diagnostic methods, while effective, often face limitations in early detection and accurate prognosis, which are critical for improving patient outcomes. In recent years, artificial intelligence (AI) and machine learning (ML) are changing the way we solve problems and make decisions in the field of medical diagnostics, enhancing the ability to detect, diagnose and predict breast cancer. However, there are still challenges, such as the need for large and diverse datasets to train these models, making AI tools work smoothly in hospitals, and addressing ethical concerns in healthcare. This paper looks at how AI and ML are used in breast cancer care, especially in analyzing real-world medical data like images, histopathology, and other datasets such as doctor notes & discharge summaries, to identify patterns that may be unnoticeable to medical experts. Large Language Models (LLMs) using embeddings, are highlighted for their capacity to improve the accuracy of image related interpretations, potentially detect early-stage tumours, and predict disease progression and treatment responses. Real-world medical datasets have been collected and analysed using different models. A publicly available Convolutional Neural Network (CNN) and a custom-built Large Language Model (LLM) with embeddings were tested. The Generative AI model achieved 98.44% accuracy, significantly higher than the traditional AI model's 61.72%. Future research can explore how Generative AI can help classify patients based on risk levels. This could lead to personalized treatment plans, reducing unnecessary treatments and improving patients' quality of life. Given the research is primarily focussed on breast cancer, there is an attempt to showcase that by harnessing the power of AI and ML, there is potential to significantly reduce the global burden of breast cancer, offering new avenues for early detection, accurate diagnosis, and tailored therapeutic strategies. Continued research and collaboration among oncologists, data scientists, and policymakers are essential to fully realize the benefits of AI in the fight against breast cancer, ultimately leading to better patient outcomes and a decrease in breast cancer-related mortality.

Index Terms: Machine Learning, Artificial Intelligence, Generative AI, Large Language Models, Embeddings, Breast Cancer

1. Introduction

1.1. Background

Breast cancer is a growing health concern worldwide, with cases increasing over the years. In India, it is now the most common cancer among women and a major health challenge. According to the Global Cancer Observatory (IARC-WHO, 2022), breast cancer was the most common cancer globally, causing 665,255 deaths among women in 2022. India had the highest number of estimated breast cancer deaths that year, with 98,337 cases. As per the National Centre for Disease Informatics and Research (NCDIR) – National Cancer Registry Programme (NCRP), during 2019-2023, the estimated number of incidences are more than 1 million and the estimated number of mortalities of breast cancer have been approximately 400,00 in the country, State/UT. The disease affects women across all age groups, though there is a noticeable shift towards younger women being diagnosed, with a significant number of cases occurring in women below the age of 50. Additionally, while breast cancer is predominantly a female disease, men are also susceptible, albeit at lower rates, further emphasizing the need for inclusive and comprehensive approaches to its management [1, 2, 3, 4].

The burden of breast cancer in India is exacerbated by several factors, including late-stage diagnosis, lack of awareness, limited access to quality healthcare, and disparities in the availability of advanced diagnostic tools. The five-year survival rate for breast cancer in India remains lower than in high-income countries, largely due to the high proportion of cases diagnosed at an advanced stage. The challenges in early detection and diagnosis are compounded by the heterogeneous nature of breast cancer, which presents with various subtypes, each with distinct biological behaviours and prognostic outcomes. This complexity necessitates a more nuanced approach to breast cancer management, one that can adapt to the diverse presentation of the disease across the population.

In recent years, artificial intelligence (AI) has garnered significant attention as a promising solution to these challenges. Generative AI, particularly through Large Language Models (LLMs), has shown potential in enhancing the accuracy and efficiency of breast cancer detection, diagnosis, and prediction [6]. Research in AI applications for breast cancer has primarily focused on improving the interpretation of images, automating histopathological analysis, and predicting treatment outcomes based on real-world data (RWD). Studies have demonstrated that AI models can achieve comparable or even superior performance to human experts in detecting breast cancer, especially in cases where the disease is subtle or atypical. Furthermore, AI-driven tools have the capability to integrate and analyse large-scale datasets, encompassing imaging, clinical, and genomic information, thereby facilitating personalized medicine approaches that can tailor treatment to the individual patient's profile [7].

1.2. Rationale and Knowledge Gap

Although AI has made great progress, its use in healthcare still faces challenges. These include the need for large and diverse datasets to train accurate models, concerns about how AI makes decisions, and the need for clear regulations to ensure AI is used safely and ethically. However, research and pilot projects around the world are working to solve these issues, aiming to fully use AI to help fight breast cancer [5].

Over time, machine learning has become a key tool in breast cancer detection and diagnosis. It can process large amounts of data and find patterns that doctors might not easily notice. Some commonly used machine learning techniques for detecting breast cancer are discussed below [8, 9, 10]-

Supervised Learning Algorithms:

- Support Vector Machines (SVM): SVM is a classification algorithm that separates data points into different classes using a hyperplane. SVM has been used for breast cancer classification based on features extracted from medical images or patient data.
- Random Forest: As an ensemble learning method, Random Forest brings together multiple decision trees to make predictions. It has been applied to breast cancer detection tasks, such as classifying benign and malignant tumours based on imaging or genomic data.
- Logistic Regression: Logistic Regression is a simple yet effective classification algorithm that models the probability of a binary outcome. It has been used in breast cancer research to predict the likelihood of malignancy based on clinical and histopathological features.

Deep Learning Techniques:

- Convolutional Neural Networks (CNNs): CNNs are deep learning models specifically designed for processing structured grids of data, such as images. CNNs have shown promising results in breast cancer detection and

classification tasks using mammography and histopathology images.

- Recurrent Neural Networks (RNNs): RNNs are well-suited for sequential data processing and have been used in breast cancer research for analysing time-series data, such as gene expression profiles or longitudinal patient data.
- Autoencoders: Autoencoders are neural network architectures used for unsupervised feature learning and data compression. They have been applied to breast cancer detection tasks for feature extraction from medical images or genomic data.

Unsupervised Learning Techniques:

- Clustering Algorithms: Clustering algorithms, such as K-means or hierarchical clustering, can identify natural groupings or clusters within breast cancer datasets based on similarities in patient characteristics, tumour features, or gene expression profiles. Clustering can help identify subtypes of breast cancer or stratify patients based on disease prognosis.
- Dimensionality Reduction Techniques: Dimensionality reduction techniques, such as Principal Component Analysis (PCA) or t-distributed Stochastic Neighbour Embedding (t-SNE), can reduce the complexity of breast cancer datasets while preserving important information. These techniques are commonly used for visualization, feature selection, or data preprocessing in machine learning tasks.

Hybrid Approaches:

- Ensemble Learning: Ensemble learning methods combine multiple machine learning models to improve predictive performance. Techniques such as bagging, boosting, or stacking can be applied to breast cancer detection tasks to enhance classification accuracy and robustness.
- Transfer Learning: Transfer learning leverages pre-trained deep learning models on large datasets and fine-tunes them for specific tasks with limited data. Transfer learning has been applied to breast cancer detection tasks, particularly in scenarios where labelled data is scarce or expensive to obtain.

All the discussed machine learning techniques can be used for variety of data types such as medical images (e.g., mammography, MRI, PET scans, ultrasound), genomic data (e.g., DNA sequencing, gene expression profiles), clinical data (e.g., medical history, patient demographics), and pathology data (e.g., tissue microarrays, histopathology images). By making use of these techniques, researchers and clinicians can develop more personalized approaches that are sophisticated, consistent, accurate and efficient for breast cancer detection, diagnosis, and treatment.

CNNs are widely used for image classification because they are great at detecting patterns in images. They work by applying filters to images in multiple layers, helping to identify edges, textures, and other important features [11]. These features are then processed through pooling layers, which reduce the image size while keeping the most useful information. The model learns by training on large, labeled datasets, adjusting its filters over time to improve accuracy. CNNs automatically learn features in a structured way. For example, early layers detect basic patterns like edges and shapes whereas, deeper layers capture more complex patterns, like tumor textures in medical images. To make CNNs more flexible, techniques like image rotation, flipping, and scaling are used to increase the variety of training images. Despite their advantages, CNNs also have several challenges [12,13]-

i. CNNs require huge datasets to perform well. If there isn't enough labeled data, they may struggle to recognize new cases and perform poorly on unseen images (overfitting).
ii. Training CNNs requires powerful GPUs and large memory, which can be expensive and difficult for smaller organizations to afford.
iii. When datasets are small or lack diversity, CNNs may memorize training examples instead of learning general patterns. This reduces their ability to work on new, real-world images.
iv. Building and tuning CNNs isn't straightforward—it often requires a lot of trial and error to find the best architecture for a given task.
v. As CNN models grow, they become harder to scale for new tasks or larger datasets. Training larger models requires more data, storage, and computing power.

1.3. Objective

Breast cancer detection presents several unique challenges, which make it difficult for traditional image analysis methods like CNNs to achieve high accuracy. There is high variability in tumor appearances, they may have irregular shapes, varying contrast, and different densities across imaging modalities (mammograms, ultrasound, MRI). In dense breasts, normal fibroglandular tissue can obscure tumors, reducing sensitivity in mammography. Early-stage cancers may lack distinct visual markers, making them hard to detect using purely pixel-based analysis. Also, malignant cases are much rarer than benign or normal cases, making deep learning models biased toward benign findings. Given these challenges, LLMs offer a range of advantages. LLM-based embeddings capture semantic and high-level contextual

features, which CNNs may overlook. For example, instead of just analyzing pixel-level patterns, an LLM-enhanced model can understand that a mass with architectural distortion is more likely malignant. LLMs adopt a robust similarity search for diagnosis using embeddings which enable efficient retrieval of similar cases for comparative analysis. For example, if a new scan is ambiguous, the model can find past cases with similar features and provide relevant diagnostic outcomes.

Given the challenges associated with breast cancer detection and limitations of traditional machine learning techniques, a differentiated and novel approach is deemed necessary. The proposed approach outlines a sophisticated innovative system for detecting malignancy in medical images using advanced Generative AI techniques. The core of the approach involves leveraging Generative AI embeddings to transform images into high-dimensional vectors that capture rich contextual meanings and semantic relationships [15]. The approach showcases the transformative potential of Generative AI in improving malignancy detection through sophisticated image analysis techniques. This can empower medical professionals with enhanced diagnostic capabilities and actionable insights. The approach not only highlights the application of Generative AI in healthcare but also underscores its role in advancing precision medicine and patient care.

2. Methods

In this research, an advanced image classification system based on a Large Language Model (LLM) has been developed to analyze real-world data (RWD) of medical scanned images (see Fig.1.). The system works by - converting images into vector embeddings, storing these embeddings in a vector database and then classifying new images based on similarity metrics. To make results easier to understand, a cluster plot is used, providing a visual representation of the classification process.

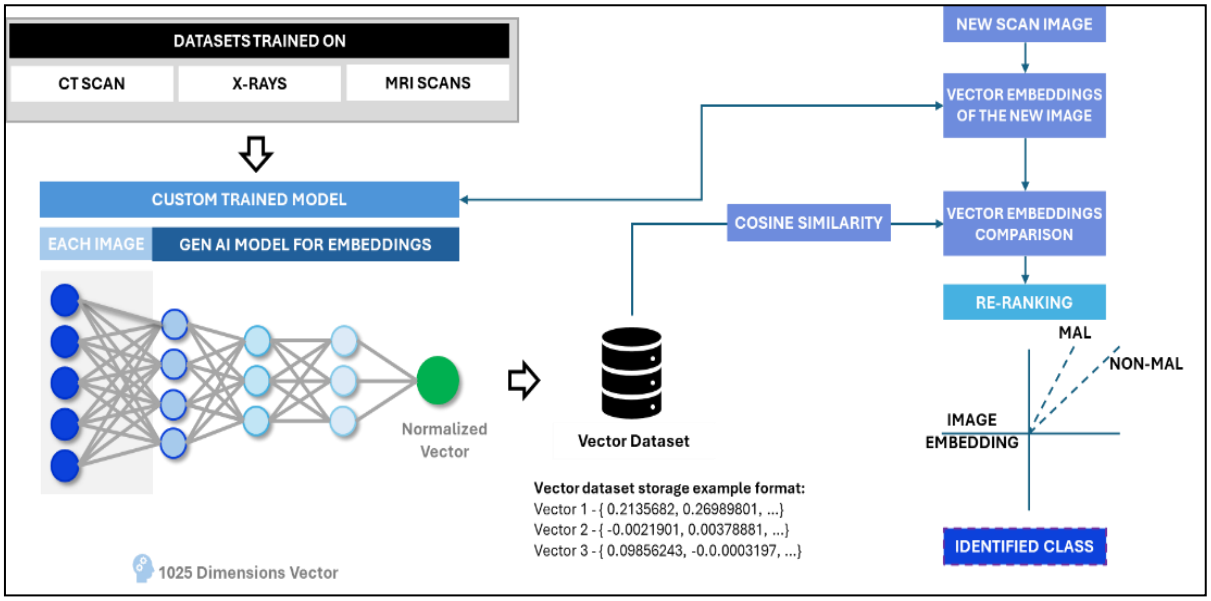

Fig. 1. GenAI Architecture for Medical Image Dataset Analysis

LLMs create embeddings—numerical vector representations of data (images, text, audio, etc.) - which are stored in a vector database. These embeddings are organized by - Origin (starting point of the vector), Direction (where the vector moves) and Magnitude (length of the vector). The similarity between two vectors determines how closely related two images are. Unlike traditional databases, which search for exact matches in rows and columns, vector databases use algorithms to find the most similar vector to a query. This process is called Retrieval Augmented Generation (RAG) in LLMs. RAG enhances LLM outputs by retrieving additional context beyond what the model was originally trained on. It helps improve classification accuracy by finding the most relevant embeddings [14, 16, 17]. Fig.2. provides a visual representation of how images are stored in a vector database.

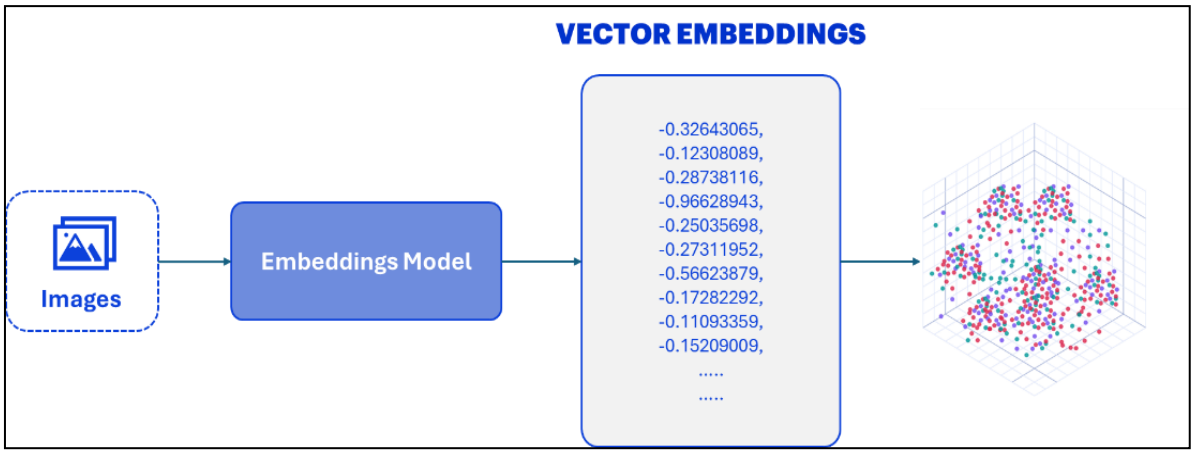

Fig. 2. Medical Images to Vectors conversion (illustrative)

This research uses Google's Gemini LLM's Generative AI image embeddings with custom adjustments. LLMs use neural networks to analyze and understand data. The process includes: Tokenization – Breaking input into small, manageable parts; Vectorization – Converting these parts into numerical values (vectors) and Embeddings – Capturing relationships between data points to allow advanced analysis. Unlike traditional models, this LLM setup does not require extra pre-processing. Unlike traditional AI models, LLMs do not require pre-processing - they handle normalization and feature extraction automatically. This means the system can process raw images directly, capturing important features without extra steps.

2.1. Proposed Methodology and Approach

A detailed methodology and approach outlining the LLM construct is as follows-

2.1.1. Vectorization Using LLM:

The first step in the methodology involves processing each image through an LLM model to generate a vector embedding. Although LLMs are originally designed for handling text, they are adapted here for image data through a specialized approach. This adaptation leverages pre-trained neural networks that can extract meaningful features from images and encoding them into high-dimensional vectors.

2.1.2. Feature Extraction:

The LLM utilizes deep convolutional layers combined with attention mechanisms to extract and encode features from images. These features include fundamental attributes such as edges, textures, shapes, and complex patterns. Beyond these basic features, the model also captures contextual meaning—how these attributes interact within the image to form coherent structures and objects. By understanding the relationships between various features, the LLM can discern intricate patterns that are critical for distinguishing between image categories.

2.1.3. Embedding Generation:

To ensure dimensionality balance, each image is represented by a 512-dimensional vector embedding – high enough to capture complex semantic information and low enough to prevent the 'curse of dimensionality', ensuring efficient similarity comparisons and reducing computational load. This high-dimensional vector not only captures low-level details and high-level abstract features but also encodes contextual relationships within the image. For instance, the vector reflects how different elements in the image relate to one another, such as the spatial arrangement of tissues in a medical scan. The choice of 512 dimensions is based on empirical testing to ensure a balance between retaining detailed contextual information and maintaining computational efficiency. 512-dimensional vectors are compact enough to allow fast indexing & retrieval in vector databases. This approach allows the embeddings to represent both the intrinsic characteristics of the image and the interrelationships between its components, enhancing the accuracy and relevance of the classification [18, 19].

2.1.4. Storage and Labelling:

The vector embeddings are stored in a vector database, which is designed to handle high-dimensional data efficiently. Along with each vector, the corresponding classification label (Malignant or Non-Malignant) is also stored.

i. Vector Database: The database is optimized for fast retrieval and similarity search. It supports efficient querying mechanisms to find vectors that are most similar to a given query vector.
ii. Labelling: Each stored vector is associated with a label that reflects the true classification of the image it represents. This labelling is crucial for accurate classification of new images.

2.1.5. Classification of New Image

When a new medical image is uploaded, it undergoes the same vectorization process as the images in the dataset. The new image is processed by the LLM to generate its 512-dimensional vector embedding. This embedding reflects the characteristics of the new image in the same feature space as the stored embeddings. The core of the classification process involves comparing the new image's vector embedding with those stored in the vector database. This is established through the concept of 'Cosine similarity' which is a metric used to measure how similar two vectors are in a high-dimensional space. It is especially useful in scenarios where the magnitude of the vectors is not as important as their direction. In other words, cosine similarity quantifies how closely the orientation of two vectors aligns, regardless of their length [20, 21].

The cosine similarity between two vectors A and B is calculated using the formula:

$$Cosine\ Similarity = \frac{A \cdot B}{\|A\|\|B\|} \tag{1}$$

where: A·B is the **dot product** of vectors A and B. ‖A‖ and ‖B‖ are the **Euclidean norms** (or magnitudes) of vectors A and B, respectively.

Dot Product: The dot product of two vectors A and B is computed as:

$$A \cdot B = \sum_{i=1}^{n} A_i \cdot B_i \tag{2}$$

where A_i and B_i are the components of vectors A and B, and n is the number of dimensions in the vectors. The dot product measures the extent to which the vectors point in the same direction.

Euclidean Norm: The Euclidean norm of a vector A is calculated as:

$$\|A\| = \sqrt{\sum_{i=1}^{n} A_i^2} \tag{3}$$

It represents the length or magnitude of the vector in the vector space. Similarly, the Euclidean norm of B is computed in the same way.

The output of Cosine Similarity score is interpreted as follows-

- Cosine Similarity = 1: The vectors are perfectly aligned, meaning they point in the same direction in the vector space. They are highly similar.
- Cosine Similarity = 0: The vectors are orthogonal (perpendicular) to each other, indicating no similarity.
- Cosine Similarity = -1: The vectors are diametrically opposed, meaning they point in exactly opposite directions.

Cosine similarity is particularly useful in this context because it focuses on the direction of the vectors rather than their magnitude, making it suitable for comparing the relative positions of image embeddings irrespective of their scale.

2.1.6. Retrieve Nearest Neighbor

Nearest Neighbor Search is a technique used to find the closest vector(s) to a given query vector in a dataset which in this case involves comparing the vector embedding of a new image with all stored vectors in the database to determine its closest match. It comprises few steps as mentioned below-

i. Vector Comparison: For a new image, its vector embedding is compared against all vectors in the vector database using the cosine similarity metric. Each stored vector represents an image from the training dataset.
ii. Finding the Closest Match: The vector with the highest cosine similarity to the new image's vector is identified as the closest match. This vector is the most similar in terms of orientation and feature representation.
iii. Label Assignment: The label associated with the closest matching vector is assigned to the new image. This label reflects the category (Malignant or Non-Malignant) of the most similar image in the database.
iv. Calculate Cosine Similarities: For the new image's vector V_{new}, calculate the cosine similarity with each vector V_i in the database.

$$Similarity_i = \frac{V_{new} \cdot V_i}{\|V_{new}\| \cdot \|V_i\|} \tag{4}$$

v. Identify Maximum Similarity: Find the vector $V_{closest}$ with the maximum cosine similarity.
vi. Retrieve and Assign Label: Retrieve the label associated with $V_{closest}$ and assign it to the new image. This label represents the category that the new image is most likely to belong to, based on the closest match.

2.1.7. Visualization of Classification Results

i. Cluster Plot Construction: To visualize the classification results and the distribution of embeddings, a cluster plot has been created. This plot provides a spatial representation of how embeddings are organized and how the new image's embedding fits into this structure [22, 23].

ii. Dimensionality Reduction: Since the vector embeddings are 512-dimensional, dimensionality reduction techniques are applied to project these high-dimensional vectors into a lower-dimensional space. The t-Distributed Stochastic Neighbor Embedding (t-SNE) has been used to maintain the relative distances between vectors while simplifying the visualization.

iii. Plot Details – Fig.3. displays the following:

- Orange Dots**:** Represent embeddings of images labelled as Malignant. These dots cluster together, showing the spatial arrangement of malignant images in the reduced-dimensional space.
- Green Dots: Represent embeddings of images labelled as Non-Malignant. Similarly, these dots cluster separately from malignant ones, illustrating the spatial separation between categories.
- Pink Dot with blue outline (Larger): Represents the vector embedding of the new image. The position of this dot relative to the red and green clusters indicates the classification result. If the blue dot is closer to the red cluster, it is classified as Malignant; if closer to the green cluster, it is classified as Non-Malignant.

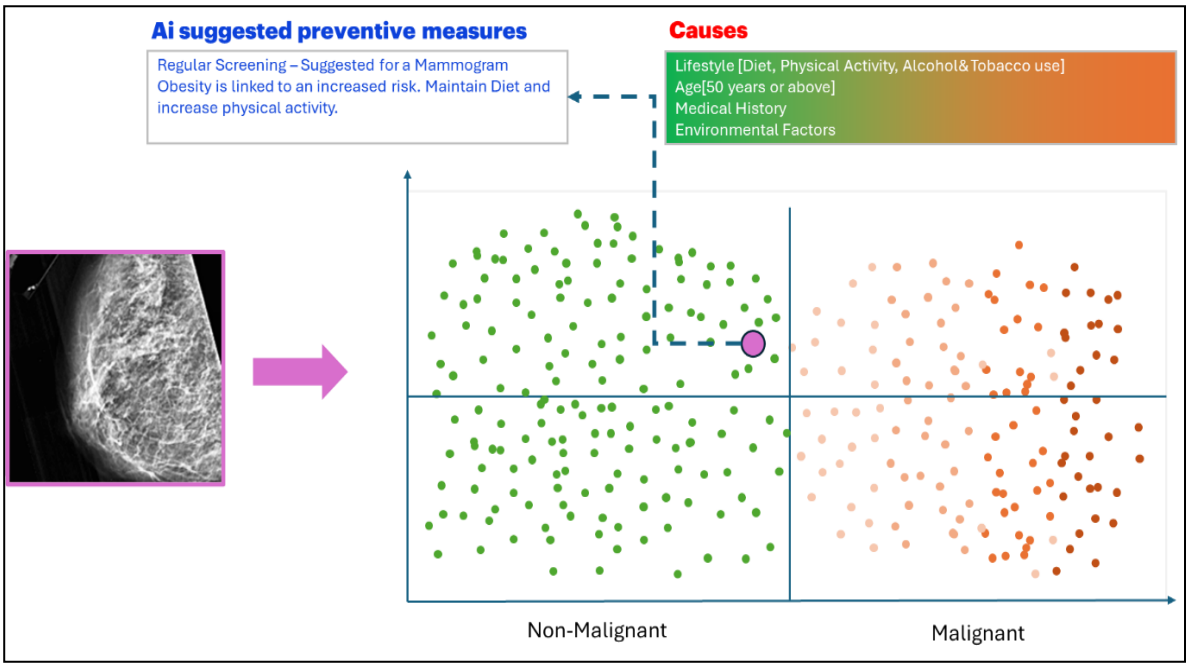

Fig. 3. Illustrative Cluster Plot for Malignancy segmentation

2.1.8. Interpretation of Results

The cluster plot enables users to visually assess the placement of the new image's vector within the context of the existing data. By examining the proximity of the blue dot to the clusters of red and green dots, users can quickly understand the classification decision and see how the new image compares to the known categories.

2.2. Advantages of LLM Embeddings over CNN

This unique approach has several advantages over CNN and other ML techniques which is summarized in Table 1. [24]-

Table 1. Differentiation between LMM Embeddings versus CNN Model

S.No.	Parameters	Vector Embeddings	CNN
1	Datasets	Vector embeddings generated by LLMs can be highly effective even with relatively small datasets. This is because LLMs, especially when pre-trained, capture rich feature representations that are transferable to specific tasks with limited data.	Traditional CNNs generally require large amounts of data to train effectively. With small datasets, CNNs are prone to overfitting, where the model learns the specifics of the training data too well and performs poorly on unseen data.
2	Feature Representation and Contextual Understanding	LLMs are designed to capture contextual meaning and complex relationships within the data. This means that vector embeddings not only represent individual features but also encode the relationships between them, leading to a richer and more nuanced understanding of the image content.	While CNNs are excellent at extracting spatial features through convolutional layers, they may struggle to capture higher-order relationships or contextual meaning without extensive design and tuning. CNNs focus more on local feature extraction and require more manual intervention to incorporate contextual understanding.
3	Training	Once the LLM model is pre-trained, generating embeddings for new images requires less computational effort compared to training a CNN from scratch. The embeddings can be directly used for similarity comparison and classification.	Training CNNs from scratch involves significant computational resources and time, particularly for deep architectures. Even with transfer learning, fine-tuning a CNN on a new dataset still requires substantial training effort and computational power.
4	Classification and Adaptability	Classification using vector embeddings and cosine similarity is straightforward. The method involves comparing the new image's vector against stored vectors and finding the closest match. This approach is easy to implement and adapt to new categories or changes in the dataset.	CNNs require a more complex classification pipeline, including training a classifier (like a softmax layer) on top of the CNN features. Adapting a CNN to new classes or modifying the classification task involves retraining or fine-tuning, which can be cumbersome.
5	Interpretability	The use of cluster plots to visualize embeddings provides clear insights into how images are grouped and classified. This visualization helps in understanding the relationships between different categories and the positioning of new images within the existing data space.	While CNNs can be interpreted to some extent through techniques like feature maps and activation visualization, these methods do not always provide an intuitive understanding of the overall classification decision or the relationships between different
6	Computational Efficiency	Once embeddings are computed, the classification process involves relatively low computational overhead. Searching for the nearest neighbor in the embedding space is efficient, especially with optimized data structures like KD-trees or approximate nearest neighbor methods.	Training and inference with CNNs are computationally intensive, requiring GPUs and large amounts of memory. Even during inference, running a deep CNN can be resource-heavy compared to the lightweight comparison of vector embeddings.
7	Flexibility and Scalability	The approach of using embeddings and cosine similarity is highly flexible and can easily scale to different types of data or classification tasks. New data can be incorporated seamlessly into the existing vector space, and the system can be updated with minimal adjustments.	Scaling CNNs to new tasks or integrating new data often requires retraining or fine-tuning the entire network. This process can be less flexible and more resource-intensive compared to updating an embedding-based system.
8	Robustness to Noise and Variability	Embeddings often capture high-level features that are robust to noise and variability in the data. They tend to focus on the overall structure and relationships within the image, making them more resilient to minor perturbations or variations.	CNNs can be sensitive to noise and variations in the input data, especially if the network has not been sufficiently trained or regularized. Ensuring robustness often requires additional training techniques and data augmentation strategies.
9	Pre-Processing Requirement	The use of LLMs for generating embeddings often does not require extensive pre-processing of images. The model is designed to handle raw image data and generate meaningful embeddings directly. This eliminates the need for complex pre-processing steps such as normalization, resizing, or augmentation before feature extraction.	Traditional CNNs typically require pre-processing steps to standardize input data. Common pre-processing tasks include resizing images to a consistent size, normalizing pixel values, and augmenting the dataset to enhance generalization. These steps are crucial for ensuring that the input data is suitable for training and inference but add extra complexity and computational overhead to the workflow.

3. Results

3.1. Study & Experimental Design

Dataset 1 - A leading Oncology treatment hospital in Bangalore, India was the basis of sourcing Real-World Data (RWD) for the research work. Given limited availability of RWD and sensitivity around the collection of patient information to be used for such a kind of research work, limited stratification criteria were applied to ensure reasonable availability of desired dataset for meaningful analysis and insights generation. The following were considered as minimum inclusion criteria – sample size of 150 images of various types such as Mammograms (MG), PET scans, CT scans, Ultrasounds (US), patients with age greater than 30 years irrespective of stage of their cancer, availability of textual information such as medical history, radiology reports, discharge summaries, etc. The attributes of the collected dataset were - 53 patients of Indian origin within age group of 32-76 years who underwent treatment in the year 2023 and 2024, with a total of 193 images of various types such as Mammograms (MG), CT scans, Ultrasounds (US), PET scans, patients' medical history, radiology reports including images, doctor notes, lab reports, treatment regimen, discharge summary etc. For future studies, Dataset 1 can be further enriched by increasing the sample size through pooling of data from multiple hospitals, stratification based on stage of cancer, types of images, age, etc.

Dataset 2 - Another dataset was sourced from the work performed by Huang ML and Lin TY [25, 26]. The characteristics of the dataset were noted as follows – dataset comprising of 250 breast mammography images which were combination of both malignant and non-malignant images. These images as accessed from the source were already pre-processed using contrast limited adaptive histogram equalization (CLAHE) method and then data augmentation was performed to prevent the problem of overfitting. The use of CLAHE method directly impacts feature extraction, model robustness, and overall diagnostic accuracy. Occasionally, Mammograms or MRIs, suffer from low contrast, noise, and artifacts, making it difficult for models to distinguish between pathological and normal regions. CLAHE locally enhances contrast, making subtle patterns like microcalcifications, lesions, or tumor boundaries more distinguishable. It also improves the visibility of faint features that might be crucial for detecting early-stage cancers. CLAHE reduces bias from overexposed or underexposed regions, ensuring that embeddings capture meaningful pathology rather than noise.

RWD dataset was first processed through the image classification system represented in Fig.1. As illustrated in the section – 'Methods', the process involves converting medical images of various types - MG, PET, CT, US, etc. into vector embeddings, and subsequently storing these embeddings in a vector database. The RWD dataset was then analyzed through Image Analysis & Visualization Application (IAVA) for insights and evidence generation. The steps involved are described as follows:

3.1.1. Data Conversion

Patient records were converted into a structured format (blinded and anonymized) using an automated extraction tool by developing a highly advanced .NET Core utility with MS SQL as the database. This app is tailored to process and organize data from unstructured PDF documents. This utility leverages cutting-edge technology to convert raw text into structured information with impressive accuracy. Below is the flow of the application:

i. PDF Upload and Text Extraction: upload the unstructured PDFs into the application. The utility extracts the full text from these documents, ensuring all relevant information is captured for further processing.

ii. Sentence Segmentation: The extracted text is divided into individual sentences, a crucial step that allows for precise and focused analysis of each sentence.

iii. Prompt Engineering and Entity Extraction [27, 28]**:**

- Prompt Engineering: This system utilizes advanced prompt engineering techniques to guide OpenAI's large language model (LLM) in extracting specific types of information. These prompts are fine tuned to ensure comprehensive coverage of various data points.
- Large Language Model (LLM): Sentences are processed using OpenAI's state-of-the-art LLM, which applies deep learning algorithms to understand context and semantics, allowing for accurate identification and classification of entities.
- Entity Types: The system extracts a wide range of entities, including:
  a. Personal Identifiers: Patient ID, age, gender, location, visit date, demographics, ethnicity, marital status, occupation.
  b. Medical Information: Lifestyle habits, medical history, medication history, scans, procedures.
  c. Doctor's Notes: The utility also effectively extracts additional, context-specific entities from doctor's notes, such as detailed observations, treatment plans, and diagnostic impressions.

iv. Data Formatting and Storage:

- JSON Conversion: Extracted data is returned in a JSON format, which includes detailed information on identified entities and their classifications.
- Structured Data Conversion: The application parses the JSON data and converts it into a structured format compatible with the relational database schema.

- Database Integration: The structured data is stored in a relational database management system (RDBMS). The data is organized across multiple tables with careful attention to relational mapping, ensuring efficient data retrieval and management.

v. Accuracy and Performance: The system achieved a remarkable accuracy rate of 90% and higher in entity extraction.

vi. Manual Review and Error Correction:

- Manual Review: Following the automatic extraction and database storage processes, a manual review phase is implemented to ensure data quality and accuracy. Trained personnel review the extracted and stored data, validating the correctness of the information and identifying any discrepancies or errors.
- Error Correction: Any inaccuracies or issues identified during the manual review are corrected, ensuring that the data stored in the database is reliable and accurate. This step is crucial for maintaining high data integrity and addressing any nuances or exceptions that may not have been captured perfectly by the automated system.

3.1.2. Image Classification System (ICS) Set-up

An advanced image classification system was developed using Google Gemini LLMs Gen AI image embedding, which were customized to analyze and classify medical scan images. As described in the section – 'Methods', the process involves converting images into vector embeddings, storing these embeddings in a vector database, and then classifying new images based on similarity metrics.

3.1.3. Image Analysis & Visualization Application (IAVA) Set-up

A real-time image analysis application has been developed to process medical images instantly. Users can upload images, analyze them, and view classification results through a cluster plot, making it easier to understand the outcomes (see Fig.4.).

The key features of the application include the following:

i. Technology Stack: A backend built with .NET Core 3 (C#) for efficient processing that uses Entity Framework Core with SQLite for data management. A frontend that uses jQuery for interactivity and Bootstrap for a responsive design. AI Integration achieved through the Gemini API which generates image embeddings for advanced processing, enabling semantic search and content recommendations.

ii. User Interface: On left side: "Upload Image" button and image viewer. On right side: three sections—Uploaded Image Details, Image Analysis, and Graphical Representation.

iii. How It Works: Users upload an image via the Browse button; the image appears in the viewer for verification; Clicking Analyze triggers the AI model to process the image and generate a vector representation; the vector is then compared with stored vectors, and similarity scores are calculated. Finally, the top match's class and score are displayed in the Image Analysis section.

iv. Visualization & Insights: Users can view image metadata (dimensions, name, file path). A cluster plot categorizes images into malignant and non-malignant groups (see Fig.3.). The uploaded image's position is highlighted, helping users interpret results easily. This application provides a fast, user-friendly way to analyze medical images with AI.

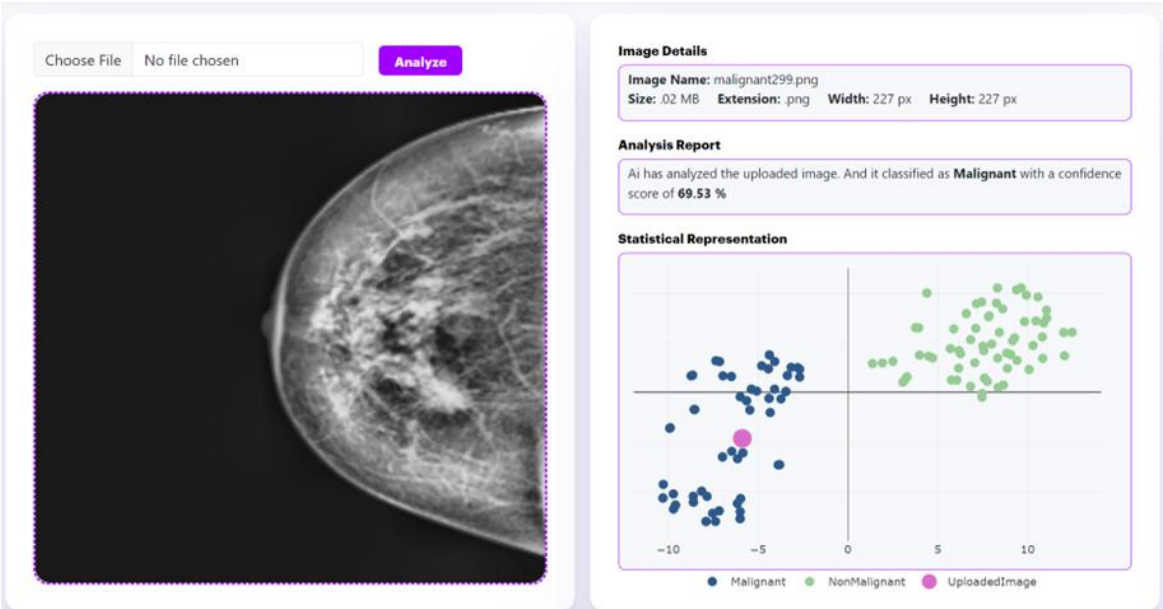

Fig. 4. RWD – Sample Cluster Plot for Malignancy Classification

3.2. Study & Experimental Analysis Outputs

In order to determine the effectiveness of using LLM embeddings and further validating the hypothesis that LLMs produce a more consistent and accurate output with relatively smaller datasets (with variety of medical scan images) versus a CNN model that may require large amount of data to train effectively, the two datasets (Dataset 1 and Dataset 2) were processed and analyzed through IAVA and then through a CNN model [33].

Approach 1: Determine the accuracy of output by segmenting the dataset into 25% training, 10% validation, 65% testing and run both LLM embeddings and CNN model

Approach 2: Determine the accuracy of output by segmenting the dataset into 70% training, 10% validation, 20% testing and run both LLM embeddings and CNN model

Using either of the above approaches, the LLM model was trained on a combination of mammogram, ultrasound, PET scans and MRI images, which represent different modalities of breast cancer detection. Cross-validation experiments were performed to evaluate the model's performance across different subgroups, including age, and other medical conditions. This approach enabled to assess how well the model generalizes to diverse patient populations and different stages of disease progression. Sections 4.1 and 4.2 further outline the outcomes of these analyses.

The below tables (Table 2a. and 2b.) summarize the analysis output of imaging dataset 2 after running the LLM embeddings model and a CNN model using approach 1 and 2.

Table 2a. Dataset 2 - LMM Embeddings versus CNN Model accuracy (Approach 1)

Dataset	**Embeddings (Count of Images)**	**CNN (Count of Images)**
Total	250	250
Train	50	50
Validate	20	20
Test	180	180
No. of errors	0	37
Accuracy	**100%**	**79.4%**

Table 2b. Dataset 2 - LMM Embeddings versus CNN Model accuracy (Approach 2)

Dataset	**Embeddings (Count of Images)**	**CNN (Count of Images)**
Total	250	250
Train	180	180
Validate	20	20
Test	50	50
No. of errors	0	1
Accuracy	**100%**	**98.0%**

In both the approaches, LLM Embeddings return 100% accurate results versus CNN model accuracy being 79.4% and 98% with approach 1 and 2 respectively. The output demonstrates superiority of LMM embedding over traditional AI methods such as CNN.

The summary of the analysis using approach 1, for Dataset 1 is captured in the Tables 3a. and 3b.

Table 3a. Dataset 1 - RWD Dataset - LMM Embeddings accuracy (Approach 1)

RWD Dataset	**Embeddings (Count of Images)**				
	Overall	**CT**	**MG**	**PT**	**US**
Total	193	74	73	16	30
Train	46	4	29	5	8
Validate	19	9	5	1	4
Test	128	61	39	10	18
No. of errors	2			1	1
Accuracy	**98.4%**	**100.0%**	**100.0%**	**90.0%**	**94.4%**

Table 3b. Dataset 1 - RWD Dataset - CNN Model accuracy (Approach 1)

RWD Dataset	**CNN (Count of Images)**				
	Overall	**CT**	**MG**	**PT**	**US**
Total	193	74	73	16	30
Train	46	4	29	5	8
Validate	19	9	5	1	4
Test	128	61	39	10	18
No. of errors	49	5	33	6	5
Accuracy	**61.7%**	**91.8%**	**15.4%**	**40.0%**	**72.2%**

In approach 1, LLM Embeddings return results in the range of 90% - 100% accuracy results versus CNN model's accuracy ranging from 40% - 91.8%.

The summary of the analysis using approach 2, for Dataset 1 is captured in the Tables 4a. and 4b.

Table 4a. Dataset 1 - RWD Dataset - LMM Embeddings accuracy (Approach 2)

RWD Dataset	Embeddings (Count of Images)				
	Overall	**CT**	**MG**	**PT**	**US**
Total	193	74	73	16	30
Train	135	57	52	7	19
Validate	19	6	7	3	3
Test	40	11	14	6	9
No. of errors	0	0	0	0	0
Accuracy	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Table 4b. Dataset 1 - RWD Dataset - CNN Model accuracy (Approach 2)

RWD Dataset	CNN (Count of Images)				
	Overall	**CT**	**MG**	**PT**	**US**
Total	193	74	73	16	30
Train	135	57	52	7	19
Validate	19	6	7	3	3
Test	40	11	14	6	9
No. of errors	17	3	6	1	7
Accuracy	**57.5%**	**72.7%**	**57.1%**	**83.3%**	**22.2%**

In approach 2, LLM Embeddings return results with 100% accuracy versus CNN model's accuracy ranging from 22.2% - 83.3%.

The output from RWD dataset further validates the hypotheses that LMM embeddings generate more accurate output over traditional AI methods such as CNN for all different kind of medical images, thereby enabling more consistent, specific and accurate determination of malignancy. Though LLM embeddings approach has yielded higher accuracy as compared to CNN model both with approach 1 and approach 2, future studies on larger datasets need to be conducted to further validate and substantiate the hypothesis.

4. Discussion

LLM embeddings-based approach has some significant benefits for this kind of research work [29, 30], enlisted below-

- **Longitudinal Patient Monitoring** -Supports longitudinal monitoring of patients over time by tracking changes in image embeddings and detecting subtle progression or regression of malignancies
- **Early Malignancy Detection** - By accurately identifying malignancy in its early stages, the system enables timely medical interventions and treatment planning, potentially improving patient outcomes and survival rates
- **Reduced dependency on image scaling or pre-processing** – For faster retrieval and classification of medical images, semantic search and cosine similarity usage outperforms traditional image processing techniques
- **Adaptability to new image types and domains** - The system's ability to generate embeddings based on image content allows for easy adaptation to new types of medical images and domains without significant retraining
- **Enhanced efficiency as compared to traditional models** - The use of Gen Ai though semantic classification of medical images is more efficient compared to traditional image processing techniques
- **Efficient training with limited datasets** - Generative AI embeddings can learn meaningful representations from limited datasets, reducing the need for extensive data annotation and training time

4.1. Key Results

For Dataset 2, as is evident (from Tables 2a. and 2b.), for both approaches 1 and 2, LMM Embeddings outcome had a 100% accuracy. The CNN model demonstrated discrepant outcome with 79.4% accuracy using approach 1 but had a significantly improved accuracy of 98% with approach 2. With approach 1, CNN model incorrectly classified the images on 37 occasions out of 180 unique images that were tested and with approach 2, incorrectly classified the images on 1 occasion out of 50 unique images that were tested. On the other hand, LLM Embeddings even with smaller dataset used for training was able to produce 100% accurate outcome as compared to the CNN model.

For Dataset 1 (RWD), with approach 1, as is evident (Table 3a.), LMM Embeddings achieved an accuracy of 98.44% (out of the 128 unique images that were tested, it incorrectly classified the images on 2 occasions). The model demonstrated 100% accuracy for both CT scans and Mammograms where 61 and 39 images were tested respectively though the training dataset was only limited to 4 and 29 respectively. PET scans and Ultrasounds had an accuracy of 90%

and 94.4% respectively, which is significant given the type of complexity associated with these kinds of scans and the fact that these images were not pre-processed using any segmentation or filtration techniques. The CNN model (Table 3b.), however, demonstrated discrepant outcomes with 61.7% accuracy and out of the 128 unique images that were tested, it incorrectly classified the images on 49 occasions. CT scans and Ultrasounds had a decent accuracy of 91.8% and 72.2% respectively but accuracy of Mammograms and PET scan was low at 15.4% and 40% respectively. The above analysis confirms the hypotheses that LLM Embeddings produce a more consistent and accurate outcome even with smaller datasets. This is a significant observation given the limited availability of medical imaging datasets to generate insights and evidence generation for informed decision making.

Again, for Dataset 1 (RWD), with approach 2, as is evident (Table 4a.), LMM Embeddings achieved an accuracy of 100% (out of 128 unique images that were tested, there were 0 errors), thereby substantiating the evaluation of the Oncologist and Radiologist proving to be consistently giving an output matching their assessment. Even with PET scans and Ultrasounds, an accuracy of 100% was observed, which is significant given the type of complexity associated with these kinds of scans and the fact that these images were not pre-processed using any segmentation or filtration techniques. The CNN model (Table 4b.), however, demonstrated discrepant outcomes with 57.5% accuracy and out of 40 unique images that were tested, it incorrectly classified the images on 17 occasions. CT scans and PET scans had a decent accuracy of 72.7% and 83.3% respectively but accuracy of Mammograms and Ultrasounds was low at 57.1% and 22.2% respectively. The above analysis further confirms the hypotheses that LLM Embeddings produce a more consistent and accurate outcome irrespective of the size of the dataset – small or large versus a CNN model which requires significantly large datasets to improve the accuracy of the outcome.

When analyzing medical images, Large Language Models (LLMs) and Convolutional Neural Networks (CNNs) interpret data in fundamentally different ways. LLM embeddings rely on similarity scores, which measure how closely related features are in a high-dimensional space, capturing subtle relationships and contextual patterns. In contrast, CNNs use confidence scores, which quantify how certain the model is about an image belonging to a specific category. While high confidence in CNNs may suggest strong certainty, it does not always correlate with accuracy, especially in complex cases where subtle patterns are crucial. Notably, LLM embeddings can provide correct assessments even with low similarity scores by focusing on relational insights rather than rigid classification. On the other hand, CNNs may misclassify images despite high confidence if they fail to capture abstract or nuanced features. This distinction is particularly important in medical imaging, where LLM-based methods can offer a more reliable way to identify intricate patterns, making them valuable for challenging diagnostic tasks.

4.2. Inferences

Table 5. presents sample blinded patient records from the Real-World Data (RWD) dataset, further supporting this hypothesis. The findings reinforce the superiority of LLMs over traditional AI approaches by demonstrating improved accuracy, reliability, and pattern recognition across multiple aspects of medical image analysis.

Table 5. LMM Embeddings versus CNN Model score comparison

Source Image	Image Category	Image Type	Embedding Gen AI detected category	Embedding Gen AI Similarity Score	CNN detected category	CNN Confidence Score
Patient #1	Malignant	CT	Malignant	89.70%	Malignant	85.28%
Patient #2	Malignant	CT	Malignant	85.68%	Malignant	97.41%
Patient #3	Malignant	MG	Malignant	81.64%	Non- Malignant	94.88%
Patient #4	Malignant	MG	Malignant	93.13%	Malignant	76.30%
Patient #5	Malignant	PT	Malignant	87.01%	Malignant	89.73%
Patient #6	Malignant	US	Malignant	86.98%	Non-Malignant	51.60%
Patient #7	Malignant	US	Malignant	72.24%	Non-Malignant	73.10%
Patient #8	Malignant	US	Malignant	90.86%	Non-Malignant	75.99%
Patient #9	Non-Malignant	PT	Non-Malignant	93.39%	Malignant	85.96%
Patient #10	Non-Malignant	US	Non-Malignant	79.93%	Non-Malignant	76.39%

Table 5. highlights that LLM embeddings achieved 100% accuracy across all 10 patient records, covering different image types—CT, MG, PT, and US. In contrast, the CNN model misclassified 50% of the cases, despite having comparable or even higher confidence scores than the similarity scores of LLM embeddings. This finding underscores the superiority of LLM embeddings, which can capture subtle patterns and relationships more effectively than traditional CNN-based approaches.

The sample output from blinded patient records including inferences from LLM embeddings, using the application is represented below for different type of medical images–

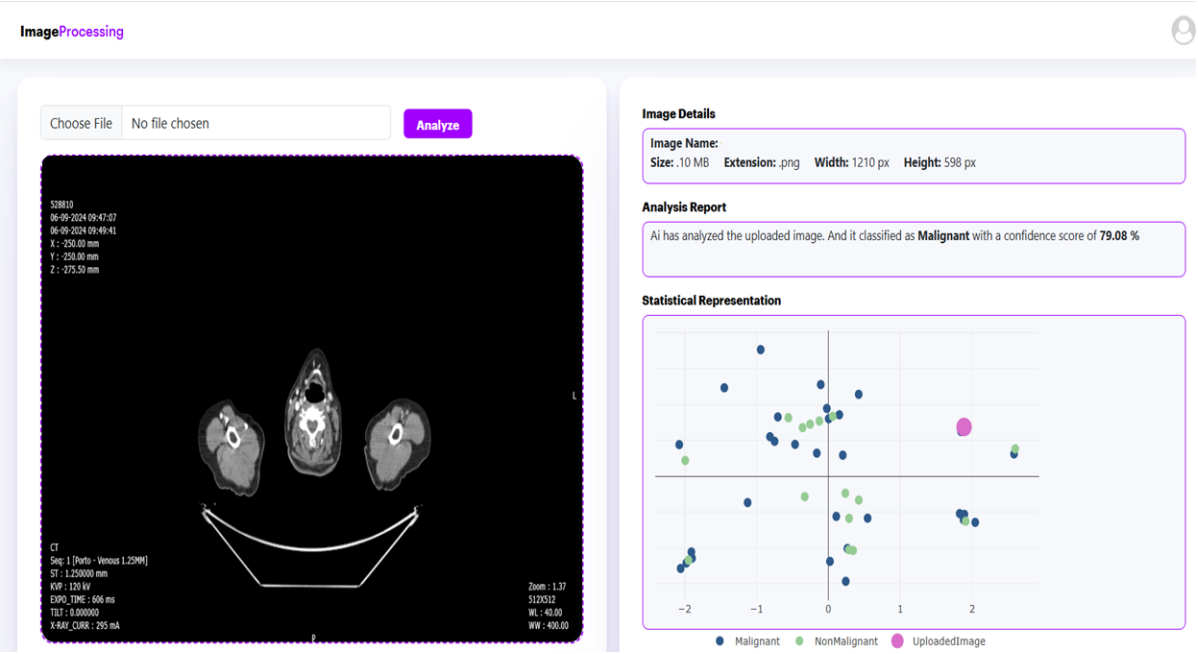

Fig. 5. RWD – CT scan with Approach #1 (Inference: Malignant – 79.08% confidence score)

Fig. 5. is a representation of a patient's CT scan (test image) that went through LLM embedding's 25% training dataset approach and yielded an outcome of malignancy with 79.08% accuracy.

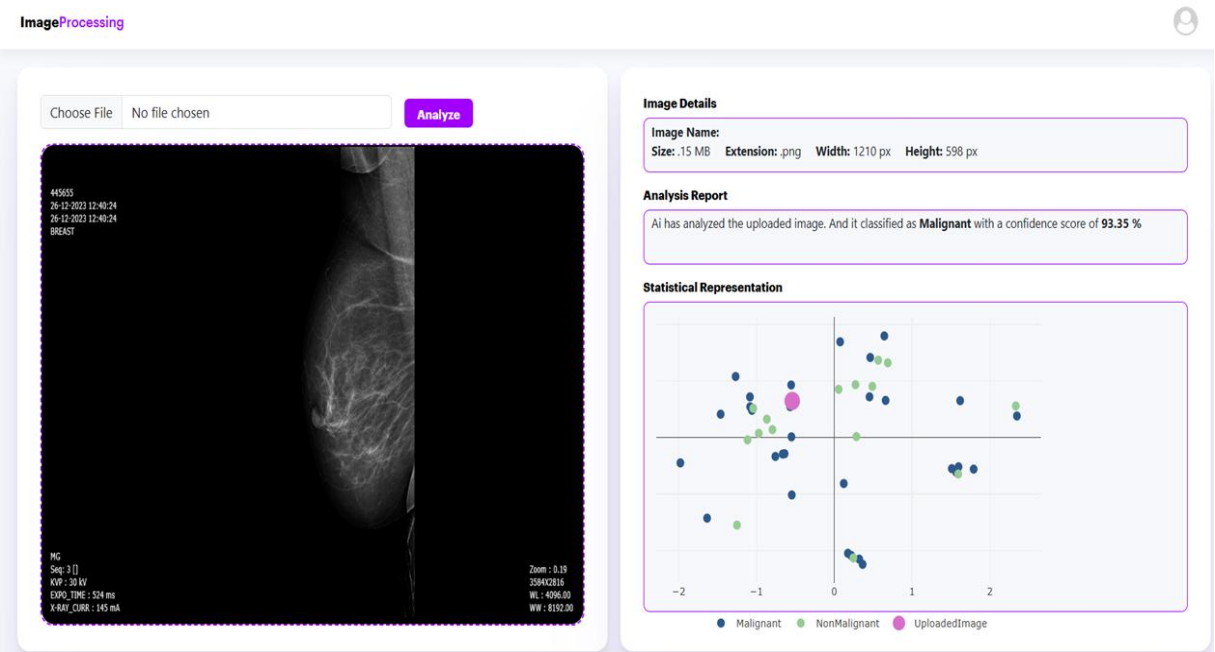

Fig. 6. RWD – MG with Approach #1 (Inference: Malignant – 93.35% confidence score)

Fig. 6. is a representation of a patient's Mammogram (test image) that went through LLM embedding's 25% training dataset approach and yielded an outcome of malignancy with 93.35% accuracy.

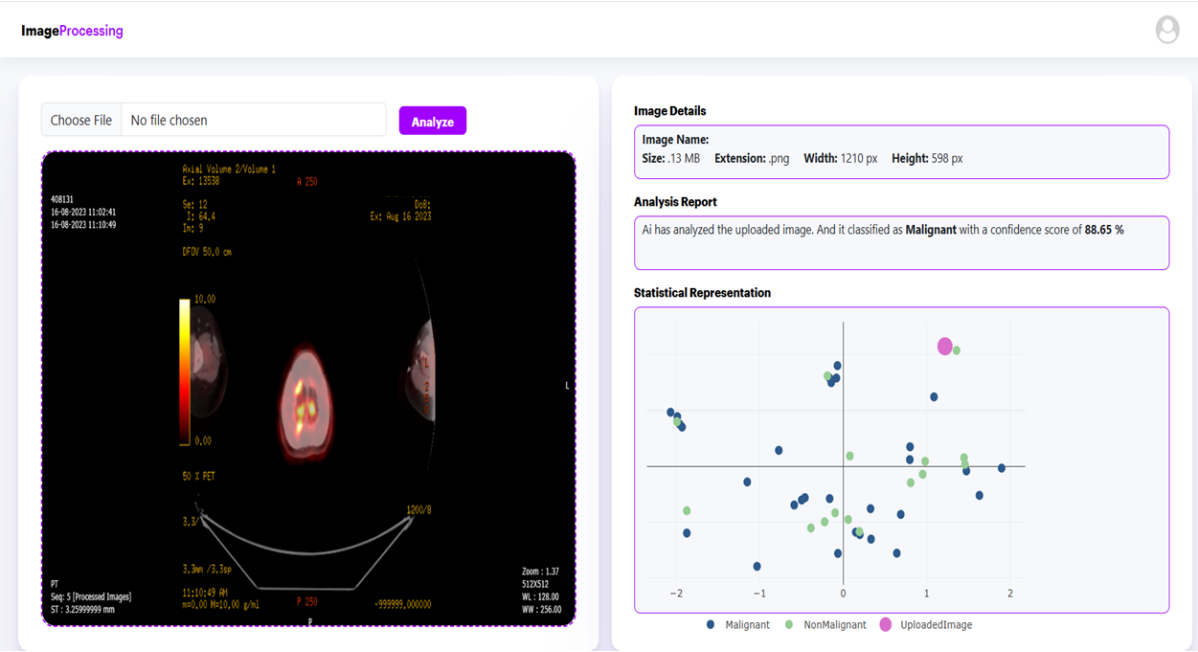

Fig. 7. RWD – PET Scan with Approach #1 (Inference: Malignant – 88.65% confidence score)

Fig. 7. is a representation of a patient's PET scan (test image) that went through LLM embedding's 25% training dataset approach and yielded an outcome of malignancy with 88.65% accuracy.

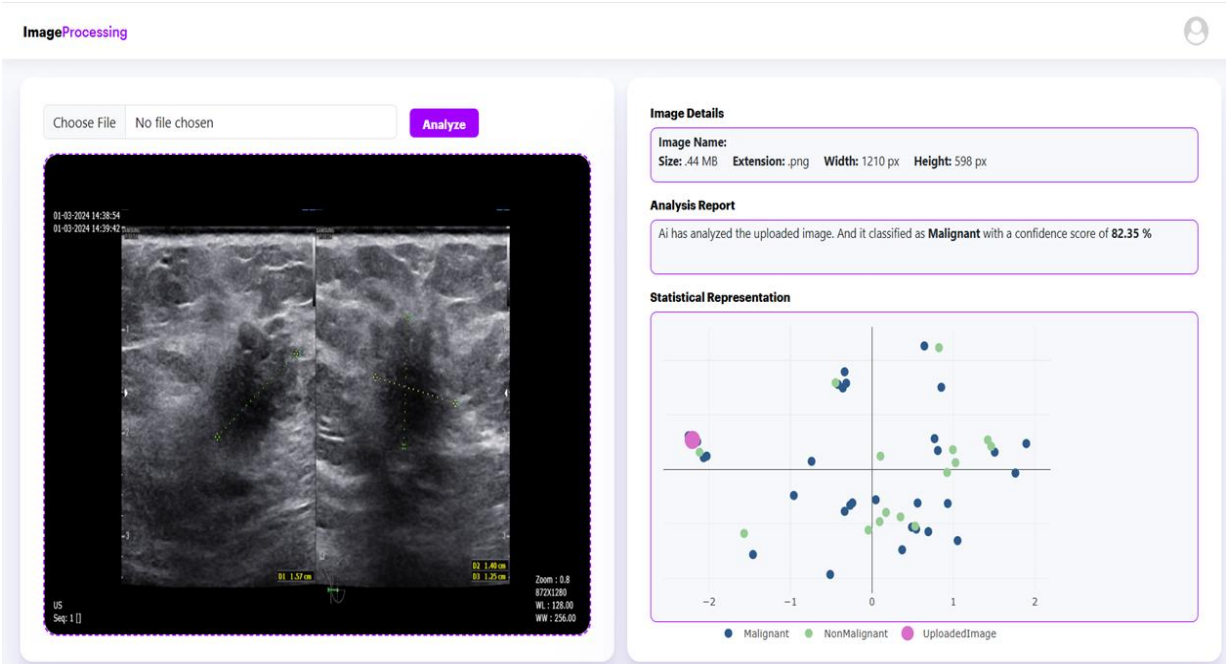

Fig. 8. RWD – Ultrasound with Approach #1(Inference: Malignant – 86.98% confidence score)

Fig. 8. is a representation of a patient's Ultrasound (test image) that went through LLM embedding's 25% training dataset approach and yielded an outcome of malignancy with 86.98% accuracy.

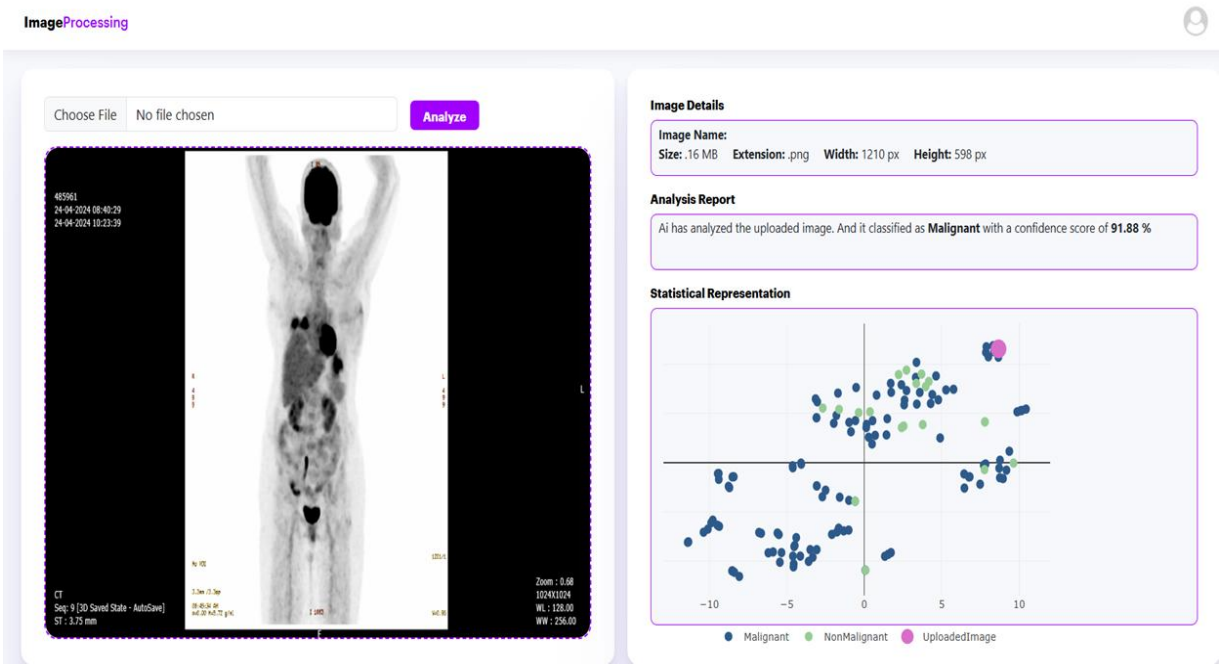

Fig. 9. RWD – CT scan with Approach #2 (Inference: Malignant – 91.88% confidence score)

Fig. 9. is a representation of a patient's CT scan (test image) that went through LLM embedding's 70% training dataset approach and yielded an outcome of malignancy with 91.88% accuracy.

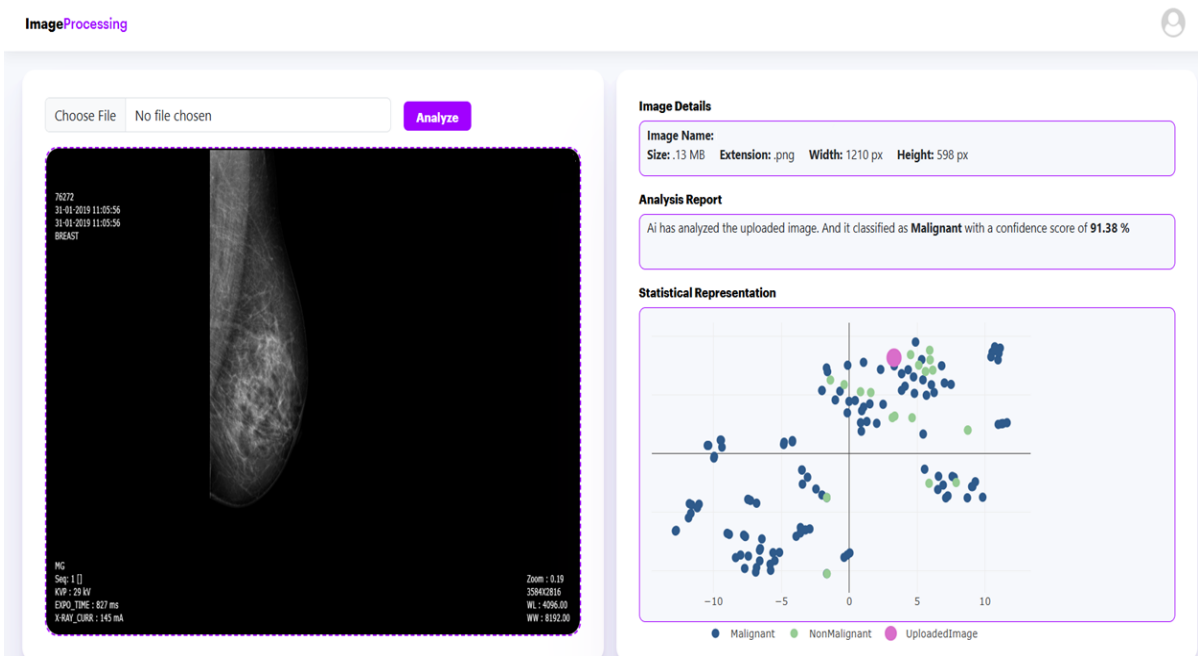

Fig. 10. RWD – MG with Approach #2 (Inference: Malignant – 91.38% confidence score)

Fig. 10. is a representation of a patient's Mammogram (test image) that went through LLM embedding's 70% training dataset approach and yielded an outcome of malignancy with 91.38% accuracy.

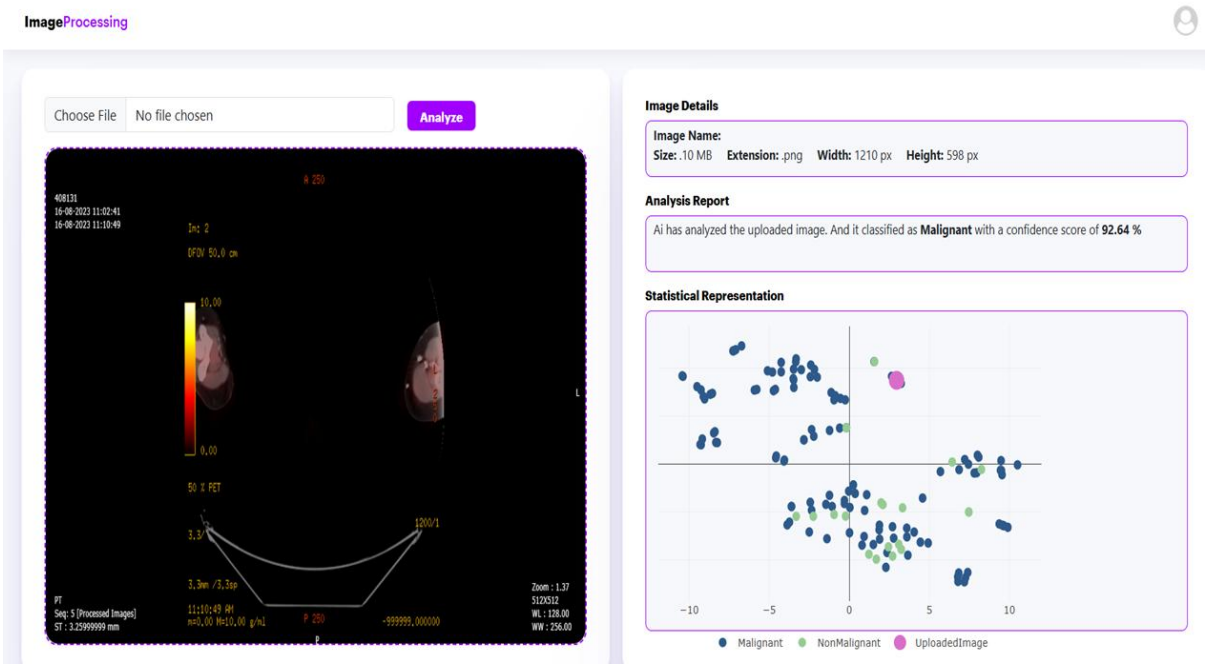

Fig. 11. RWD - PET scan with Approach #2 (Inference: Malignant – 92.64% confidence score)

Fig. 11. is a representation of a patient's PET scan (test image) that went through LLM embedding's 70% training dataset approach and yielded an outcome of malignancy with 92.64% accuracy.

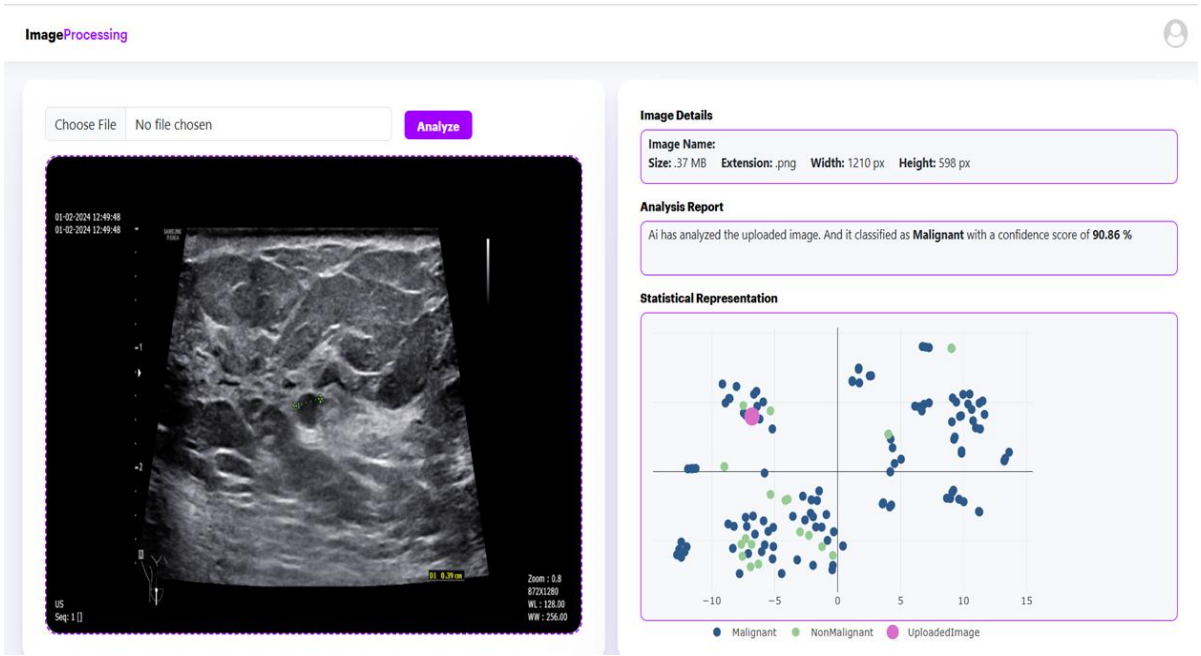

Fig. 12. RWD - Ultrasound with Approach #2 (Inference: Malignant – 90.86% confidence score)

Fig. 12. is a representation of a patient's Ultrasound (test image) that went through LLM embedding's 70% training dataset approach and yielded an outcome of malignancy with 90.86% accuracy.

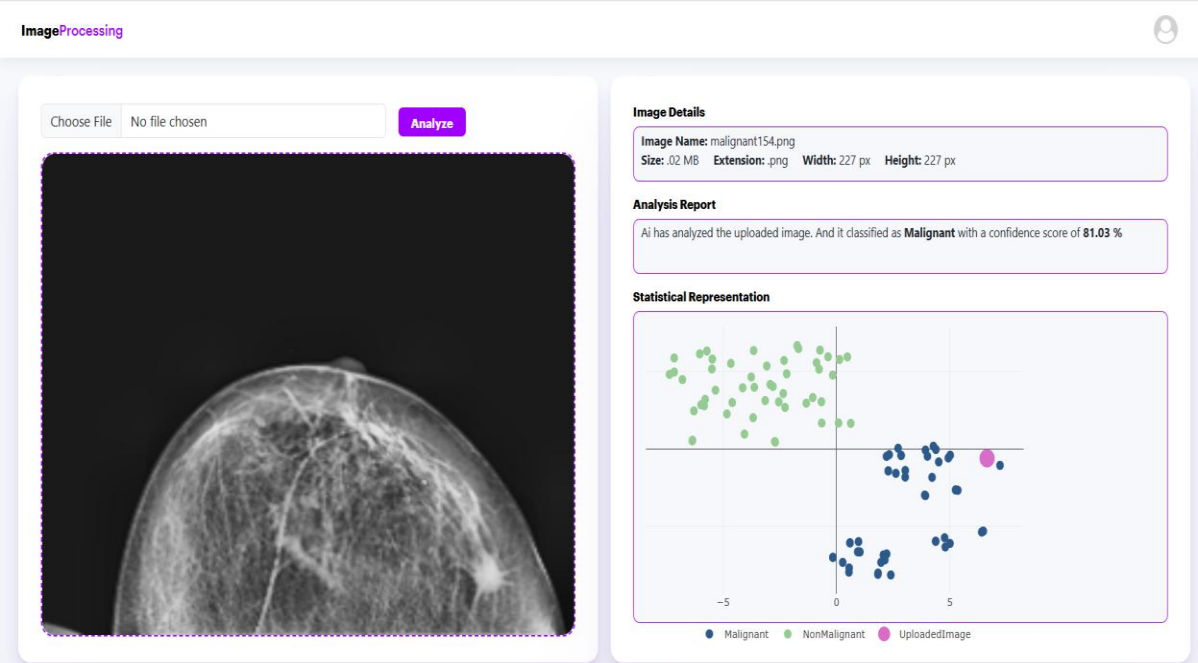

Fig. 13. Dataset 2 - MG with Approach #1 (Inference: Malignant – 81.03% confidence score)

Fig. 13. is a representation of a patient's Mammogram (test image) that went through LLM embedding's 25% training dataset approach and yielded an outcome of malignancy with 81.03% accuracy.

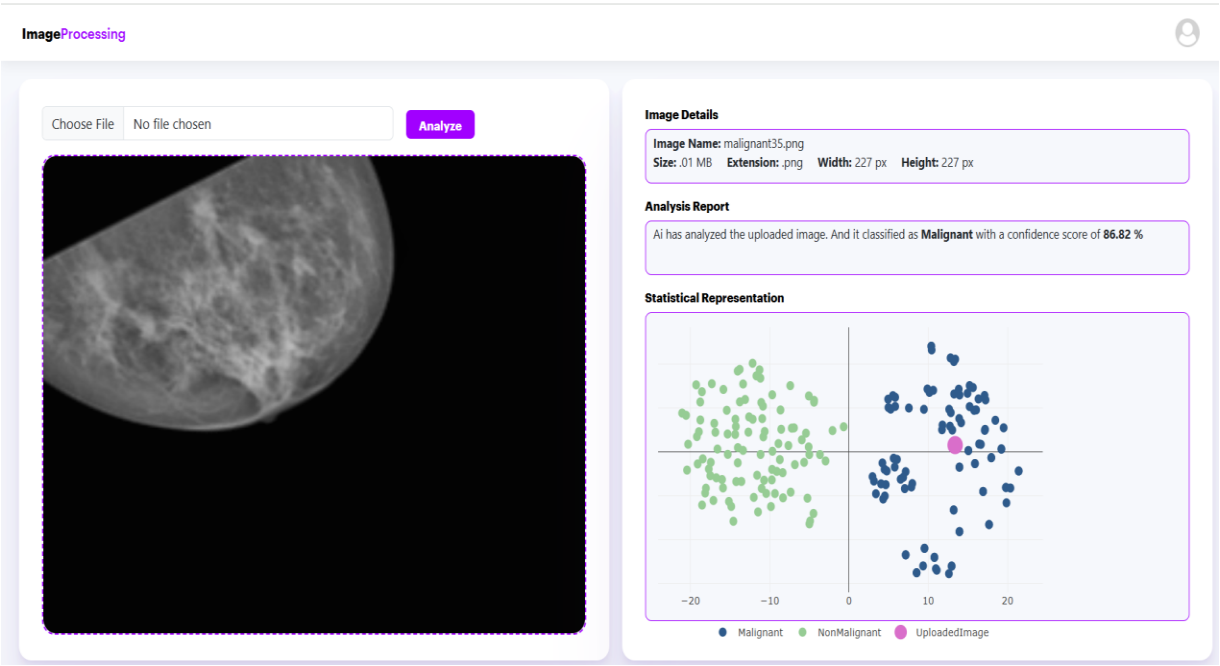

Fig. 14. Dataset 2 - MG with Approach #2 (Inference: Malignant – 86.82% confidence score)

Fig. 14. is a representation of a patient's Mammogram (test image) that went through LLM embedding's 25% training dataset approach and yielded an outcome of malignancy with 86.82% accuracy.

For the RWD of 53 patient records, a segmentation approach was adopted to understand the correlation between age groups and LLM Embeddings confidence score. Table 6. presents blinded patient records from the RWD dataset and Fig. 15. is a scatter plot for the same dataset. As is evident, for patients in age group between 45-65 years, LLM model demonstrates a high level of confidence score in the range of 75% to 94% across different types of medical images. The plot also demonstrates that irrespective of the age group, LLM Embeddings can classify and segment all types of medical images with a confidence/similarity score of minimum 40% and above (100% accuracy in malignancy detection).

Table 6. Age Group v/s LLM Embeddings Confidence Score

Age Group (years)	Confidence Score						
	40%-49%	50%-59%	60%-69%	70%-79%	80%-89%	90%-99%	Grand Total
30-34			2	1	1	1	5
35-39			1		7	2	10
40-44	3	1	2		2		8
45-49	5	1	1	2	5	3	17
50-54	2		1	6	8	7	24
55-59	1	1	2	3	8	4	19
60-64		4	3	2	7	5	21
65-69	1	3			3	1	8
70-74				4	1		5
Grand Total	**12**	**10**	**12**	**18**	**42**	**23**	**117**

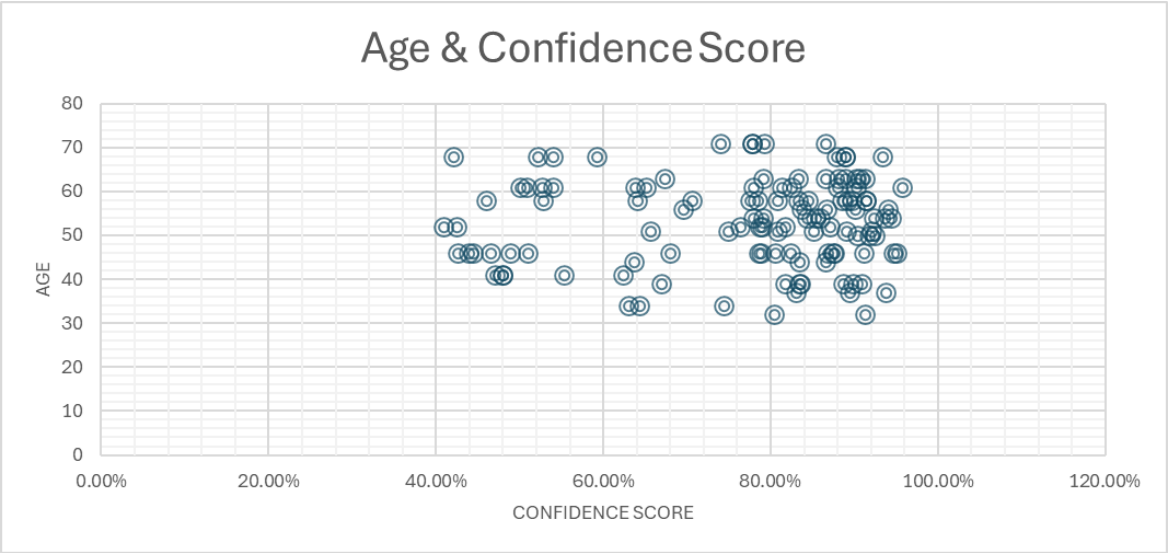

Fig. 15. RWD – Scatter Plot (Age Group versus Confidence Score)

A further analysis was conducted to establish correlation between patient's chronic condition, if any, to the LLM Embeddings confidence/similarity score. Table 7. presents blinded patient records from the RWD dataset and Fig.16. is a scatter plot for the same dataset. As is evident from the table and the corresponding scatter plot, almost 75% of the patients neither had any chronic conditions nor known drug allergies (NKDA). 25% of the patients had a chronic condition such as Hypertension or a combination of Hypertension (HTN) and Diabetes Mellitus (DM). As an inference, no concrete hypothesis can be established that malignancy can be necessarily associated with any other chronic condition as a combination.

Table 7. Chronic condition v/s LLM Embeddings Confidence Score

Chronic Condition	Confidence Score						
	40%-49%	50%-59%	60%-69%	70%-79%	80%-89%	90%-99%	Grand Total
HTN/ DM			2	1	2	2	7
Hypertension	2	4	1	3	9	3	22
NA			1		7	2	10
NKDA			2	5	1		8
#N/A	10	6	6	9	23	16	70
Grand Total	**12**	**10**	**12**	**18**	**42**	**23**	**117**

After analyzing similarity/confidence scores from LLM embeddings for malignancy detection, future research could expand to segmenting patients by age groups. This approach could help identify which age groups are more susceptible to malignancy, allowing for targeted prevention and early intervention. Additionally, the study could explore differences in treatment responses across age groups, leading to personalized recommendations for medical treatment, lifestyle adjustments, dietary interventions, and preventive strategies to improve patient outcomes [31, 32].

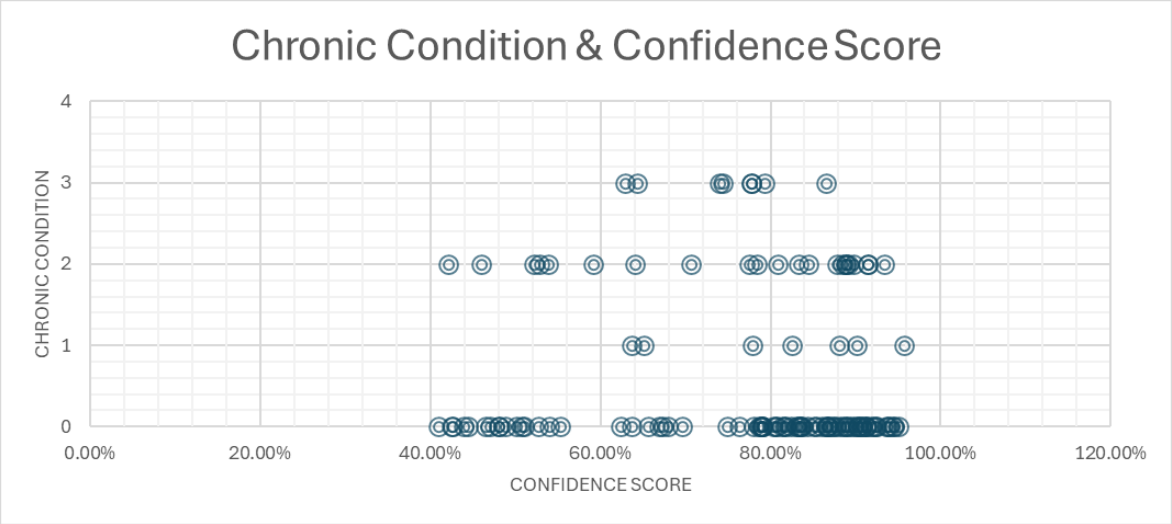

Fig. 16. RWD – Scatter Plot (Chronic condition versus Confidence Score)

5. Conclusion

The proposed model (ICS and IAVA) that integrates embeddings from different types of data sources—such as combining image embeddings with clinical data embeddings like patient history, demographics, geography, etc. to create a comprehensive, multi-modal representation enhances the classification accuracy and provides richer insights by leveraging contextual information from multiple sources. This is particularly novel since the model significantly improves performance compared to traditional single-modal embeddings and offers a practical solution for complex diagnostic scenarios. It's an adaptive clustering framework that dynamically updates its clustering structure based on new incoming data.

Unlike traditional classification models that offer a static approach to image analysis, which lacks the flexibility to dynamically filter results based on specific attributes such as age group, geography, or other demographic factors, The proposed approach in this paper leverages advanced vector embeddings stored in a high-dimensional vector database. This system allows for sophisticated similarity searches that can be tailored to various contextual filters. By applying attribute-based filters—such as age group, geographical location, or other relevant demographic details—search results can be refined to identify the closest matches more precisely. This capability provides enhanced customization and relevance in the classification process, facilitating targeted analysis and improving the accuracy of diagnostic insights based on specific patient profiles or contextual parameters. This level of flexibility and adaptability represents a significant advancement over traditional methods, offering personalized and context-aware classification that is directly aligned with the needs of diverse patient populations.

Furthermore, the model has a capability to establish disease progression in any individual over time, from the first-time screening at day 1 (year 1) to day n (year n), thereby enabling early detection and diagnosis of certain conditions pro-actively. Given the Indian context, people do not visit hospitals for regular check-ups unlike western countries where this is an established norm. In India, hospital visits and check-ups are only done when a problem surfaces and sickness is determined. Therefore, availability of datasets from an early stage prior to onset of a disease to actual progression over time is limited or non-existent. These datasets may be available in western countries and our model can be validated/tested as a part of any 'future-studies' to determine its accuracy and consistency including application of LLMs with other advanced models, such as Vision Transformers or hybrid models combining CNNs with LLMs. Additionally, the learnings from historical datasets will enable the proposed model to provide potential recommendations/suggestions for line of treatment, lifestyle and dietary controls, etc. to the patients which can be used by the Oncologists at their discretion. This is a true example of how a Human + Machine can work in tandem delivering outcomes that have never been imagined in the healthcare industries.

References

[1] Ferlay J, Colombet M, Soerjomataram I, Parkin DM, Piñeros M, Znaor A, Bray F. Cancer statistics for the year 2020: An overview. Int J Cancer. 2021 Apr 5. doi: 10.1002/ijc.33588. Epub ahead of print. PMID: 33818764

[2] Ferlay J, Ervik M, Lam F, Laversanne M, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F (2024). Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. Available from: https://gco.iarc.who.int/today.

[3] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021 Feb 4. doi: 10.3322/caac.21660. Epub ahead of print. PMID: 33538338.
[4] Roy, J.K, Singh Deo, S.K., Yadav, D., Naik, R. A., Sonkar, V. K., Singh, B., Ministry of Health and Family Welfare, Department of Health and Family Welfare, & Prof. Singh Baghel, S.P. (2024). Unstarred Question no. 1227. In Lok Sabha [Report]
[5] Ahmed MI, Spooner B, Isherwood J, Lane M, Orrock E, Dennison A. A Systematic Review of the Barriers to the Implementation of Artificial Intelligence in Healthcare. Cureus. 2023 Oct 4;15(10):e46454. doi: 10.7759/cureus.46454. PMID: 37927664; PMCID: PMC10623210.
[6] Deng, L., Wang, T., Yangzhang, N., Zhai, Z., Tao, W., Li, J., Zhao, Y., Luo, S., & Xu, J. (2024). Evaluation of large language models in breast cancer clinical scenarios: a comparative analysis based on ChatGPT-3.5, ChatGPT-4.0, and Claude2. *International Journal of Surgery*, *110*(4), 1941–1950. https://doi.org/10.1097/js9.0000000000001066
[7] Ahn JS, Shin S, Yang SA, Park EK, Kim KH, Cho SI, Ock CY, Kim S. Artificial Intelligence in Breast Cancer Diagnosis and Personalized Medicine. J Breast Cancer. 2023 Oct;26(5):405-435. doi: 10.4048/jbc.2023.26.e45. PMID: 37926067; PMCID: PMC10625863.
[8] Quazi S. Artificial intelligence and machine learning in precision and genomic medicine. Med Oncol. 2022 Jun 15;39(8):120. doi: 10.1007/s12032-022-01711-1. PMID: 35704152; PMCID: PMC9198206.
[9] Varsha Nemade, Vishal Fegade, Machine Learning Techniques for Breast Cancer Prediction, Procedia Computer Science, Volume 218, 2023, Pages 1314-1320, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2023.01.110.
[10] Khalid A, Mehmood A, Alabrah A, Alkhamees BF, Amin F, AlSalman H, Choi GS. Breast Cancer Detection and Prevention Using Machine Learning. Diagnostics (Basel). 2023 Oct 2;13(19):3113. doi: 10.3390/diagnostics13193113. PMID: 37835856; PMCID: PMC10572157.
[11] Yamashita, R., Nishio, M., Do, R.K.G. *et al.* Convolutional neural networks: an overview and application in radiology. *Insights Imaging* 9, 611–629 (2018). https://doi.org/10.1007/s13244-018-0639-9
[12] Alzubaidi, L., Zhang, J., Humaidi, A.J. *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 8, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8
[13] Ahmed, S.F., Alam, M.S.B., Hassan, M. *et al.* Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artif Intell Rev* 56, 13521–13617 (2023). https://doi.org/10.1007/s10462-023-10466-8
[14] Besen, S. (2024, March 14). LLM Embeddings — Explained Simply - AI Mind. *Medium*. https://pub.aimind.so/llm-embeddings-explained-simply-f7536d3d0e4b
[15] Saleem, A. (2024, April 29). Explore the Role of Vector Embeddings in Generative AI. *Data Science Dojo*. https://datasciencedojo.com/blog/vector-embeddings-generative-ai/
[16] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., The University of Sydney, University of Engineering and Technology (UET), Lahore, Pakistan, The Chinese University of Hong Kong (CUHK), HKSAR, China, University of Technology Sydney (UTS), Sydney, Australia, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Sydney, Australia, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia, SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRCAI), Dhahran, Saudi Arabia, The University of Melbourne (UoM), Melbourne, Australia, Australian National University (ANU), Canberra, Australia, The University of Western Australia (UWA), Perth, Australia, Barnes, N., & Mian, A. (2024). A Comprehensive Overview of Large Language Models. In *Preprint*.
[17] *Demystifying Embedding Spaces using Large Language Models*. (n.d.). https://arxiv.org/html/2310.04475v2
[18] Monir, S. S., Lau, I., Yang, S., & Zhao, D. (2024). VectorSearch: Enhancing Document Retrieval with Semantic Embeddings and Optimized Search. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2409.17383
[19] Dandamudi, H. (2025, January 26). Vector databases: data storage, querying, and embeddings. The ByteDoodle Blog. https://blog.bytedoodle.com/vector-databases-data-storage-querying-and-embeddings/
[20] PingCAP. (2024, July 16). *Understanding the Cosine Similarity Formula*. TiDB. https://www.pingcap.com/article/understanding-the-cosine-similarity-formula/#:~:text=It%20quantifies%20how%20closely%20two,image%20recognition%2C%20and%20recommendation%20systems.
[21] *The Role of Cosine Similarity in Vector Space and its Relevance in SEO*. (n.d.). https://marketbrew.ai/a/cosine-similarity
[22] Grønne, M. (2022, October 18). Introduction to Embedding, Clustering, and Similarity. *Medium*. https://towardsdatascience.com/introduction-to-embedding-clustering-and-similarity-11dd80b00061
[23] S. Bhattacharjee *et al*., "Cluster Analysis: Unsupervised Classification for Identifying Benign and Malignant Tumors on Whole Slide Image of Prostate Cancer," *2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS)*, Genova, Italy, 2022, pp. 1-5, doi: 10.1109/IPAS55744.2022.10052952.
[24] Goncalves, D. (2024, October 30). ML vs. LLM: Is one “better” than the other? | Superwise ML Observability. *Model Observability Platform | Observe, Monitor & Improve ML*. https://superwise.ai/blog/ml-vs-llm-is-one-better-than-the-other/#:~:text=While%20LLMs%20shine%20in%20generative,efficiency%20and%20lower%20resource%20demands.
[25] Lin, Ting-Yu; Huang, Mei-Ling (2020), “Dataset of Breast mammography images with Masses”, Mendeley Data, V2, doi: 10.17632/ywsbh3ndr8.2
[26] Mammography Dataset from INbreast, MIAS, and DDSM. (2024, May 31). Kaggle. https://www.kaggle.com/datasets/emiliovenegas1/mammography-dataset-from-inbreast-mias-and-ddsm/data
[27] *OpenAI Platform*. (n.d.). https://platform.openai.com/docs/guides/prompt-engineering
[28] *OpenAI Cookbook*. (n.d.). https://cookbook.openai.com/
[29] *Understanding LLM Embeddings: A Comprehensive Guide*. (n.d.). https://irisagent.com/blog/understanding-llm-embeddings-a-comprehensive-guide/
[30] Khawaja, R. (2024, October 15). Embeddings 101: The foundation of large language models. *Data Science Dojo*. https://datasciencedojo.com/blog/embeddings-and-llm/
[31] LLM-driven Multimodal Target Volume Contouring in Radiation Oncology. (n.d.). https://arxiv.org/html/2311.01908v3

[32] Oh, Y., Park, S., Byun, H.K. *et al*. LLM-driven multimodal target volume contouring in radiation oncology. *Nat Commun* 15, 9186 (2024). https://doi.org/10.1038/s41467-024-53387-y

[33] DeepranjanG.(n.d.). *Breast_Cancer_Classification/Breast_Cancer_Classification.ipynb at main DeepranjanG/Breast_Cancer_Classification*. GitHub. https://github.com/DeepranjanG/Breast_Cancer_Classification/blob/main/Breast_Cancer_Classification.ipynb

Authors' Profiles

Shobhit Shrotriya is a Managing Director at Accenture and is the Global Life Sciences R&D Operations Lead. He is an Engineer by training with a Masters in Industrial & Management Engineering from India's premier institute – Indian Institute of Technology, Kanpur. Currently he is pursuing his Ph.D. in Data Science from CHRIST (Deemed to be University), Bangalore Central Campus, Hosur Road, Near Dairy Circle, Bangalore. Shobhit has a total of 22+ years of industry experience and has worked in several industries such as Automatove, Consumer Goods, Software and Life Sciences. He is currently a board member of one of the most prestigious industry consortium – Society for Clinical Data Management (SCDM). He has has been a speaker at several international conferences and contributes actively to the industry.

Nizar Banu P K is an Associate Professor at Department of Computer Science, CHRIST (Deemed to be University), Bangalore Central Campus, Hosur Road, Near Dairy Circle, Bangalore. She is a Gold Medalist in Master of Philosphy Computer Science from Periyar University and has earned several other degrees such as Master of Science, Master of Computer Application and Ph,D. Nizar has written and contributed to several research papers, presented and participated in several conference.

Avi Kulkarni is the chief executive of ThoughtSphere. Before ThoughtSphere he was SBU head of Cognizant's life sciences business and Accenture's clinical business. Avi has been at the forefront of the drugs and diagnostics industry for three decades, managing clinical trials, launching new diagnostic tests, building cellular medicine capabilities, and publishing extensively in personalized medicine. Avi's prior roles included: SVP and leader of IQVIA, a clinical research organization, head of Life Sciences R&D at Booz & Co., a management consulting company, and CXO roles at multiple biotechnology start-ups spun out from Stanford University and funded by Sand Hill Road venture capital. Avi currently serves on the Board of Directors of Diaceutics. Avi is a pharmacist who holds a Ph.D. degree in pharmaceutics from Temple University, Pennsylvania, and an MBA from Stanford University. He lives in the San Francisco Bay area with his wife and daughter.

Vinod G. Kumar is an Associate Vice President at Accenture. He received his Bachelor of Science degree from Kristu Jayanti College, Bangalore. He is an accomplished Technology Leader and Software Architect dedicated to optimizing and enhancing Life Sciences R&D Operations at Accenture. In his current role, he spearheads the development of advanced software applications, harnessing the power of automation, workflows, Artificial Intelligence, Machine Learning, and Data Analytics. His commitment lies in delivering bespoke software solutions that elevate client business value. Specializing in Artificial Intelligence and Data Analytics, he consistently led teams to create impactful results.

How to cite this paper: Shobhit Shrotriya, Nizar Banu P. K., Avi Kulkarni, Vinod G. Kumar, "Application of Large Language Models for Data-Driven Analytics in Oncology: Insights and Evidence Generation from Real-World Imaging Data", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.17, No.6, pp. 38-59, 2025. DOI:10.5815/ijigsp.2025.06.03