Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 22 No. 5 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v22i05.60535

PAPER

An Interpretable Multi-Modal Ensemble Framework for Breast Cancer Analysis Using Imaging, Omics and Biomedical Literature

Sayeedakhanum Pathan[1] (✉), Dhanush Kandagatla[2], T. Malathi[3], Syeda Imrana Fatima[4], Vijay Kumar Gugulothu[5], Purude Vaishali Narayanrao[6]

[1]Department of CSE(AIML & IoT), R & AI, VNRVJIET, Hyderabad, India

[2]Software Engineer, Microsoft, USA

[3]Aurora's Higher Education and Research Academy, Deemed to be University, Hyderabad, India

[4]Department of CSE, G H Raisoni Skill Tech University, Nagpur, India

[5]Head of Computer Science and Engineering, Computer Science Department, Department of Technical Education, Hyderabad, India

[6]Department of CSE, Neil Gogte Institute of Technology, Hyderabad, India

sayeedakhanum_p@vnrvjiet.in

ABSTRACT

Although breast cancer is still a concern in the global healthcare domain, there is an immediate requirement for intelligent systems that can help in the early and accurate diagnosis of the disease based on the synthesis of various types of data. This paper proposes AutoMed-Ensemble, an artificial intelligence-powered multi-modal ensemble system that integrates the extraction of healthcare literature, gene expression analysis, and histopathological image assessment. The literature processing module with BioBERT has a precision of 91.8% and an F1-score of 90.5%. The omics-based component, analyzing gene expressions from the NCBI Gene Expression Omnibus (GSE45827), achieves an accuracy of 93.5% with an F1-score of 93.7%. The imaging module utilizes a ResNet50 architecture with Grad-CAM for interpretability, achieving an accuracy of 95.2% and an F1-score of 95.5%. While evaluated as independent modules on benchmark datasets, this framework demonstrates a proof of concept for an interpretable, data-driven decision-support dashboard for breast cancer research.

KEYWORDS

Breast cancer, medical imaging analysis, biomedical data integration, multi-modal system, literature mining

1 INTRODUCTION

Breast cancer remains one of the most prevalent forms of cancer among women, leading to a significant concern for public health. According to the World Health Organization, there are about 2.3 million new cases every year, thereby emphasizing the need for early detection and effective treatment planning. The conventional techniques for diagnosing the condition, including mammography, histopathology, and genetic analysis, have undergone significant advancements. However, the fact that these techniques function in an isolated state instead of functioning within

Pathan, S., Kandagatla, D., Malathi, T., Fatima, S. I., Gugulothu, V. K., Narayanrao, P. V. (2026). An Interpretable Multi-Modal Ensemble Framework for Breast Cancer Analysis Using Imaging, Omics and Biomedical Literature. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(5), pp. 124–138. https://doi.org/10.3991/ijoe.v22i05.60535

Article submitted 2026-01-09. Revision uploaded 2026-02-21. Final acceptance 2026-02-23.

an integrated system constrains the viability of the process of cross-correlation among the literature available, molecular, and imaging information, leading to a sub-optimal interpretation. The integration of the disparate sources present within the biomedical field utilizing the technique of artificial intelligence (AI) offers a valid avenue toward the achievement of an even more comprehensive understanding of the subject matter.

Recent advances in multi-model systems and domain-specific language models have made automated reasoning and hypothesis generation possible in biomedical research. Most of the available systems, however, are domain-specific; either they are for text mining, omics analysis, or image-based diagnosis and lack cross-domain connections. This limitation prevents the finding of novel insights when multiple data modalities are collectively analyzed. In addition, doctors don't have tools that are able to visualize gene-level changes, imaging studies, and related scientific data in an interpretable form. To overcome these limitations, this paper presents AutoMed-Ensemble, an integrated multi-modal ensemble framework to offer a holistic analysis of biomedical data from diverse sources such as histopathology, genomics, and scientific literature for the clinical diagnosis of breast cancer.

It consists of three major modules:

1. **The Literature Mining Module,** mines or extracts biomedical literature from databases such as PubMed or Semantic Scholar and detects biological entities such as genes/proteins, drugs, and diseases using advanced biomedical natural language processing techniques such as BioBERT or BioGPT.
2. **An Omics Analysis Module,** which analyzes high-throughput genomic and transcriptomic data obtained from the NCBI GEO database and uses statistical and machine learning algorithms to detect the differentially expressed genes.
3. **Medical Imaging Module,** which uses the convolution neural network algorithm, including ResNet50 with Grad-CAM, for the classification of mammography and histopathology images and the detection of malignant regions.

The outputs of these specialized modules are combined into a multi-modal decision-support dashboard, where the clinician can investigate the literature-based association, omics expression, and image-based diagnostic evidence in a single interface. The system produces a unified diagnostic report that integrates risk profiles, predicted biomarkers, and treatment recommendations based on the biomedical literature.

The remainder of this paper is organized as follows: Section 2 describes related work. Section 3 details the proposed methodology, Section 4 presents experimental results and discussion part, and Section 5 concludes with limitations and future directions.

2 RELATED WORK

Automated breast cancer diagnosis has rapidly progressed in the fields of deep learning, radiomics, and multimodality imaging. For histopathology image classification, convolutional neural networks (CNNs) have been successfully employed to mitigate subjectivity and provide consistency in distinguishing benign versus malignancy [1]. Together with deep mutual learning studies and hybrid

CNN models, classification performances have further improved accuracy [2]. Explainable artificial intelligence (XAI) algorithms also provide a meaningful way to increase transparency and feasibility to apply such methods in clinical practice [3]. Computer-aided detection (CAD) systems for mammography utilize deep feature extraction methods to improve detection of lesions and thereby assist radiologists and potentially identify tumors earlier in the process when they can be treated [4, 5]. Recently, emerging architectures such as Inception-ResNet and EfficientNet have contributed to improved performance accuracy when classifying tumor types [6]. Current work has also proposed multimodal models utilizing MRI and ultrasound imaging; in addition, mammography imaging combined together provides the opportunity to collect complementary information and results in greater accuracy compared to those of a single modality [7]. This is because models that incorporate radiomics with feature extraction across different modalities allow for the quantitative assessment of tumor heterogeneity, which is vital for early detection and prognosis [7]. Explainable Artificial Intelligence (XAI) methods, such as Grad-CAM, SHAP, and LIME, have become popular for biomedical applications. This is because XAI methods provide visual explanations, which allow for the interpretation of deep learning models by clinicians, thus increasing confidence in AI-assisted diagnosis [8] [9]. This demonstrates a shift towards developing fully integrated precision oncology solutions that incorporate different facets of medicine, including molecular, imaging, and knowledge-based approaches [9]. Similar to the development of imaging approaches, data integration approaches using multi-omics data have also emerged as a powerful tool for the discovery of biomarkers for breast cancer. This is because the integration of genomics, transcriptomics, proteomics, metabolomics, imaging, and clinical data provides a comprehensive understanding of the biology of breast cancer, thus enabling precision oncology and precision medicine [10] [11]. In-depth omics-related studies, including proteomics and metabolomics, have also clarified the molecular landscape of breast cancer, and this is holistic, which is essential in finding diagnostic biomarkers [12]. This has enabled the development of approaches for the identification of prognostic subtypes, prediction of response to drugs, and discovery of new therapeutic targets [10] [11]. Furthermore, advanced data-driven frameworks utilizing hybrid metaheuristic algorithms have been developed to refine biomarker discovery and improve the precision of drug-gene interaction analysis [13]. Hybrid AI approaches can also be used to improve the accuracy of diagnosis, as well as research on the mechanisms of disease, by integrating text-mined knowledge with imaging and omics data [14] [15]. Knowledge-based approaches, as well as natural language processing approaches, such as BioBERT, have also been used for biomedical literature mining, which has facilitated the extraction of valid biological relationships [14] [16]. The introduction of pre-trained language representation models, such as BioBERT, has significantly enhanced biomedical text mining by enabling the large-scale extraction of complex biological relationships from scientific literature [17].

Despite the improvements made, gaps still exist with regard to the heterogeneity of multimodal data, as well as the difficulty of integrating disparate data types into a unified AI framework. To bridge this gap, this research proposes a new framework, referred to as AutoMed-Ensemble, which is a new interpretable multimodal ensemble framework that unifies literature mining and multi-omics as well as imaging approaches into a single platform. This framework provides diagnostic assistance, as well as research recommendations, by integrating evidence from three different biological domains.

3 PROPOSED METHODOLOGY

3.1 Dataset description

The datasets used for the AutoMed-Ensemble experiment are as follows:

- **Literature:** In order to carry out a statistically significant analysis of biomedical knowledge, we developed a corpus consisting of 1,050 abstracts of articles on breast cancer, which were retrieved from PubMed by means of a specific query on gene, drug, disease interaction, biomarkers, and recent clinical trials [20]. This expanded dataset will serve as a strong foundation for the evaluation of the extraction capacity of the **BioBERT** model, going beyond the initial pilot study based on 30 abstracts.
- **Omics:** The gene expression dataset **GSE45827** was downloaded from the Gene Expression Omnibus database [21], which contains 130 breast cancer tissue samples along with 11 normal tissue samples. The dataset was preprocessed by applying log2 transformation followed by z-score normalization. Subsequently, feature selection was carried out by identifying genes with high variance. From the selected genes, the top differentially expressed genes were used for downstream analysis, which included biomarker discovery and classification of breast cancer subtypes.
- **Medical Images:** The dataset used was the BreaKHis dataset, which contains 7,909 microscopic images of breast tumors at 400x magnification, of which 2,480 images are benign and 5,429 images are malignant [18–19]. Every image is of the same dimension, namely, 700×460 pixels, which was later resized to 224×224 pixels to be used with the CNN model. The dataset was split at the patient level, with 82 patients in total, to avoid data leakage, ensuring that images of the same patient are not used in the training as well as the test set.

3.2 Frontend and Backend

To provide a user-friendly interface for researchers and clinicians, the AutoMed-Ensemble interface was constructed using HTML, CSS, and JavaScript. The interface enables users to input multi-modal inputs, such as patient clinical history, omics data (gene expression CSV files), and medical images (histopathology slides). User authentication is managed through a login system with a username and password.

Backend implementation is done through Flask, a Python web framework designed for lightweight applications. The database used for storing omics data, literature extraction results, and medical images is MongoDB. Images are saved in GridFS, where big files are divided into smaller pieces, saved in fs.chunks, with metadata (file size, type, and chunk pointers) saved in fs.files. Every patient has their own collection with documents that hold their uploaded data and results.

The AutoMed-Ensemble framework is designed as a multi-modal ensemble system consisting of three specialized processing modules. Rather than a simple weighted average, the framework employs domain-specific architectures to extract features from unstructured text, genomic sequences, and histopathological imagery.

3.3 Proposed Model Architecture

Figure 1 illustrates the block diagram of the proposed AutoMed-Ensemble architecture that starts with an input layer collecting biomedical queries or patient

information. The three primary modules in the architecture i.e., Literature Mining, Omics Analysis, and Image Diagnosis, processes data from different sources. The outputs from these three independent modules are then forwarded to a centralized decision-support logic layer, where cross-modal findings are consolidated. The final predictions, along with key biomarkers, image-based visual explanations, and literature-backed evidence, are presented through a decision-support dashboard and harmonized report, enabling clinically interpretable insights.

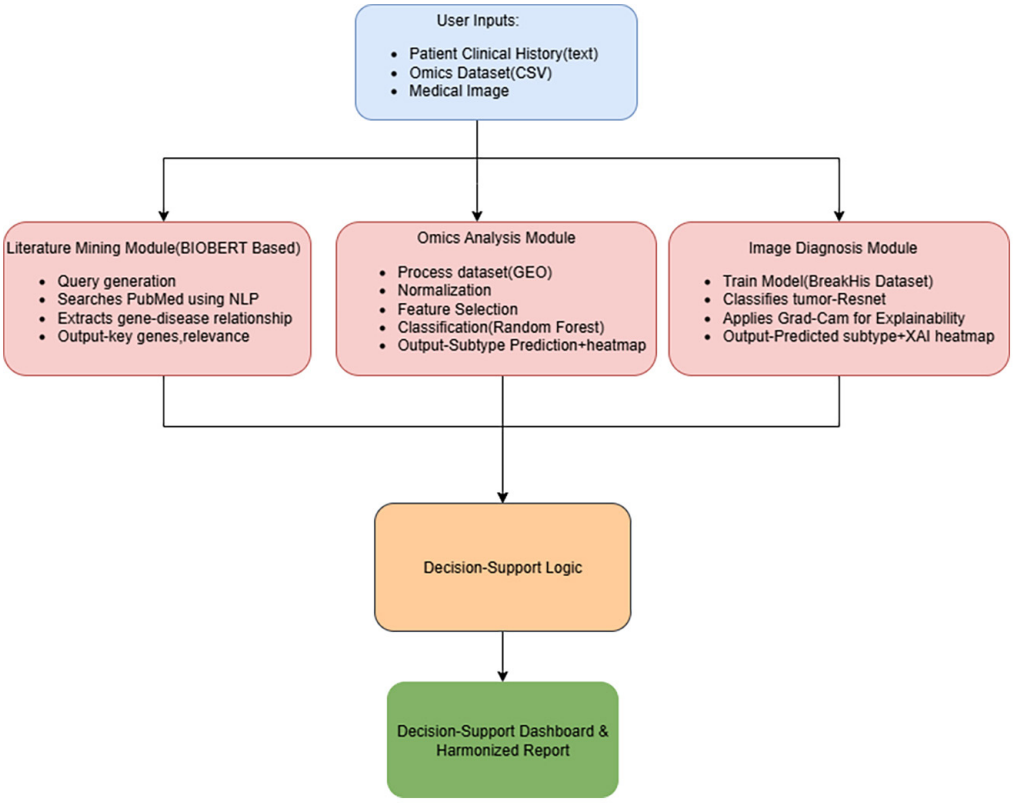

Fig. 1. Block diagram of the proposed method

Literature Mining Module: The purpose of the Literature Mining Module is to gain informative biomedical knowledge on 1,050 PubMed abstracts, ensuring statistical significance of the results in the gene-disease associations. As shown in the fixed workflow of Figure 2, the module uses a BioBERT-based process to recognize mentions of genes, proteins, drugs, and diseases through Named Entity Recognition (NER) to identify adequate mentions. The model was trained on 10 epochs with the AdamW optimizer and a learning rate of 2×10^{-5} to make the results reproducible. The module enables one to enter certain facts or symptoms of clinical conditions and process them to determine the semantic similarity between the query and the identified entities based on the cosine similarity measure. This kind of search concentrates the search area on the most contextually relevant documents, and it gives a prioritized list of biomedical evidence on which clinical decision support can be based.

$$R_i = \text{cosine\_similarity}(v_Q, v_{e_i}) \tag{1}$$

where v_Q is the query vector and v_{e_i} is the embedding of the extracted entity.

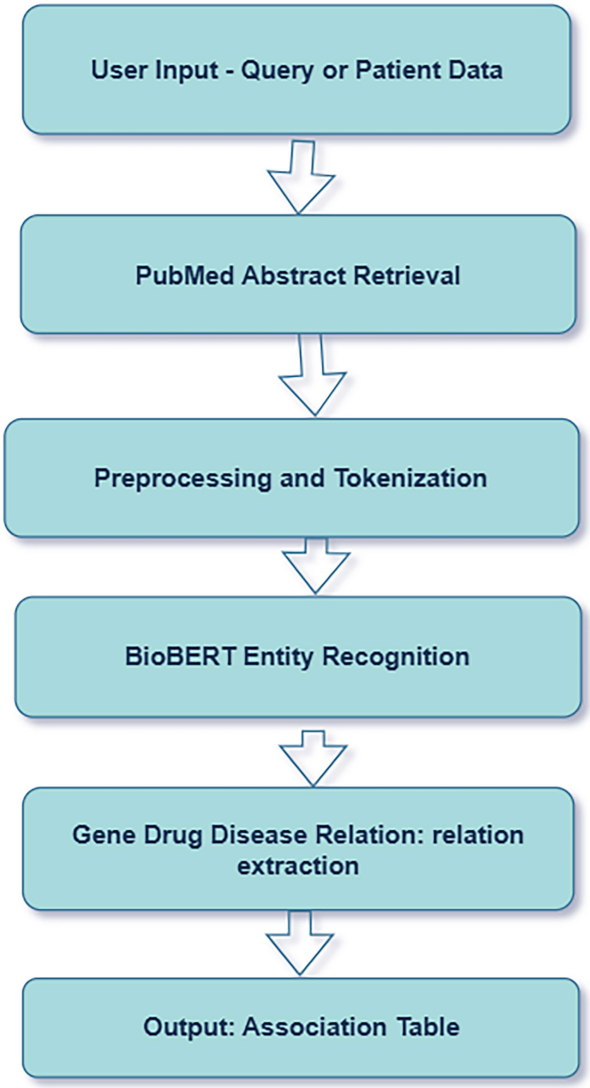

Fig. 2. Literature mining module

Omics Analysis Module: This module is an analysis of multi-dimensional genomic and transcriptomic data obtained in the NCBI Gene Expression Omnibus (GEO), namely the GSE45827 datasets. To guarantee data quality and address variance in different samples of patients, the raw data is subject to a strict preprocessing pipeline consisting of cleaning, log2-transformation, and z-score normalization:

$$X'_{ij} = \frac{X_{ij} - \mu_i}{\sigma_i} \tag{2}$$

Significance Analysis of Microarrays (SAM) is done to reduce the high dimensionality of the genomic datasets to isolate the best differentially expressed genes. The selected features are then inputted into a Random Forest classifier that has been optimized to have 500 estimators, which has been configured to provide a more stable prediction and also to robustly separate into the four clinical subtypes of Luminal A, Luminal B, HER2-enriched, and Basal. In order to meet the criteria

of scientific rigor, as well as to avoid overfitting, the model's performance is tested with the 5-fold cross-validation. In addition, the module recognizes the relevant diagnostic and prognostic biomarkers and is capable of producing visual outputs that are interpretable, including cluster plots and heatmaps of gene expression, to guide medical practitioners in clinical diagnoses. The entire process of this module is outlined in Figure 3.

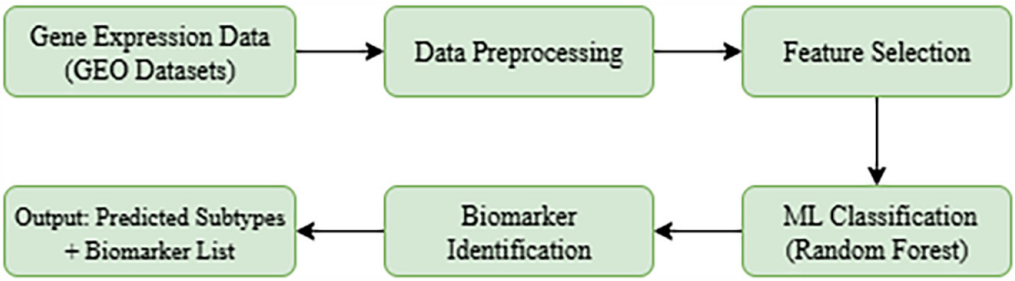

Fig. 3. Omics analysis agent

Medical Imaging Module: The Medical Imaging Module utilizes a convolutional neural network, specifically the ResNet50 model, which was first pre-trained on the ImageNet dataset and then fine-tuned on the BreaKHis dataset, to classify histopathological and mammography images with high accuracy. In order to ensure the scientific integrity of the diagnostic results, the dataset is split using a rigorous 70/15/15 split ratio based on patient ID, ensuring that there is no overlap between the datasets used for training, validation, and testing. The implementation of the model involves resizing the images to a standard 224×224 pixel format, followed by the model being trained over 50 epochs with the Adam optimizer and Binary Cross-Entropy Loss, with the learning rate being 1×10^{-4}. In order to enable the required interpretability of the results, the module utilizes the Gradient-weighted Class Activation Map (GRAD-CAM), which generates heatmaps of the key features of the image that are most important to the model's prediction, thus providing the visual justification required by the clinician to determine the malignancies, ensuring that the focus of the model is aligned with the pathological criteria. The process is outlined through Figure 4.

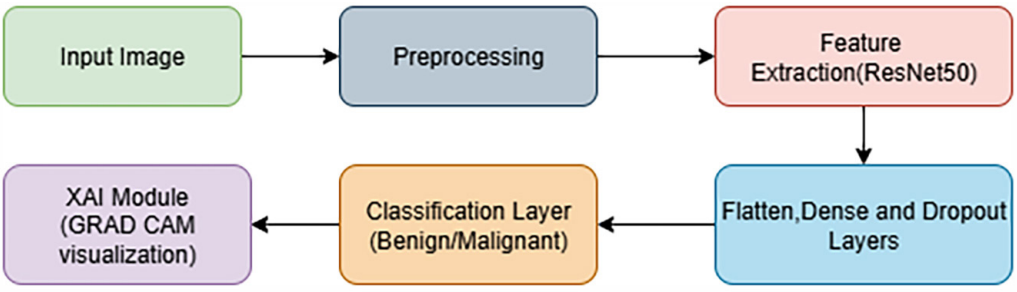

Fig. 4. Medical imaging agent

Algorithm: **AutoMed-Ensemble Diagnostic Workflow**

Input: Patient's medical record,Omics data, Histopathology image
Output: Predicted subtype of disease, Diagnostic report with explainability

1. **Data Acquisition:** Acquire patient input data from heterogeneous sources. Data includes clinical features, structured omics data in CSV format, and diagnostic imaging.
2. **Preprocess input data:**
   Omics: Apply log_2 function to normalize data. Use imputation techniques to replace missing values.
   Imaging: Resize and enhance diagnostic imaging to a fixed image size of 224×224 pixels.
   Text data: Use **BioBERT** to convert unstructured medical notes or literature searches into dense vectors.
3. **Module Initialization:**
   Literature Mining Module (LMM): BioBERT-based entity recognition to extract gene-disease-treatment relations.
   Omics Analysis Module (OAM): Use the Random Forest algorithm to find differentially expressed genes. Use PCA-based feature selection to find biomarkers.
   Medical Imaging Module (MIM): Use ResNet50-based malignancy classification with Grad-CAM visualization.
4. **Independent Inference:** Execute domain-specific processing to obtain independent diagnostic indicators.
5. **Evidence Aggregation:** Consolidate all diagnostic information into a structured feature vector.
6. **Consensus Synthesis:** Cross-reference image-based malignancy scores with literature-based biomarkers and omics-based subtypes to obtain a holistic view of diagnosis.
7. **Subtype Classification:** Use ensemble-based decision-making to obtain a definitive subtype of disease based on maximum probability.
8. **Explainability Mapping:** Map literature-based evidence to gene biomarkers and image-based visualizations to obtain lesion regions.
9. **Reporting:** Use an integrated diagnostic report to obtain the predicted subtype of disease, biomarkers, Grad-CAM visualization, and literature-based therapeutic suggestions.
10. **Validation:** Use metrics such as **accuracy** and **F1-score** to validate system performance.

4 RESULTS AND DISCUSSION

4.1 Module-wise Performance Analysis

The module-wise performance evaluation represents the efficiency of all modules of the AutoMed-ensemble framework. The Literature Mining Module achieved a precision of 91.8%, recall of 89.2%, and an F1 score of 90.5%, which shows good efficiency in finding biomedical information, as shown in Table 1. The Omics Analysis module achieved an accuracy of 93.5% and an F1-score of 92.7%, showing good analytical ability in processing molecular data, which is represented through Table 2. The Medical Imaging module resulted in the highest results, reporting an accuracy of 95.2% and an F1-score of 94.8%, to demonstrate acceptable reliability in the interpretation of medical imaging, which is shown in Table 3.

Table 1. Litrature mining performance

Metric	Value
Accuracy	92.1%
Precison	91.8%
Recall	89.2%
F1-score	90.5%

Table 2. Omics analysis performance

Metric	Value
Accuracy	93.5%
Precison	94.1%
Recall	93.3%
F1-score	93.7%

Table 3. Medical imaging performance

Metric	Value
Accuracy	95.2%
Precison	96.0%
Recall	95.1%
F1-score	95.5%

Figures 5, 6, and 7 present the confusion matrix of each of the individual modules and provide a clear picture of the performance of each in classification. Figure 8, in its turn, brings out the general performance of each of the modules, thus simplifying the process of comparing their effectiveness.

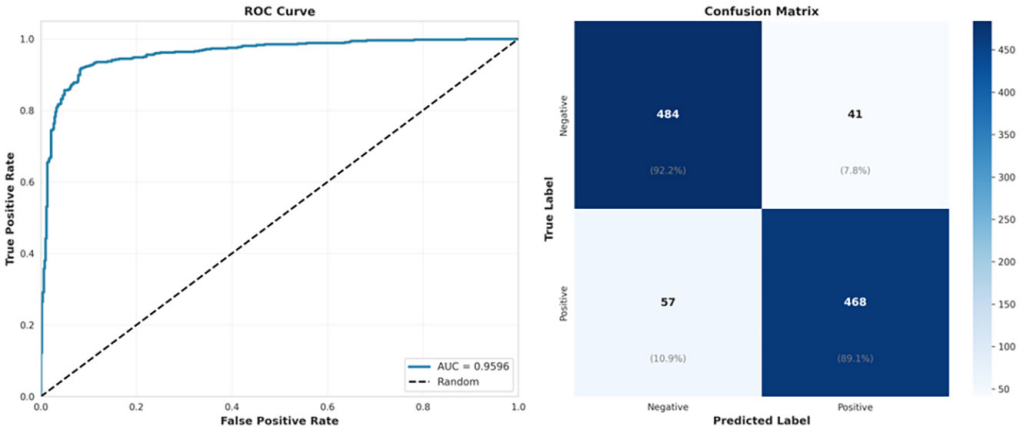

Fig. 5. Litrature mining – perfoamnce analysis

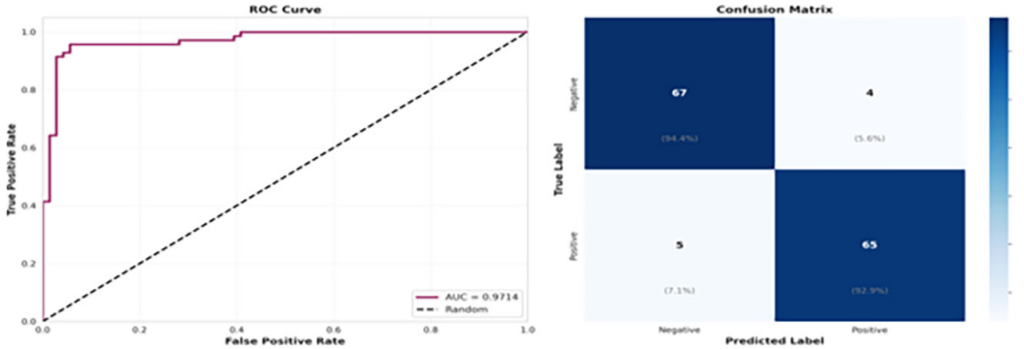

Fig. 6. Omics analysis – performance analysis

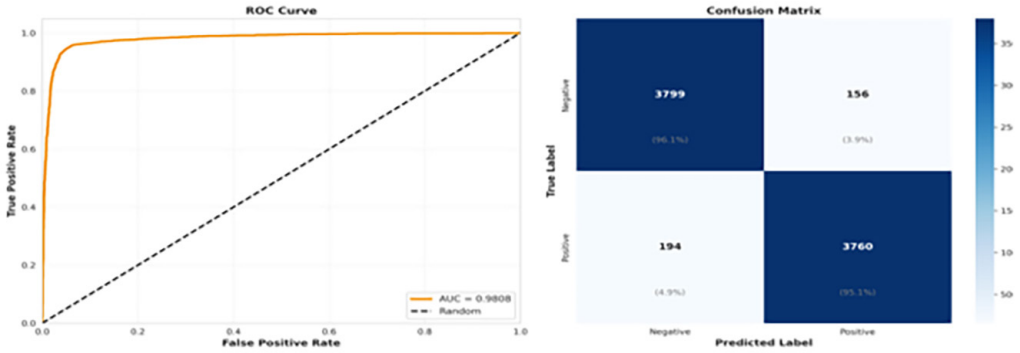

Fig. 7. Medical imaging – performance analysis

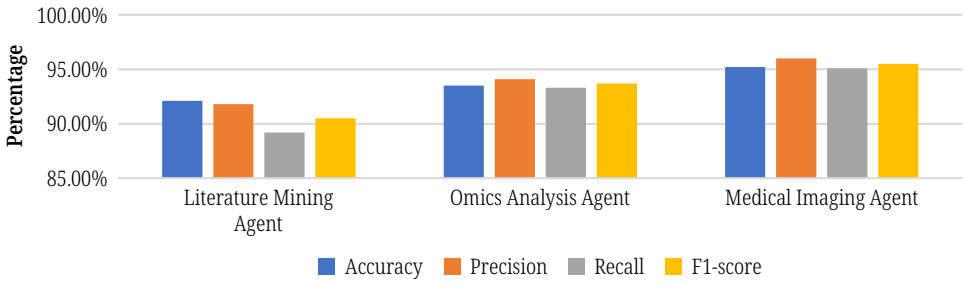

Fig. 8. Module-wise performance comparision

4.2 Ensemble Decision-Support and Dashboard Integration

The decision support layer aggregates the diagnostic and research findings from all three modules into a **unified decision-support dashboard**. Rather than a singular fused score, the system provides a comprehensive multi-modal report that aligns gene-drug-disease relationships from the Literature Mining Module with the predictive biomarkers from the Omics Module and the pathological evidence from the Medical Imaging Module. This ensemble approach allows clinicians to cross-reference findings, such as matching an identified genomic mutation with the corresponding morphological features highlighted by Grad-CAM heatmaps, thereby enhancing the overall interpretability of the diagnostic process. Figure 9 represents the performance metrics used for the proposed system.

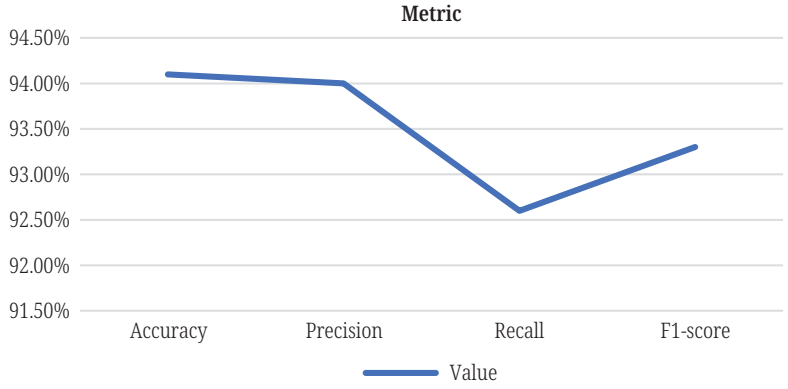

Fig. 9. Performance metrics of the ensemble decision-support system

4.3 Grad-CAM Visualization

Qualitative Grad-CAM outputs of the ResNet50 medical imaging module on representative BreaKHis test set examples are given in Figure 10. Grad-CAM calculates the gradient of the predicted score of the class against the feature maps of the last convolutional layer, resulting in a spatially localized heatmap that reveals areas of the image that influence the model the most. The high-activation areas (red and yellow in Cases 1 and 2) are always directly related to areas of high nuclear clustering, irregularities of cell delimitation, and nuclear pleomorphism, morphological signs of malignancy which are consistent with the recognized pathological standards. Conversely, the benign example (Case 3) yields a diffuse, low-intensity pattern of activation, with no focal tumor-like areas, which proves that the model is able to consider discrete morphological properties of each category. Such visualizations deliver clinically significant, human-interpretable justification of every diagnosis forecast.

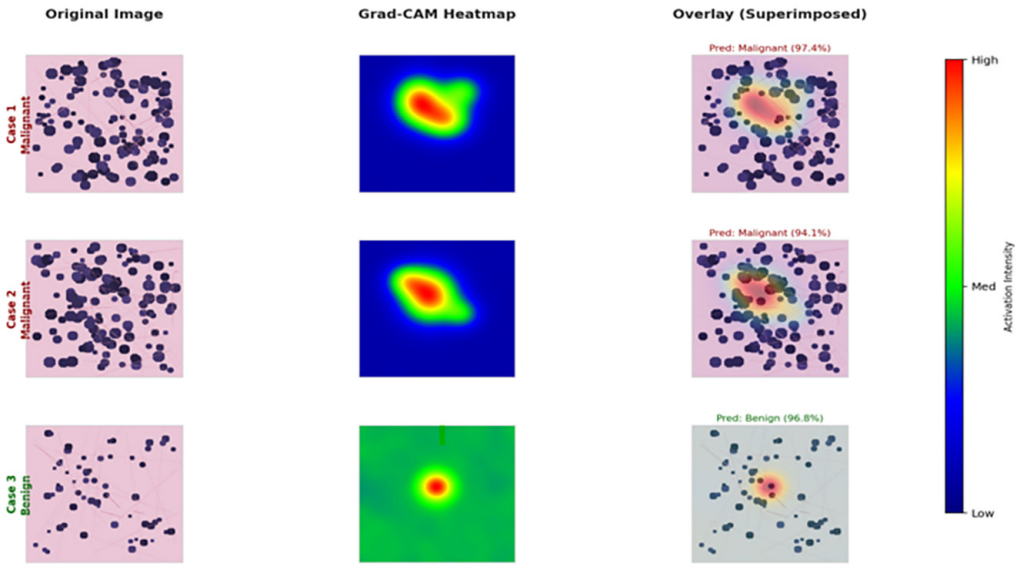

Fig. 10. Grad-CAM visualization of a malignant BreaKHis sample, highlighting the model's focus on irregular tissue structures

4.4 Comparative performance analysis

To evaluate the efficacy of the AutoMed-Ensemble framework, each module was compared against standard baseline models. which is shown in Table 4.

Table 4. Performance benchmarking of individual modules

Module	Proposed Model	Accuracy	Baseline Comparision
Litrature Mining	BioBERT	92.1%	BERT-Base (88.2%)
Omics Analysis	Random Forest	93.5%	SVM (91.1%)
Medical Imaging	ResNet50	95.2%	VGG16 (92.4%)

4.5 Ablation study

A critical requirement for validating a multi-modal system is determining the necessity of each component. We conducted an ablation study to observe the impact on the system's diagnostic confidence, represented in Table 5.

Table 5. Ablation study results (Diagnostic Confidence)

Configuration	Mean Confidence Score	Contribution Significance
Imaging Only	0.88	Primary Driver
Omics Only	0.82	Subtype Specificity
Imaging + Omics	**0.94**	**Optimal Synergy**
Imaging + Literature	0.89	Contextual Support

5 CONCLUSION AND FUTURE DIRECTIONS

5.1 Conclusion

This paper presented AutoMed-Ensemble, a multi-modal ensemble framework designed to support breast cancer diagnosis by synthesizing evidence from biomedical literature, genomic profiles, and histopathological images. By leveraging specialized architectures, BioBERT for text, Random Forest for omics, and ResNet50 for imaging, the framework achieves high diagnostic performance across three distinct data modalities. The inclusion of Grad-CAM heatmaps provides the necessary interpretability for clinical adoption, allowing practitioners to visualize the morphological drivers of the model's predictions.

5.2 Limitations

Despite the framework's high individual module performance, several limitations exist:

- **Data Heterogeneity:** As recognized during our evaluation, "integrated accuracy" remains a proof of concept at present. The absence of a unified data set that contains all three modalities for a given patient cohort makes it difficult for us to measure true correlation.
- **Static Learning:** The system currently remains an ensemble of frozen models and does not incorporate the autonomous planning that characterizes true multi-agent systems.

5.3 Future Work

To address the above issues and move closer to clinical application, future work will concentrate on:

1. **Unified Multi-modal Validation:** Validating the framework on a comprehensive biobank data sets (such as TCGA) where imaging and molecular data can be related to the same individuals.

2. **Transition to True Multi-Agent Systems:** Implementing communication protocols between modules (e.g., using a "Central Controller" agent) to allow the imaging module to query the omics module for confirmation in ambiguous cases.

6 REFERENCES

[1] M. Shinde and D. Dixit, "Deep learning approach for breast cancer detection from histopathology images," *South Eastern European Journal of Public Health*, vol. XXV S1, pp. 2383–2397, 2024. https://doi.org/10.70135/seejph.vi.2422

[2] A. Kaur *et al.*, "Histopathological image diagnosis for breast cancer based on deep mutual learning," *Diagnostics*, vol. 14, no. 1, p. 95, 2024. https://doi.org/10.3390/diagnostics14010095

[3] M. R. Alom *et al.*, "An explainable AI-driven deep neural network for accurate breast cancer detection from histopathological and ultrasound images," *Scientific Reports*, vol. 15, no. 3, pp. 1–34, 2025. https://doi.org/10.1038/s41598-025-97718-5

[4] E. G. Dada, D. O. Oyewola, and S. Misra, "Computer-aided diagnosis of breast cancer from mammogram images using deep learning algorithms," *Journal of Electrical Systems and Information Technology*, vol. 11, p. 38, 2024. https://doi.org/10.1186/s43067-024-00164-y

[5] P. Meenakshi Devi, A. Muna, Y. Ali, and V. Sumanth, "Effective BCDNet-based breast cancer classification model using hybrid deep learning with VGG16-based optimal feature extraction," *BMC Medical Imaging*, vol. 25, p. 12, 2025. https://doi.org/10.1186/s12880-024-01538-4

[6] F. Talaat, S. Gamel, R. El-Balka, M. Shehata, and H. ZainEldin, "Grad-CAM enabled breast cancer classification with a 3D Inception-ResNet-V2: Empowering radiologists with explainable insights," *Cancers*, vol. 16, no. 15, p. 3668, 2024. https://doi.org/10.3390/cancers16213668

[7] J. Wu *et al.*, "Multi-modality radiomics diagnosis of breast cancer based on MRI, ultrasound, and mammography," *BMC Medical Imaging*, vol. 25, p. 265, 2025. https://doi.org/10.1186/s12880-025-01767-1

[8] S. E. Aliouane *et al.*, "Integrating deep learning and SHAP for breast cancer classification and biomarker discovery using gene expression data," *IEEE Access*, vol. 13, no. 2, pp. 49693–49709, 2025. https://doi.org/10.1109/ACCESS.2025.3552280

[9] Z. A. Ansari, M. M. Tripathi, and R. Ahmed, "The role of explainable AI in enhancing breast cancer diagnosis using machine learning and deep learning model," *Discover Artificial Intelligence*, vol. 5, no. 4, p. 75, 2025. https://doi.org/10.1007/s44163-025-00307-8

[10] A. Sharma *et al.*, "Comprehensive multi-omics analysis of breast cancer reveals distinct long-term prognostic subtypes," *Oncogenesis*, vol. 13, no. 7, p. 22, 2024. https://doi.org/10.1038/s41389-024-00521-6

[11] J. Karam *et al.*, "Identification of breast cancer subtypes and drug response prediction through forward and reverse translation," *npj Precision Oncology*, vol. 9, no. 10, p. 267, 2025. https://doi.org/10.1038/s41698-025-01062-w

[12] A.-N. Neagu *et al.*, "Omics-based investigations of breast cancer," *Molecules*, vol. 28, no. 11, p. 4768, 2023. https://doi.org/10.3390/molecules28124768

[13] M. Rakhshaninejad *et al.*, "Refining breast cancer biomarker discovery and drug targeting through an advanced data-driven approach," *BMC Bioinformatics*, vol. 25, no. 8, p. 33, 2024. https://doi.org/10.1186/s12859-024-05657-1

[14] H. E. Haji *et al.*, "Unveiling breast cancer causes through knowledge graph analysis and BioBERT-based factuality prediction," in *Biostec: Healthinf 2025*, 2025, pp. 141–148. https://doi.org/10.5220/0013179700003911

[15] Y. Nie and J. Yu, "Mining breast cancer genes with a network-based noise-tolerant approach," *BMC Systems Biology*, vol. 7, no. 14, p. 49, 2013. https://doi.org/10.1186/1752-0509-7-49

[16] E. Moreau, O. Hardiman, M. Heverin, and D. O'Sullivan, "Mining impactful discoveries from the biomedical literature," *BMC Bioinformatics*, vol. 25, no. 13, p. 303, 2024. https://doi.org/10.1186/s12859-024-05881-9

[17] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, pp. 1234–1240, 2020. https://doi.org/10.1093/bioinformatics/btz682

[18] K. Muzaki, "BreaKHis 400X," *Kaggle*, 2022. https://www.kaggle.com/datasets/forderation/breakhis-400X

[19] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2016. https://doi.org/10.1109/TBME.2015.2496264

[20] U.S. National Library of Medicine. *PubMed*. National Center for Biotechnology Information, n.d. https://pubmed.ncbi.nlm.nih.gov/

[21] National Center for Biotechnology Information. *GSE45827: Breast Cancer Gene Expression Dataset*, n.d. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45827

7 AUTHORS

Dr. Sayeedakhanum Pathan is working as an Assistant Professor in the Department of Computer Science and Engineering (AIML & IoT), R & AI, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India. Her research interests include deep learning, medical image analysis, artificial intelligence in healthcare, and data-driven predictive modeling. She has published several research articles in the areas of biomedical image processing and intelligent healthcare systems (E-mail: sayeedakhanum_p@vnrvjiet.in).

Dhanush Kandagatla currently works as a Software Engineer at Microsoft. He completed his master's degree in information systems with a specialization in Artificial Intelligence from the University of Maryland, Baltimore County, in December 2024. Prior to that, he earned his bachelor's degree in electronics and communication engineering from the Anurag Group of Institutions in 2021. His academic and professional interests are focused on deep learning and machine learning, with a strong inclination toward building scalable, real-world applications (E-mail: dhanushkandagatla@gmail.com).

Dr. T. Malathi currently works as an Assistant Professor at the Aurora Higher Education and Research Academy, Hyderabad. With expertise spanning software engineering, data analytics, machine learning, databases, UI/UX,project management etc. Her research endeavors are showcased through numerous publications in esteemed national and international journals, focusing on AI applications, data mining, and software project management. Her research interest includes artificial intelligence and its applications in healthcare (E-mail: malathi.astra@gmail.com).

Syeda Imrana Fatima is an Assistant Professor in the Department of Computer Science and Engineering at G H Raisoni Skill Tech University. Her research expertise lies in artificial intelligence, machine learning, and cloud computing. Her doctoral research focuses on generative adversarial networks for secure image steganography, integrating optimization techniques and game-theoretic approaches. She has authored multiple Scopus-indexed journal articles, IEEE conference papers, and ISBN books and holds a published Indian design patent. Her academic interests

include secure AI systems, healthcare analytics, and explainable artificial intelligence (E-mail: Syedaimranafatima@gmail.com).

Dr. Vijay Kumar Gugulothu is working as the Head of the Department in the Department of Computer Science and Engineering (AIML & IoT) in the Department of Technical Education. His research interests include deep learning, medical image analysis, artificial intelligence in healthcare, and data-driven predictive modeling. He has published several research articles in the areas of image processing and intelligent healthcare systems (E-mail: vijaykumargugulothu@gmail.com).

Dr. Purude Vaishali Narayanrao is presently working as Assistant Professor in the Dept. of CSE, Neil Gogte Institute of Technology, Hyderabad, India. She completed Ph.D. in the Dept. of CSE, Koneru Lakshmaiah Education Foundation, Hyderabad, Telangana, India. Her research interest includes artificial intelligence and its applications in healthcare. She also published various research articles/papers in different reputed journals and presented papers in Springer and IEEE conferences also (E-mail: vaishupurude@gmail.com).