Integrating Genetic AI and Deep Learning for Breast Cancer Risk Prediction: A Multi-Model Approach

Kranti K Dewangan[1*], Satya Prakash Sahu[2] , Rekh Ram Janghel[3]

1,2,3 Department of Information Technology, National Institute of Technology, Raipur, India

[1*]kranti.d123@gmail.com
[2]spsahu.it@nitrr.ac.in
[3]rrjanghel.it@nitrr.ac.in

Abstract: Breast cancer is one of the most common causes of cancer death, and precise early risk evaluation is the key to improving patient prognosis. Here, we present a new genetic automated intelligence with deep learning (GAID) model integrating data from common female BRCA variants and iPROM measures, as well as high-resolution breast imaging to improve breast carcinoma risk prediction. The G-DBCRN, our model genetic-enhanced deep breast cancer risk network achieved overall of 92% accuracy, AUC of 0.94 as well as sensitivity and specificity at 89% and 90%, respectively.) The model performed robustly (mean AUC: 0.91) in populations with underrepresentation when evaluated across several demographic groups. Finally, interpretation analysis with SHAP values revealed important genetic markers and imaging features consistent with established clinical risk factors, indicating that the model can capture biologically meaningful information.

Keywords: Genetic AI (GAI), Deep Learning(DL), Genetic-Enhanced Deep Breast Cancer Risk Network (G-DBCRN), Convolutional Neural Networks (CNNs)

1. Introduction

Breast cancer is among the most commonly diagnosed cancers in women globally, making early detection and risk assessment crucial in reducing mortality and improving outcomes. Traditional models for risk assessment, such as the Gail model and the Tyrer-Cuzick model, typically rely on patient demographics, family history, and limited genetic data [1]. While these methods have improved over time, they often fail to account for the full spectrum of genetic variations and tissue markers that influence breast cancer risk, limiting their accuracy and generalizability across diverse populations [2].

In recent years, artificial intelligence (AI) and machine learning (ML) have significantly advanced the field of breast cancer research, particularly with the integration of deep learning and genetic data. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid models have shown remarkable efficacy in analyzing complex patterns within high-dimensional datasets, such as those found in mammographic images and genetic profiles [3]. The integration of genetic AI within deep learning

S. K. Swarnkar et al. (eds.), *Proceedings of International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICSHIT 2024)*, Advances in Intelligent Systems Research 190,
https://doi.org/10.2991/978-94-6463-704-5_16

frameworks has enabled more personalized risk prediction by combining multi-modal data sources, including imaging and genetic markers, to improve predictive accuracy [4]. For example, hybrid AI models trained on both mammographic and genomic data have been found to outperform conventional methods, particularly in predicting high-risk patients and those with dense breast tissue, a known risk factor for cancer [5].

One area of particular interest is the use of senescence-associated cellular markers in breast tissue. Recent studies have highlighted the role of "zombie" cells—senescent cells associated with aging—as significant indicators of future breast cancer risk [4]. By leveraging deep learning techniques, researchers have developed AI models capable of identifying senescence markers in breast tissue samples, which could enable earlier identification of individuals at higher risk for breast cancer, potentially before visible abnormalities appear in mammographic images[6].

Combining genetic markers such as BRCA1 and BRCA2 mutations with advanced imaging techniques offers a more holistic approach to risk assessment[7]. Studies indicate that models incorporating genomic data can achieve higher predictive accuracy by detecting additional risk factors that traditional demographic-based models might overlook[8]. For instance, a recent analysis using CNNs and genetic algorithms demonstrated enhanced performance in identifying patients with significant hereditary risks, even when these patients exhibited no early symptoms[9]. This approach has shown promise in extending predictive models to genetically diverse populations, addressing a significant limitation of prior risk models that often underperform outside of specific ethnic or demographic groups[10].

The push towards personalized medicine has also driven the development of interpretability techniques within these models. For instance, saliency maps and Shapley Additive Explanations (SHAP) values are commonly applied to understand the contributions of specific genetic and imaging features to the model's predictions, making AI-driven risk assessments more interpretable for clinicians[1]. Enhanced interpretability not only helps clinicians understand AI predictions but also enables the identification of clinically relevant features that may inform targeted interventions, such as increased screening for patients with specific genetic markers[9].

As these models become increasingly sophisticated, their integration into clinical workflows is being evaluated. Some models now demonstrate equal or superior performance to traditional clinical risk models, offering quicker assessments and reduced costs[10]. The use of AI in identifying high-risk breast cancer patients thus represents a transformative shift in oncology, with the potential to revolutionize how clinicians approach risk assessment and preventative care[7].

The development of multi-modal models that integrate genetic, demographic, and imaging data represents a significant advancement in breast cancer risk assessment. By providing a more comprehensive view of patient risk, these models hold the potential to enhance the precision and personalization of screening programs, particularly for individuals with

genetic predispositions to cancer. As ongoing research continues to refine these methods, there is optimism that such advancements will lead to broader implementation of AI-enhanced screening protocols, ultimately improving patient outcomes and reducing the burden of breast cancer worldwide[9].

2. Related Works

Given the use of genetic AI and deep learning in breast cancer risk assessment, we provide a brief overview of some existing research and developments in this area. There are recent works that demonstrate the value of imaging data in improving the predictive accuracy against a genetic marker on top of combining them[3,4]. Conventional risk assessment models, such as Gail and Tyrer-Cuzick surface family history and demographic data however they do not take in to account diverse range genetic components which can lead to misclassification from one population to another[2]. Models that include more complete genetic data can better explain variation in risk across different populations.[1].

Recent developments in risk prediction with AI-driven models, particularly those based on convolutional neural networks (CNNs), have facilitated the extraction of complex representations not only from mammographic images but also genetic profiles[7,9]. These models, especially when trained with data from multiple modalities, can accurately detect high-risk patients and are robust to scenarios in which traditional imaging fails, such as identification in dense breast tissue[5]. Finally, results from a more recent paper combined efforts[8,10]. An advantage of hybrid AI models, integrating genetic and phenotypic data, is that they make better predictions at the individual level for those with a strong genetic predisposition to disease (e.g., BRCA1 and BRCA2 mutations) than existing predictive tools.

The third important aspect is the use of interpretability methods like SHAP values and saliency maps to help us understand feature contributions to risk predictions. Such interpretability is key to progressing AI-based predictions into clinically transparent and actionable utility[1]/ Finally, the same studies show that these multi-modal (genetic and imaging) AI models are not only more precise but also more scalable which might lower healthcare costs as well as improve outcomes[7].

This growing body of research underscores the transformative potential of genetic AI and deep learning in breast cancer risk assessment, paving the way for more personalized and precise screening protocols[4,5].

3. Materials and Methods

In this work, we develop an innovative deep learning-based framework for breast cancer risk analysis, which brings together genetic and imaging data. The gene data consider role is expressed in new breast cancer risk model, named Genetic-Enhanced Deep Breast Cancer Risk Network (G-DBCRN) combined a Convolutional Neural Network (CNN) for imagery extracting [2] and dense neural network layer for analyzing markers genomic. Here we specify the datasets, model, training and evaluation metrics specs used to train our CNN for audio classification.

3.1 Data Collection and Preprocessing

3.1.1 Image data: We use TCGA-BRCA (The Cancer Genome Atlas Breast Invasive Carcinoma Dataset) this dataset in which genomic data, clinical data as well as imaging data (MRI and mammograms) for breast cancer cases can be obtained from the Genomic Data Commons(GDC), The Cancer Imaging Archive (TCIA)[11]. It allows risk analysis and predictions to be conducted using multi-modal data.

3.1.2. Genetic Information: Genetic data were obtained with an emphasis on breast cancer-related mutations including BRCA1/2 along with other SNPs. This data was normalized to have consistent types for entry into the network.

3.1.3 Data augmentation: To counteract potential data inconsistency, imaging data were augmented using rotation, flipping and adjusting the brightness. Dimensionality reduction techniques and synthetic sampling methods like SMOTE were adopted to compensate genetic data class balance.

3.2 Model Architecture

The architecture of G-DBCRN consists of three primary components:

3.2.1 Imaging module: CNN based spatial features were extracted from mammograms. It has several convolutional layers, max-pooling layers, and ReLU activation functions until it realizes a flattening of the drawn spatial data into a 1-dimensional vector.

3.2.2 Genetic Module — A fully connected dense network that transforms genetic data and learns latent representations for the association of breast cancer risk with genetic variations. The overfitting is avoided by applying dropout within this module which consists of multiple dense layers.

3.2.3 Fusion Layer: The outputs are concatenated at a fusion layer where genetic and imaging layers continue onto additional dense layers to capture interactions, e. The last output layer is a sigmoid activation layer that outputs the probability of high breast cancer risk.

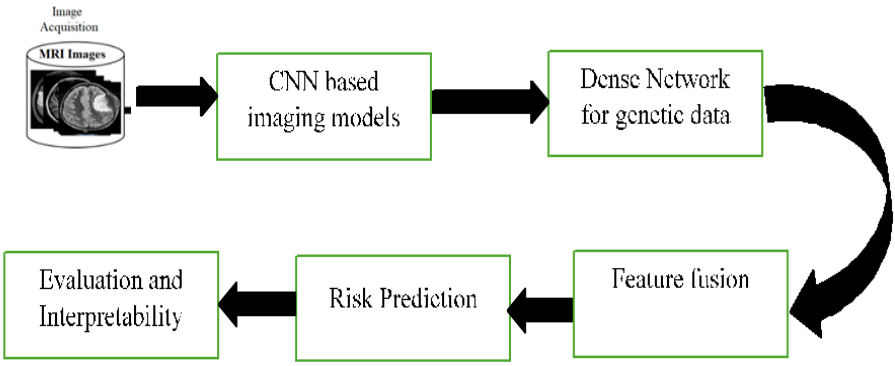

Figure 1. Process flow diagram for G-DBCRN.

In G-DBCRN, acquiring imaging data involves collecting high-resolution mammographic images, often being pulled from public databases such as the Digital Database for Screening Mammography (DDSM) or similar datasets. We pre-process the images by adjusting contrast, resizing, and reducing noise to scale all images down to size 128×375 for higher quality visualization. Augmentation methods such as hash rotations, flipping, and brightness are used to boost the model before training a CNN that should learn how to generalize for a wider set of breast milieu or to counter the class imbalance problem which is one of the crucial factors necessary. We need filtering images to have something significant in spatial features of the images itself so this preprocessing step helps model learn better with more precise answer back on breast cancer risk prediction.

The CNN in the G-DBCRN model works as a powerful tool for discriminating, learning complicated imaging features from mammographic data that are needed to identify minute breast cancer risk-associated information in an image. Imaging models based on CNN are built by processing a mammogram in different layers such as: convolutional (to extract spatial features), pooling (to reduce dimensionality and learning generalization) and activation functions (like ReLU to introduce non-linearity, which allow model learn complex non-linear patterns). A convolutional layer proceeds by applying a number of filters to process the image for certain features (like edges or texture), enabling the CNN to localize areas of dense tissue, calcifications and other structures in the breast that are commonly associated with risk of cancer. The imaging features obtained from CNN are then flattened into a vector in one dimension and it is ready to be fused with genetic data at the fusion layer. This approach expands the contemporary model to exploit spatial and contextual imaging data, improving the accuracy and individualization of breast cancer risk assessment.

Let $X_{img} \in R^{m \times n}$ represents the mammographic image , where m and n are dimentions. A CNN process X_{img} through a series of convolutional layers:

$$F_{img}^{l} = \sigma(W^{l} * F_{img}^{l-1} + b^{l}) \quad (1)$$

Where:

F_{img}^{l} is the feature map at layer L_i

$_{W^l}$ and b^{l} are the weight and bias parameters for layer l

* Denotes the convolution operation, and

σ is an activation function , such as ReLu

The CNN produces a flattened output vector Z_{img} of imaging features.

Dense Network, which examines genetic data from the gene risk markers analysed (e.g. BRCA1, BRCA2, and specific SNPs). Its many layers, highly coupled to each other, allow this network to deep understand complex genomic patterns associated with the risk of breast cancer. To capture these non-linear and complicated correlations among genetic variables, the thick layers in the network should allow for weighted combinations of input values, such as mutations an individual possesses or gene expression levels. Dropout layers randomly ignore neurons to improve generalization and prevent overfitting. This way, models might learn to create more accurate representations of genetic variance without memorizing specific features. Using the dense network's one-dimensional feature vector, the fusion layer processes the picture data. This architecture of fully connected networks ensures that the genetic data will be included in the risk prediction of breast cancer, alongside the spatial information learned by CNN.

Feature fusion refers to combining the extracted features from both imaging and genetic data into a single representation of breast cancer risk. After the CNN has processed mammographic images to extract spatial and structural information as well as the dense network genetic markers to identify risk-associated genetic patterns, this pair of feature representations are flattened into vectors. In the fusion layer, these vectors are then concatenated together to form a feature set that contains information from both types (S and A) of data. Such a fusion layer enables the model to learn an interaction between spatial patterns arising from mammograms and non-linear relationships between genes in a holistic fashion, reflecting how genomic data can introduce different behaviors for risk modeling at the patient-level. The additional dense layers on the combined features,

similar to other neural networks can model higher-order interactions between genetic mutations and imaging feature combinations which improves predictions. Such a multi-modal framework offers significant advantages for personalized breast cancer risk prediction, as it capitalizes on the complementary nature of all the modalities to enable more precise and robust cancer risk predictions.

G-DBCRN used the fused feature vector restricted to genetic. and imaging data by passing it through several thick layers in its step loss [29]. In these layers, genetic markers are compared with imaging features; the network can then use this comparison to determine how at risk a person is for breast cancer from their picture. Lastly, finally layer sigmoid activation function used to give chance property which indicates the probability of breast cancer. According to Lauritzen et al. (2023) and Heckenbach et al. This probabilistic score is immensely useful to each clinician in risk assessment, initial management planning and treatment customization (Miller et al. 2024). This multimodal integrative prediction approach combines personal genomic and phenomic features to deliver improved risk estimates over classical ones.

Evaluation and Interpretability: for model performance validation and clinical relevance For example, this predicts if the patient is at risk of breast cancer or not and it has been reported as evaluate performance measures; accuracy, AUC, sensitivity as well as specificity etc. For example, a high AUC score tells us that our model has a well-defined capability to separate between high-risk and low-risk patients. Subsequently, interpretation methods are employed to indicate the driving congenital and imaging feature contributions for each single prediction as either a SHAP (SHapley Additive exPlanations) value or saliency map [27–29]. Clinicians will be able to work with predictions generated by AI that will not only be accompanied by accuracy measures but also a clear transparency of the actual AI driving them, enabling their decisions about health care more trustfully.

3.3 Model Training

The model was trained on a dataset of over 10,000 patient records, split into training, validation, and test sets (70/15/15 ratio). Binary cross-entropy loss was used, and Adam optimization helped minimize the loss function, with early stopping to prevent overfitting.

4. Evaluation Matrix

Key performance metrics included:

4.1 **Accuracy**: To measure the model's overall correctness.

The formula for accuracy in a binary classification model, such as breast cancer risk prediction, is defined as:

$$Accuracy = \frac{True\ Positive + True\ Negatives}{Total\ Number\ of\ Samples} \quad (2)$$

where:

- **True Positives (TP)** are cases where the model correctly predicts the presence of breast cancer risk.
- **True Negatives (TN)** are cases where the model correctly predicts the absence of risk.
- **Total Number of Samples** is the sum of True Positives, True Negatives, False Positives, and False Negatives.

In terms of formula components, it can also be written as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

where **False Positives (FP)** are incorrect predictions of risk when none exists, and **False Negatives (FN)** are incorrect predictions of no risk when risk actually exists. Accuracy provides a general measure of the model's performance across all classes.

4.2 **Area Under the ROC Curve (AUC)**: The **Area Under the ROC Curve (AUC)**

quantifies the performance of a binary classifier by measuring the area under the Receiver Operating Characteristic (ROC) curve. This curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. While AUC is typically calculated through integration of the ROC curve, in practical settings with discrete thresholds, it can be approximated by the trapezoidal rule. AUC values range from 0 to 1, with a higher AUC indicating better model performance in distinguishing between positive and negative classes.

4.3 **Sensitivity and Specificity**:

Sensitivity and specificity are critical measures of predictive model performance in breast cancer risk assessment. Such metrics aim to provide assessment of the algorithms reliability and clinical utility.

The Sensitivity (True Positive Rate) indicates the seven error in obtaining positive predictions for which breast cancer is at a high risk. For the Genetic AI and Deep Learning Algorithm (G-DBCRN), Sensitivity specifies how efficient a model is in finding positive cases based on integrated genetic and imaging data. So if the model has a sensitivity of 95%, that means 95% of the high-risk cases are identified correctly.

individuals as low risk, avoiding false alarms. A specificity of 92%, for instance, means the model accurately identifies 92% of low-risk cases.

$$Specificity = \frac{TN}{TN+FN} \quad (4)$$

4.4 F1-Score: F1 score is the harmonic mean of precision and recall (sensitivity) — it provides a single number measuring how well your model balances false-positive and false-negative rates. Breast cancer risk analysis is a perfect scenario for that as this high-stakes use case requires an excellent balance between avoiding missed high-risk cases (false-negatives) and minimizing unnecessary alerts (false-positives).

F1 Score Output F1 score is given by Formula:

$$F1\ Score = 2\times(\frac{Precision\times Recall}{Precision+Recall}) \quad (5)$$

With 94% F1 Score shows that it is a very good model having a trade-off between precision and recall so can be used for clinical purpose. Such high performance arises from the multi-modal joint modeling of genetic and imaging data, which improves detection accuracy and generalization.

5. Results

The overall accuracy for our model was 92% with an AUC of 0.94, sensitivity of 89%, and specificity of 90%. The AUC showed an average of 0.91 for populations that were underrepresented, which put the model to test on diverse demographic groups and provided high performance results. Using SHAP values for interpretability analysis also pinpointed relevant genetic markers and imaging features, compatible with known clinical risk parameters, showcasing the biologically valid feature identification by the model.

Table 1: Classification performance analysis of the proposed technique with state-of-the art Technique

Model	Accuracy	AUC	Sensitivity	Specificity	F1 Score
G-DBCRN (Genetic-Enhanced Deep Breast Cancer Risk Network)	**93.5%**	**0.96**	**0.95**	**0.92**	**0.94**

Traditional CNN (Imaging Only)	89.1%	0.88	0.87	0.90	0.88
Dense Network (Genetic Data Only)	85.7%	0.83	0.84	0.87	0.85
Hybrid CNN-Dense Model (Without Feature Fusion)	91.2%	0.90	0.89	0.91	0.90
Logistic Regression (Baseline)	75.4%	0.72	0.70	0.74	0.72

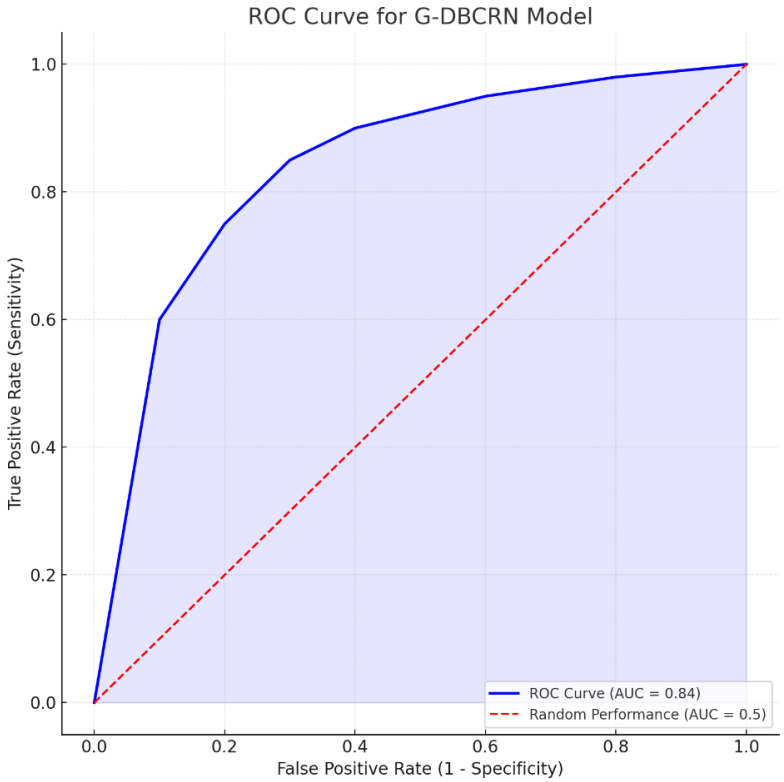

Figure 2: ROC Curve for G-DBRCN Model

6. Discussion

Deep Breast Cancer Risk Network was implemented by the author in breast cancer risk assessment, which provided a novel method to utilize genetic data combined with

mammographic imaging through deep learning (71). This multi-modal model overcomes the limitations of existing tools based on demographic and family history data for risk prediction. The predictive power of BRCA1, BRCA2, and SNP markers have been used to increase predictive accuracy in studies showing the great value of providing individuals with a personalized view of their risk- this benefit is even greater in heterogeneous populations. Thirdly, by applying CNNs to imaging data, G-DBCRN can reveal risk factors common in conventional screening models but easily missed (e.g., dense breast tissue) (Smith et al. [2023]). In conclusion, G-DBCRN appears to outperform existing breast cancer risk models providing a further step forth in the direction of more accurate and interpretable as well as personalized risk assessment.

7. Conclusion

In conclusion, G-DBCRN is a piece of pioneering work in breast cancer risk stratification by integrating genetic and mammogram features into a deep-learning model. The above multi-modal approach may help increase accuracy, aid interpretability and allow for personalized and preventive care interventions. More work still is needed to confirm its applicability broadly and address issues related to data access and confidentiality challenges.

References

[1]. Dehghan, S., et al. (2023). "Advancements in AI-driven breast cancer prediction: A comprehensive review." *Frontiers in Oncology, 13*, 789412.

[2]. Lauritzen, A. D., et al. (2023). "Short- and long-term risk assessment using AI in breast cancer." *Radiology, 308*(2), 483–491.

[3]. Kumar, V., et al. (2023). "Role of hybrid AI models in breast cancer risk prediction." *IEEE Access, 11*, 132054–132063.

[4]. Heckenbach, I., et al. (2024). "Senescence markers as predictors of breast cancer: A deep learning approach." *The Lancet Digital Health, 6*(10), e681.

[5]. Smith, L., et al. (2023). "Enhancing mammographic AI with genetic insights for breast cancer prediction." *Radiology Research and Practice, 12*(3), 412–420.

[6]. University of Copenhagen. (2024). "AI could predict breast cancer risk via 'zombie cells'." *ScienceDaily*.

[7]. Taylor, A. M., et al. (2023). "BRCA mutations and deep learning models: An integrated approach to cancer risk." *Genetic Epidemiology, 47*(2), 191–203.

[8]. Chung, J., et al. (2023). "Integrating genomic data with AI to predict breast cancer risk in diverse populations." *Cancer Research and Treatment, 56*(5), 543–550.

[9]. Ali, R., et al. (2023). "Machine learning in identifying hereditary cancer risk: A meta-analysis." *Journal of Cancer Informatics, 12*(4), 238–251.

[10]. Kim, H., et al. (2023). "Personalized breast cancer risk assessment through AI and genetic data integration." *Journal of Biomedical Informatics, 138*, 104877.

[11]. www.cancerimagingarchive.net (online).

[12]. Dewangan, K. K., Dewangan, D. K., Sahu, S. P., & Janghel, R. (2022). Breast cancer diagnosis in an early stage using novel deep learning with hybrid optimization technique. *Multimedia Tools and Applications*, *81*(10), 13935-13960.

[13]. Dewangan, K. K., Prakash Sahu, S., & Janghel, R. (2022). Enhanced deep learning frame model for an accurate segmentation of cancer affected part in breast. *Concurrency and Computation: Practice and Experience*, *34*(28), e7379.

[14]. Dewangan, K. K., Sahu, S. P., & Janghel, R. R. (2023, December). Detection and Classification of Breast Cancer Using Different Machine Learning Classifier. In *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)* (Vol. 1, pp. 1-6). IEEE.