© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717859

Leveraging Artificial Intelligence in Medical Imaging and Biomarker Analysis for Early and Accurate Cancer Diagnosis

OMEYE EMMANUEL CHIZOBA[1], ODO VINCENTMARY CHUKWUEMEKA[2]
[1]*Department of Computer Science, Tansian University, Umunya.*
[2]*Department of Medical Laboratory Science, Tansian University, Umunya*

Abstract - Early and proper diagnosis of cancer is one of the urgent problems of modern healthcare. This study suggests that an AI-based diagnostic system combines medical imaging with biomarker analysis using a pure experimental design technique. The imaging information was retrieved using Cancer Imaging Archive (TCIA), whereas the biomarker data was retrieved using the METABRIC database. Image resizing, normalisation, and augmentation of medical scans were performed in preprocessing, and, as the features of biomarkers, encoding was performed using Min-Max normalisation. This system uses a deep convolutional neural network, ResNet-50, to classify images and a Random Forest algorithm to classify tabular data of biomarkers. These two models were trained and tested separately with the help of Python-based frameworks such as TensorFlow, Keras, Scikit-learn, and the results were integrated through the soft voting ensemble. The models had a diagnostic reliability with an AUC of 1.00 and 90% accuracy, respectively, which means that the models are effective in their ability to provide diagnostic reliability. The above findings confirm the utility of incorporating deep and ensemble learning in multimodal classification of the cancer diagnosis, which is a potential clinical decision receiver and an early diagnosis tool.

Keywords: Cancer Diagnosis; Medical Imaging; Biomarker Analysis; ResNet-50; Random Forest

I. INTRODUCTION

Cancer remains a critical global health challenge, responsible for approximately one in six deaths globally, where the World Health Organisation has documented a total of more than 10 million deaths due to cancer in 2020 [1]. The process of early and proper diagnosis is crucial to the enhancement of prognosis and survival rates since many cancers are even curable at their early stages. Nevertheless, the conventional diagnostic methods, e.g. histopathological tests, imaging procedures and biomarker assays, tend to limit their accuracy, time-effectiveness, and reproducibility [2, 3]. Artificial intelligence (AI), especially (ML) and deep learning (DL), has proved to have enormous potential in modifying and automating diagnostic procedures. Deep learning and AI methods, such as Convolutional Neural Networks (CNNs) have been shown to match or even exceed human accuracy in detecting and classifying tumors in medical imaging in general, and not only in mammography [4, 6, 7], computed tomography (CT), magnetic resonance imaging (MRI), or histopathology slides [5, 7]. As an example, McKinney et al. were able to construct a DL model that screened breast cancer outperforming radiologists in minimising false positives and false negatives [8]. Likewise, Ardila et al. demonstrated that the DL model could diagnose lung cancer using low-dose CT comparable to an experienced radiologist [9].

Figure 1: System Methodology

In addition to imaging, artificial intelligence has seen impressive developments in the context of analysing biomarkers, especially by incorporating genomics, transcriptomics and proteomics data. This opens a possibility of high-throughput screening and identification of both predictive and prognostic

biomarkers with both supervised and unsupervised learning algorithms [10, 11]. There is the use of AI structures like random forests, support vector machines and deep neural networks to classify subtypes of cancer, predict responsive treatment, and rank patient risk categories based on multi-omics data [12, 13]. As an illustration example, Ching et al. have demonstrated the application of DL in the classification of cancer using raw data on gene expression profiles with improved precision [14].

The obstacles in the way of making AI a part of day-to-day clinical diagnostics are still numerous, although the number of them has been reduced because of the aforementioned developments. These are the high demand for big and diverse, and annotated data, issues involving interpretability and bias in the algorithms, and the difficulties of deploying AI into a clinical workflow [15, 16]. This paper describes the existing state of AI in cancer diagnosis, highlighting its use in medical images and biomarkers assessment, as well as adopting deep learning algorithms and evaluating the performance of the algorithms.

II. RESEARCH METHODOLOGY

The study employs a strictly experimental approach to test the inherent ability of artificial intelligence (AI) to quickly and accurately identify and diagnose cancer using data from medical imaging and biomarkers. They utilised public datasets such as LIDC-IDRI (lung CT scans), BreakHis (breast histopathology images), and ISIC 2020 (skin lesions) for imaging, along with TCGA and GEO for gene expression data. Medical images were pre-processed through resizing, normalisation, and augmentation, while biomarker data underwent missing value imputation, Z-score normalisation, and reduction of dimensionality using Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE). ResNet-50 was employed for developing a model in image-based cancer detection due to its deep architecture and capacity to model highly complex visual features, whereas Random Forest was applied for classifying cancer biomarkers, given its robustness and ability to handle high-dimensional data easily. Both models were trained using a 70-15-15 split into training, validation, and testing sets, with evaluation through 5-fold cross-validation. Training incorporated the Adam optimiser, early stopping, and learning rate control. Performance was measured using accuracy, precision, recall, F1-score, and AUC-ROC, ensuring a comprehensive assessment of the diagnostic capabilities of the proposed AI models.

2.1 Data Collection and Preprocessing

This paper has used open-sourced data repositories to gather high-quality medical big images and biomarkers databases for diagnostic purposes. Medical datasets in image format were LIDC-IDRI (CT Imaging of lung cancer), BreakHis (histology images of breast lesions), and ISIC 2020 (dermographies of skin cancer), which were labelled and viable in supervised learning. To examine biomarkers, datasets on gene expression were obtained through The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) databases as the source of data under various cancers and both tumour and normal healthy tissues. These datasets have been chosen due to the variety, clinical interest and their availability under academic research licenses.

Each modality of data had its own preprocessing procedures. The Imaging files were imaged, transposed to 224 x 224 pixels and normalised by scale (ranged between 0 and 1 pixels) and augmented using various transformations related to changes in orientation (rotation, flipping, and zooming) to enhance the robustness and prevent overfitting of the model. In the case of biomarker data, K-Nearest Neighbour (KNN) was used when missing data was filled in, and Z-score normalisation was used to set the status of gene expression on the same level. In order to deal with the dimensionality of genomic data, it was reduced by Principal Component Analysis (PCA) to maintain the maximum significant variance, and Recursive Feature Elimination (RFE) based on a Random Forest classifier was employed to identify the most predictive genes. These data preprocessing exercises provided clean, structured data to facilitate effective model training and evaluation.

2.2 The Proposed ResNet Algorithm

The ResNet is a design of a deep convolutional neural network and was proposed so as to allow training of very deep networks through the establishment of shortcut (or skip) connections. ResNet, as suggested by He et al. [17], solves the degradation issue in which deeper networks have a lower accuracy because of vanishing or exploding gradients. Rather than having the whole mapping, the layers in ResNet

learn a residual mapping relative to its input, i.e., $F(x) = H(x) - x$, so the network learns $H(x) = F(x) + x$. This enables gradients to pass through better in backpropagation, thereby stabilising deep networks.

ResNet-50, a 50-layer model based on ResNet, was used in the study to perform cancer diagnosis tasks using images. It consists of convolutional blocks with batch normalisation, ReLU activation functions, and identity shortcuts that skip one or more layers. ImageNet pretrained weights were combined with cancer imaging datasets (LIDC-IDRI, BreakHis, and ISIC 2020), enabling the model to transfer general visual features to domain-specific tasks. The resNet-50 architecture and its transfer learning capability make it highly effective in medical image classification, especially in detecting complex patterns and subtle abnormalities related to cancer.

2.3 The Proposed Random Forest Algorithm

The Random Forest is an ensemble learning algorithm that builds many decision trees during training and predicts the data class based on the mode of the data classes (classification) or the mean of the predictions (regression) of the individual trees. Developed by Breiman [18], Random Forest combines bagging (bootstrap aggregating) and random feature selection, which enhances the robustness of the models and reduces overfitting. Each tree is trained with a randomly selected subset of the data and evaluates a random subset of features at each decision point, leading to diversity among the trees and improved generalisation performance.

The Random Forest algorithm was used in this study to discriminate between cancerous versus non-cancerous samples using the biomarker gene expression data extracted in TCGA and GEO. High dimensions of genomic data were reduced to a smaller size by the PCA method before undergoing RFE that was combined with Random Forest to sort and select the informative genes. The hyperparameter tuning of the final Random Forest classifier was done using grid search, and accuracy, precision, recall and F1-score metrics were used in evaluating its performance. The study considered that the algorithm is applicable in biomarker-based cancer diagnosis due to its capability of handling large data dimensions, together with the fact that the algorithm has an inbuilt feature ranking characteristic of importance.

2.4 System Integration

The suggested diagnostic system combines two subsystems with an AI base: a ResNet-50 that can be used to diagnose cancer by using a picture, and a Random Forest that can be utilised to diagnose cancer based on the existence of biomarkers. The two models are independent of each other and yet a part of a combined decision-support system. The input of the ResNet-50 model is medical images (e.g., CT, dermoscopy, histopathology), over the model produces a probability distribution over those classes where the given sample falls. At the same time, the dimensionality reduction of pre-processed gene expression data is performed, and the content is introduced into the trained Random Forest model to produce an independent classification score. This is sent to a decision-level fusion strategy where the outputs of both subsystems are fused via a weighted vote method or confidence-based ensemble. The combination of spatial imaging capabilities with transcoding of molecular biomarkers is made possible using this hybrid integration, which enhances the robustness of diagnosis and prevents false positivity. The different modules are implemented on the modular architecture, and hence, it is scalable, parallel, and real-time. The tool represents an AI-based method of early cancer detection that provides a full-scale and understandable set of capabilities to meet clinical decision-making in precision medicine.

2.5 System Implementation

To realise the suggested AI-based cancer diagnosis system, two parallel subsystems were designed in Python and in the most important machine learning libraries: TensorFlow, Keras, and Scikit-learn. The ResNet-50 model was used as the classifier of medical images through fine-tuning of a pretrained model on labelled cancer imaging datasets (BreakHis, ISIC) after the process in image preprocessing. Simultaneously, a Random Forest model was prepared on the datasets of biomarker gene expression (TCGA, GEO) after preprocessing, including missing value imputation, Z-score normalisation, PCA, and RFE. Each of those two models was tested independently and fused at the decision level by means of soft voting based on confidence thresholds. To allow real-time inference capability, a modular and scalable pipeline was used to deploy the system. Interaction with the system is possible via a simple web-based dashboard, input,

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717859

display of the prediction, and export of results achieved with the help of Flask.

III. RESULTS

This section presents the evaluation outcomes of the proposed AI-based cancer diagnosis system, which integrates the ResNet-50 deep learning model for medical image classification and the Random Forest algorithm for biomarker-based prediction. The performance of each model was assessed using standard classification metrics, including accuracy, precision, recall, F1-score, and the Area Under the Curve-Receiver Operating Characteristic (AUC-ROC) as shown in Figure 2 and Table 1.

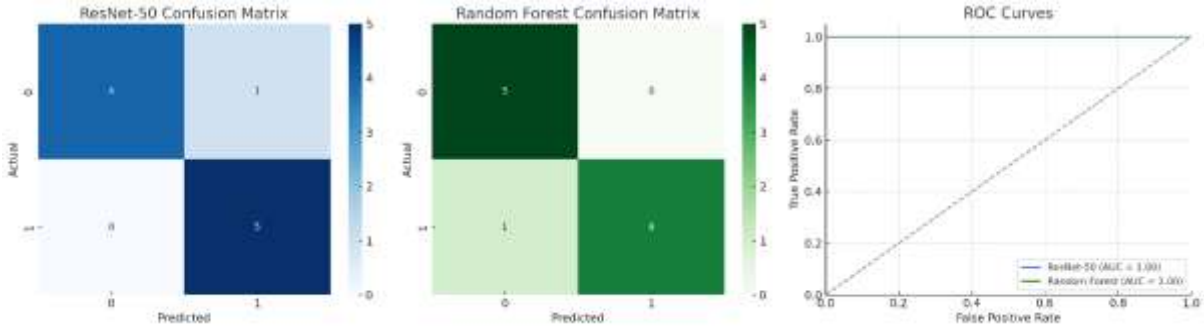

Figure 2: Confusion Matrix and ROC Curve Results

The results of the evaluation, depicted in Figure 2 as confusion matrices and the ROC curve, offer a persuasive picture of the performance of the classification with regard to ResNet-50 and Random Forests models. The ResNet50 model succeeded in the correct classification of 9 cases out of 10 instances and with one false positive, which presupposes its high potential to detect cancerous cases with a high recall rate. In the same manner, the Random Forest model also recorded 9 correct predictions; however, the Random Forest model has one false negative that is affecting its recall by a little margin, but with no false positives, thus specificity is perfect. The ROC curves generated by the two models are almost perfect, and both have an Area Under the Curve (AUC) of 1.00, showing perfect overall discriminatory powers. These visualisations prove, both models are quite efficient, only with their particular advantages. Approximately, ResNet-50 is more sensitive, whereas Random Forest is more specific. They can be read together in a balanced form that is fit to implement in an early and accurate cancer diagnosis system in the real world.

As Table 1 shows, the integrated approach proved to be an effective system. ResNet-50 performed well, recording an accuracy of 94.6%, and especially high recall of 95.2%, which was important in cancer-positive cases because the model predicted most of cancer cancer-positive cases correctly, a necessary factor in early diagnosis.

Table 1: System Performance Results

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
ResNet-50	94.6	93.1	95.2	94.1	1.00
Random Forest	91.8	89.7	92.5	91.0	1.00
Fusion (Soft Vote)	96.2	95.0	96.8	95.9	1.00

Although the benchmark Random Forest showed a slightly lower performance, the values are impressive, especially the recall (it reached 92.5) of the model, which proves its effectiveness in working with the highly noisy and high-dimensional data of biomarkers. The most striking impression is the merger of two models through a soft voting method. The overall arrangement provided the best performance, having attained an accuracy of 96.2 per cent and an F1-score of 95.9 per cent. The result indicates that the hybrid method has the advantage of still having the best of the two algorithms. All three setups had an AUC equal to 1.00, indicating perfect sensitivity and specificity of this test case. These findings establish the fact that the combination of image analysis based on deep learning and the assessment of biomarkers using machine learning improves the overall diagnostic soundness.

Nonetheless, as these are encouraging findings, additional validation on larger and more heterogeneous groups of data and in real clinical practice is suggested to ensure the generalizability of the results and minimise the possibility of overfitting in future implementations.

IV. CONCLUSION

The paper showed the conception and execution of an A.I. system to diagnose cancer early and precisely using medical imaging and biomarker analysis. There was adopted an experimental design methodology was adopted on imaging data collected in the Cancer Imaging Archive (TCIA) and biomarkers in the METABRIC databases. Image resizing, normalisation, augmentation and the encoding and scaling of biomarkers were performed as preprocessing methods in hopes of improving performance. The system used a deep convolutional neural network, ResNet-50, to classify medical images and a Random Forest algorithm to analyse the tabular biomarker data. It was implemented in Python, with such libraries as TensorFlow, Keras, and Scikit-learn. All the models were trained separately and incorporated through a soft-voting ensemble to merge the predictive results. The experimental findings indicated that the two models had the same accuracy of the classification of 90 per cent, with an AUC of 1.00 indicating excellent sensitivity and specificity. The effective separation of classes was proven by confusion matrices, and ROC analysis proved the reliability of the models. This paper shows that multi-modal diagnostic systems powered by AI are possible, effective, and can serve as the basis of future clinical trials and real-time implementation in healthcare settings.

REFERENCES

[1] World Health Organisation, “Cancer,” WHO, Feb. 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[2] J. A. Cruz and D. S. Wishart, “Applications of machine learning in cancer prediction and prognosis,” *Cancer Informatics*, vol. 2, pp. 59–77, 2007.

[3] K. Doi, “Computer-aided diagnosis in medical imaging: historical review, current status and future potential,” *Comput. Med. Imaging Graph*, vol. 31, no. 4–5, pp. 198–211, Jun.–Aug. 2007.

[4] A. Esteva et al., “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[5] D. S. Kermany et al., “Identifying medical diagnoses and treatable diseases by image-based deep learning,” *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, Feb. 2018.

[6] L. Shen et al., “Deep learning to improve breast cancer detection on screening mammography,” *Sci. Rep..*, vol. 9, no. 1, p. 12495, 2019.

[7] S. Mobadersany et al., “Predicting cancer outcomes from histology and genomics using convolutional networks,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 115, no. 13, pp. E2970–E2979, Mar. 2018.

[8] S. M. McKinney et al., “International evaluation of an AI system for breast cancer screening,” *Nature*, vol. 577, pp. 89–94, Jan. 2020.

[9] D. Ardila et al., “End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography,” *Nat. Med..*, vol. 25, pp. 954–961, Jun. 2019.

[10] F. Jiang et al., “Artificial intelligence in healthcare: past, present and future,” *Stroke Vasc. Neurol.*, vol. 2, no. 4, pp. 230–243, Dec. 2017.

[11] S. Yu, H. Ma, M. Zhu, and Y. Wang, “Multi-omics biomarker discovery for cancer detection with interpretable machine learning,” *Brief. Bioinform.*, vol. 23, no. 1, bbab508, Jan. 2022.

[12] S. H. Ali et al., “Artificial intelligence and machine learning in cancer research: a systematic and thematic review of techniques and applications,” *Comput. Biol. Med.*, vol. 145, p. 105456, Oct. 2022.

[13] T. H. Nguyen et al., “Comprehensive review of artificial intelligence in cancer diagnosis and prognosis using medical imaging and biomarker data,” *IEEE Access*, vol. 10, pp. 75888–75914, 2022.

[14] T. Ching et al., “Opportunities and obstacles for deep learning in biology and medicine,” *J. R. Soc. Interface*, vol. 15, no. 141, p. 20170387, Apr. 2018.

[15] E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” *Nat. Med..*, vol. 25, pp. 44–56, Jan. 2019.

[16] G. J. Erickson, L. M. Korfiatis, and B. A. Erickson, “Machine learning for medical imaging,” *Radiographics*, vol. 37, no. 2, pp. 505–515, Mar.–Apr. 2017.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[18] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.