A Self-Supervised Learning Approach for Detecting BRCA Mutations in Breast Cancer Histopathological Images

Faycal Touazi[1,*,†], Djamel Gaceb[1,†], Chaima Belkadi[1,†] and Besma Loubar[1,†]

[1] *LIMOSE Laboratory, Computer Science Department, University M'hamed Bougara, Independence Avenue, 35000 Boumerdes, Algeria*

Abstract

Breast and ovarian cancers are among the most pressing health issues affecting women globally, with genetic mutations, particularly in the BRCA1 and BRCA2 genes, significantly influencing their development. This thesis offers a comprehensive overview of these cancers, emphasizing the genetic, anatomical, and histopathological factors that contribute to their onset and progression. A detailed examination of the anatomy of the female breast and ovaries provides insight into the origins of these malignancies. The critical role of histopathology in identifying specific cancer subtypes and gene mutations is explored, underscoring its vital importance in diagnosis and treatment. Our results demonstrate that the developed deep learning framework, integrating Vector Quantized-Variational Autoencoders (VQ-VAE) and DBSCAN for clustering, achieved an accuracy of 95% in classifying BRCA mutation-positive and negative cases, outperforming traditional diagnostic methods. By investigating the interplay between genetic predisposition and histopathological analysis, this thesis aims to enhance the understanding of breast and ovarian cancers and their implications for public health.

Keywords

Breast Cancer, BRCA mutation, deep learning, Self supervised learning, BRCA 1, BRCA 2

1. Introduction

Breast cancer remains one of the most prevalent cancers globally, affecting millions of women each year. Early detection is a critical factor in improving survival rates, as it allows for timely intervention and treatment. Traditional methods for breast cancer detection, such as mammography, have long been the gold standard in screening programs. In recent years, deep learning has emerged as a powerful tool for enhancing breast cancer detection, particularly in medical imaging tasks such as mammography interpretation.

Deep learning algorithms, particularly Convolutional Neural Networks (CNNs), have demonstrated remarkable success in breast cancer detection from mammography images. Studies have shown that deep learning models can achieve accuracy levels comparable to radiologists in identifying tumors [1, 2, 3, 4].

Mutations in the BRCA1 and BRCA2 genes are among the most well-known genetic risk factors for breast cancer. These mutations, which can be inherited, significantly increase a woman's lifetime risk of developing breast cancer. Women who carry a BRCA1 or BRCA2 mutation have an elevated risk of 50 to 85% of developing breast cancer by the age of 70, compared to a 12% risk in the general population [5].

The discovery of a BRCA mutation in a patient is of crucial importance, not only for early diagnosis and cancer management, but also for informed decision making regarding preventive measures and treatment options. Identifying such mutations can lead to personalized surveillance strategies, risk-reducing surgeries, and targeted therapies, thus improving the overall prognosis and quality of life of high-risk patients [6] [7] [8].

IDDM'2024: 7th International Conference on Informatics & Data-Driven Medicine

[*] Corresponding author.

[†] These authors contributed equally.

f.touazi@univ-boumerdes.dz (F. Touazi); d.gaceb@univ-boumerdes.dz (D. Gaceb); c.belkadi@univ-boumerdes.dz (C. Belkadi); b.loubar@univ-boumerdes.dz (B. Loubar)

0000-0001-5949-5421 (F. Touazi)

Deep learning has revolutionized various domains and its impact on the medical field is particularly profound. The ability of deep learning algorithms to analyze complex patterns in large datasets has led to significant advances in medical diagnostics, treatment planning, and personalized medicine. In the context of medical imaging, deep learning models have demonstrated exceptional accuracy in tasks such as detecting abnormalities, classifying diseases, and predicting patient outcomes. These models, which often outperform traditional methods, have the potential to assist clinicians in making more informed decisions and improving patient care.

Our study leverages deep learning techniques to address the challenges associated with detecting BRCA1 and BRCA2 mutations in histopathological images of breast and ovarian cancer. By training a robust model on a curated dataset, we aim to provide a reliable tool for identifying these genetic mutations. The results presented in this work highlight the effectiveness of our approach and demonstrate the potential of deep learning in enhancing the accuracy of cancer detection and prognosis. Through this research, we contribute to the growing body of evidence supporting the integration of deep learning into clinical practice, ultimately aiming to improve outcomes for patients with hereditary cancer risks.

This paper is organized as follows: Section 2 reviews related works. Section 3 outlines our proposed approach, focusing on the Vector Quantized Variational AutoEncoder (VQ-VAE). Section 4 describes the experimental setup, including the TCGA-BRCA dataset, preprocessing, and evaluation metrics. Section 5 presents the results, covering clustering, BRCA patch classification, and SVS image classification, along with comparisons to related work. Finally, Section 6 concludes with a summary of findings and future research directions.

2. Related Works

In this section, we offer a comprehensive review of recent studies that focus on detecting BRCA mutations in breast cancer using deep learning methods.

Shen Zhao et al. [9] developed a deep learning framework for comprehensive molecular and prognostic stratifications of triple-negative breast cancer (TNBC). The framework features two CNNs in series: the first, a tissue type classifier based on ResNet-18, achieved a weighted F1 score of 0.96, classifying tissue types with near 90% accuracy. The second CNN predicted molecular features and relapse risks with AUCs ranging from 0.71 to 0.76.

Xiaoxiao Wang et al. [10] proposed a deep learning model based on CNNs to predict BRCA gene mutations from histopathological images. Trained on the JSPHCM and JSCH datasets, their model demonstrated high performance with AUC values ranging from 79%.

Tristan Lazard et al. [11] employed multiple instance learning (MIL) techniques to identify morphological patterns indicative of homologous recombination deficiency in luminal breast cancers. Their model, tested on a dataset of 673 WSIs from TCGA and an in-house dataset, achieved an AUC of 71%.

Nam Nhut Phan et al. [12] developed a deep learning pipeline for classifying breast cancer molecular subtypes from unannotated pathological images. Their approach utilized a two-step transfer learning process with CNNs such as ResNet50, ResNet101, VGG16, and Xception. Initially, the models were pre-trained on ImageNet and then fine-tuned on an internal dataset. They were subsequently trained on the TCGA-BRCA dataset to classify breast cancer into basal, HER2, luminal A, and luminal B subtypes. The images were normalized to 512x512 pixels, and patches were extracted from WSIs. The models achieved average AUCs ranging from 88 to 92%.

Kurian et al. [13] proposed a semi-supervised learning approach to classify breast cancer subtypes using histopathological images from the TCGA-BRCA dataset. They focused on differentiating between Basal and Luminal A PAM50 subtypes by analyzing a curated subset of 180 whole slide images (WSIs) selected to minimize heterogeneity. Their model leveraged a Deep Neural Network (DNN) architecture based on SimCLR with a ResNet18 backbone for out-of-distribution (OOD) detection, pre-trained on a large histology image dataset. Patch extraction from annotated tumor regions enabled the model to focus on relevant regions, although it introduced potential label noise. They achieved a patient-level accuracy of 81.43%.

The methodology employed by Valieris et al. [14] involved developing a deep learning framework to detect homologous recombination (HR) deficiency in breast tumors using the TCGA-BRCA dataset. The model leveraged whole slide images (WSI), utilizing advanced image processing techniques to extract histopathological features indicative of HR deficiency. To address the complexity and variability in these images, the authors implemented a multiple instance learning (MIL) approach, allowing the model to learn from entire tumor samples without the need for manual segmentation. Their model achieved an area under the curve (AUC) of 80%.

Table 1 provides a comparative summary of the performance achieved in state-of-the-art studies for breast cancer classification, highlighting different datasets, methods, and evaluation metrics used.

Table 1
Performance Achieved in State-of-the-Art Breast Cancer Studies

Reference	Dataset	Year	Methods	Metrics
Tristan Lazard et al. [11]	TCGA	2022	ResNet-18	AUC 71%
Xiaoxiao Wang et al. [10]	JSPHCM, JSCH	2021n	ResNet-18	AUC 79%
Kurian et al. [13]	TCGA-BRCA	2023	SimCLR	81.34% accuracy
Valieris et al. [14]	TCGA	2020	Resnet34	AUC 80%
Nam Nhut Phan et al. [12]	TCGA-BRCA	2021	2-Step ResNet50,101, VGG16, Xception	AUC 92%

3. Proposed approach

In this section, we describe our proposed approach for the detection and diagnosis of breast cancer using advanced deep learning techniques. Our approach is designed to address the challenges of analyzing histopathological images and aims to provide a comprehensive solution to detect and classify breast masses. But first we will introduce a key architecture in our proposal.

3.1. Vector Quantized Variational AutoEncoder

The Vector Quantized Variational AutoEncoder (VQ-VAE)[15] is a type of variational autoencoder that introduces vector quantization to obtain a discrete latent representation, distinguishing itself from traditional VAEs, which produce continuous latent codes. The VQ-VAE uses a codebook, a matrix $\mathbf{e}$ of dimensions $K \times D$, where K represents the number of embeddings and D is the dimensionality of each embedding (see Figure 1). This architecture enables the encoding of data into discrete codes, which helps in learning compact and structured representations. The model consists of three main components:

- **An encoder:** that maps input data x (such as images) into continuous latent representations z.
- **A vector quantizer:** that transforms these continuous representations into discrete vectors e_k using the codebook, by selecting the closest embedding through minimizing the Euclidean distance:

$$\text{quantization}(z) = \underset{e_k \in E}{\text{argmin}} ||z - e_k||^2 \tag{1}$$

- **A decoder:** that reconstructs the original data from the discrete latent codes.

The codebook, which contains the learned embeddings, plays a critical role in the quantization process. It allows the continuous output of the encoder to be mapped to discrete codes, facilitating the generation of data from these discrete codes. By using a discrete latent space, VQ-VAE simplifies model optimization and enables the use of generative models based on discrete distributions, such as PixelCNN or other autoregressive models, to model the latent codes.

One of the key advantages of VQ-VAE is its ability to capture meaningful structural representations in the data, making it particularly useful for high-quality generation tasks, especially in areas like medical imaging and discrete signal modeling.

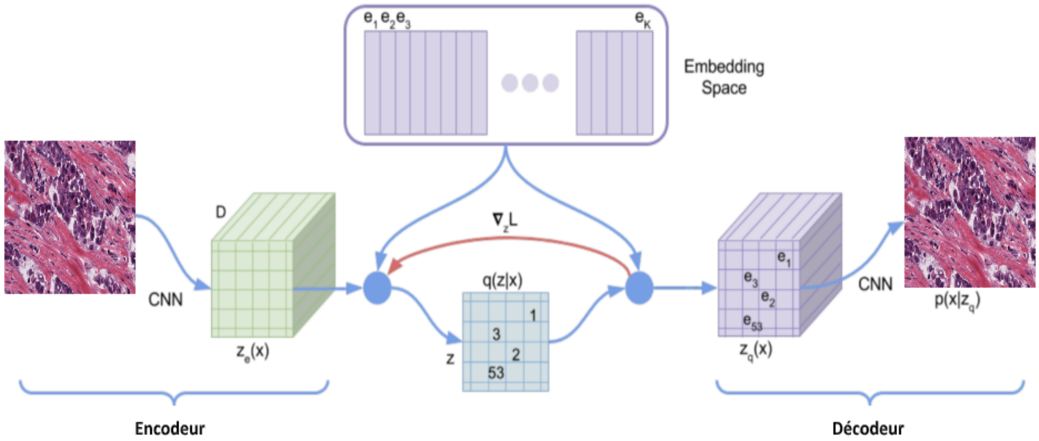


Figure 1: VQ-VAE Architecture: The image is encoded into a grid of latent vectors. These vectors are replaced by the nearest codebook vector at the bottleneck. Finally, the quantized vectors pass through the decoder to reconstruct the image [15].

3.2. Methodology

In this section, we detail the steps of our methodology for analyzing histopathological images to identify BRCA mutations and classify cancerous versus normal cells. Our approach employs image preprocessing, feature extraction using Vector Quantized-Variational Autoencoder (VQ-VAE), and clustering techniques to organize images based on mutation status, enabling precise classification and improving diagnostic accuracy (see Figure 2).

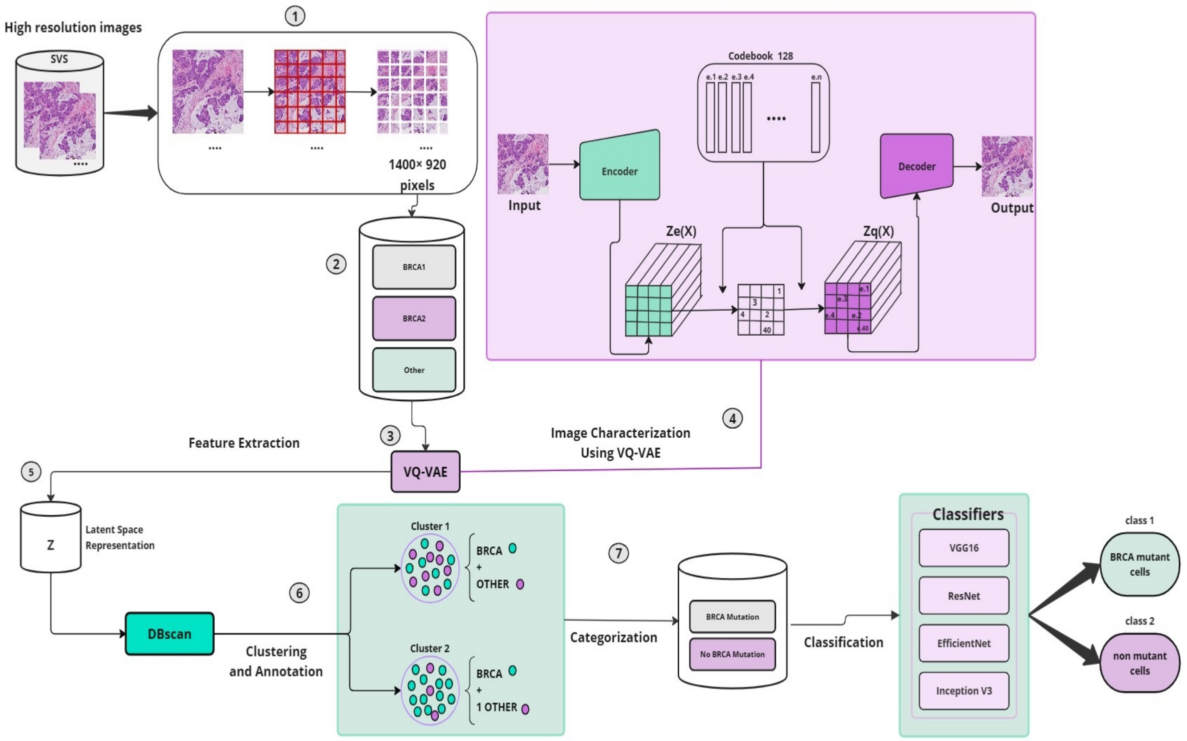


Figure 2: Overview of the proposed architecture for histopathological image analysis.

- **Step 1: Input Images** The process begins with the acquisition of large histopathology images in

SVS (Scalable Vector Graphics) format. These images are particularly challenging to handle due to their high resolution and substantial size, which necessitates advanced processing techniques. To manage this, the SVS images are divided into smaller patches of size 1400 × 920 pixels. This approach simplifies the analysis and processing of the images while retaining important details. The dataset comprises a diverse set of patches,

- **Step 2: Dataset Categorization:** After patchifing the dataset into small images were CNN models can process theme, further refinement involves categorizing patches based on BRCA mutation status. The patches are divided into three distinct categories: those related to SVS images where BRCA1 is identified, those where BRCA2 is identified, and those with no identified BRCA mutations. This detailed categorization enables a more focused analysis of IDC patches in relation to specific BRCA mutations, enhancing the understanding of their histopathological features.
- **Step 3: Image Characterization Using Vector Quantized-Variational Autoencoder (VQ-VAE)**
  In our approach, we utilize a Vector Quantized-Variational Autoencoder (VQ-VAE) [16] with a codebook of 128 discrete vectors to handle and analyze high-resolution histopathological images. This model is crucial for effectively managing the complexity of these images through its encoder-decoder architecture.

  - **Encoder Network:** The encoder network transforms high-resolution input images into a continuous latent representation. It consists of multiple neural network layers that extract significant features and reduce the dimensionality of the images while retaining important details.
  - **Vector Quantization:** Following the generation of the continuous latent representation, VQ-VAE applies vector quantization. This process maps the continuous latent vectors to the nearest discrete vectors in a predefined codebook of 1024 entries. This quantization step converts the latent space into a more manageable and structured form, which simplifies further analysis.
  - **Codebook:** The codebook, comprising 1024 discrete vectors, is updated during training to minimize reconstruction loss. This ensures that the codebook effectively captures the essential characteristics of the input images.
  - **Decoder Network:** The decoder network reconstructs the high-resolution images from the quantized latent representation. Using the discrete codes produced by the encoder, the decoder aims to accurately recreate the original images, preserving critical features and details.
  - **Dimensionality Reduction and Efficient Analysis:** The combination of the encoder, vector quantization, and decoder facilitates dimensionality reduction of high-resolution images. This reduction compresses the data into a latent space that retains essential information, making the data more suitable for efficient analysis and processing.

- **Step 4: Feature Extraction with VQ-VAE:** To extract meaningful features from the images, we utilize the VQ-VAE model. The VQ-VAE's encoder network processes the histopathological images to generate continuous latent representations, which are then quantized into discrete vectors using the codebook. This approach captures intricate patterns and features within the images. The aim is to characterize the images with a reduced dimensionality representation, which simplifies and enhances the clustering operation. This method provides a comprehensive and structured feature representation by reducing the dimensionality of the high-resolution images, making it easier to perform effective clustering.
- **Step 5: Latent Space Representation:** After feature extraction, each image is encoded into a latent vector. These latent vectors collectively form a dataset that is used for subsequent analysis. This latent space representation simplifies the data and prepares it for clustering and other operations.

- **Step 6: Clustering and Annotation:** After encoding the images into latent vectors, we apply the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm to group similar vectors into clusters. This clustering approach organizes the images into meaningful groups based on their feature vectors, distinguishing between BRCA mutation-positive and BRCA mutation-negative cases.
- **Step 7: Classification:** The final stage involves classifying the images into two main categories: normal cells and cancerous cells. This classification is based on the previously obtained clusters and latent space representations. The model aims to enhance the accuracy of differentiating between various types of cancerous and non-cancerous tissues, thereby improving diagnostic capabilities.

By integrating VQ-VAE with advanced CNNs and clustering techniques, our approach provides a robust framework for analyzing histopathological images. This methodology is designed to improve the performance of breast cancer detection and diagnosis, offering a more accurate and comprehensive analysis of histological samples.

4. Experimentations and results

4.1. TCGA-BRCA Dataset

The TCGA-BRCA dataset, referenced in [17], is part of the Cancer Genome Atlas (TCGA) project, which aims to enhance the understanding of cancer through comprehensive genomic studies. This dataset includes RNA sequencing, somatic mutation profiles, and gene-level copy number variation data from 1,098 breast invasive carcinoma cases. It contains 1978 images from these 1,098 patients, with 763 tumor samples that include single nucleotide polymorphism (SNP) and copy number variation (CNV) data generated using the Affymetrix 6.0 SNP array, alongside somatic mutation information obtained from the Illumina sequencing platform. Data sources include the Genomic Data Commons (GDC) Data Portal, Pan-Cancer Atlas, and The Broad Institute's TCGA GDAC Firehose. The dataset is publicly available through both the GDC Data Portal and the Cancer Imaging Archive (TCIA).

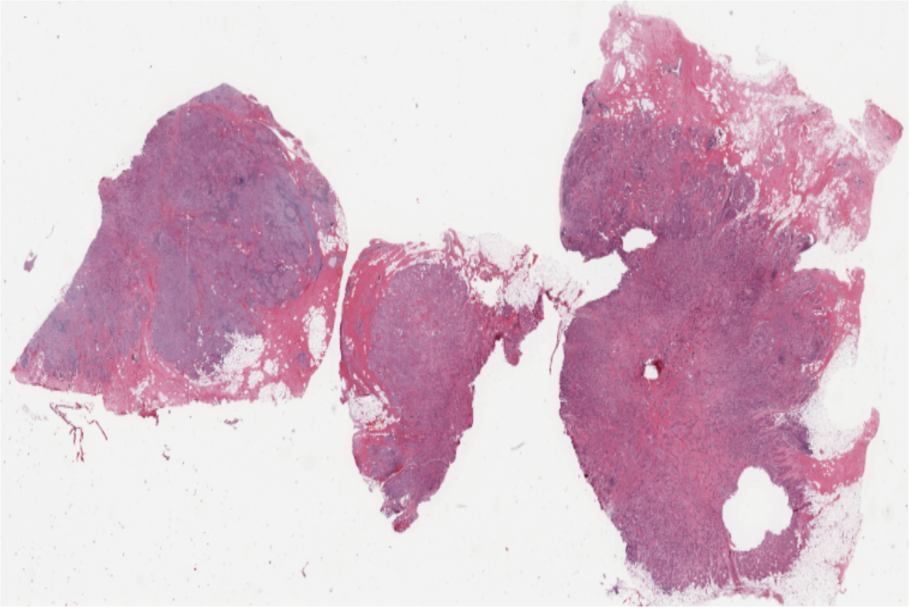


Figure 3: Examples of SVS image from TCGA-BRCA dataset

4.2. Data Pre-processing

In this study, we undertake a comprehensive pre-processing procedure to prepare histopathology images for deep learning analysis.

First, 200 SVS image files were collected from the TCGA-BRCA dataset, including images with BRCA1 or BRCA2 mutations, as well as some without these mutations. These images can be as large as 130,000x99,000 pixels. To facilitate efficient processing and analysis, the images were divided into smaller patches of 1400 x 920 pixels (see Figure 4 for exemples). Then each patch was classified according to the status of the BRCA mutation, distinguishing between the BRCA mutation positive and BRCA mutation negative cases. This classification is critical for investigating the role of BRCA mutations in breast cancer. The dataset was organized according to the BRCA mutation status, ensuring a comprehensive range of examples for model training. Subsequently, the dataset was split into training and validation sets to prepare for model evaluation (see Table 2 for the distribution of images).

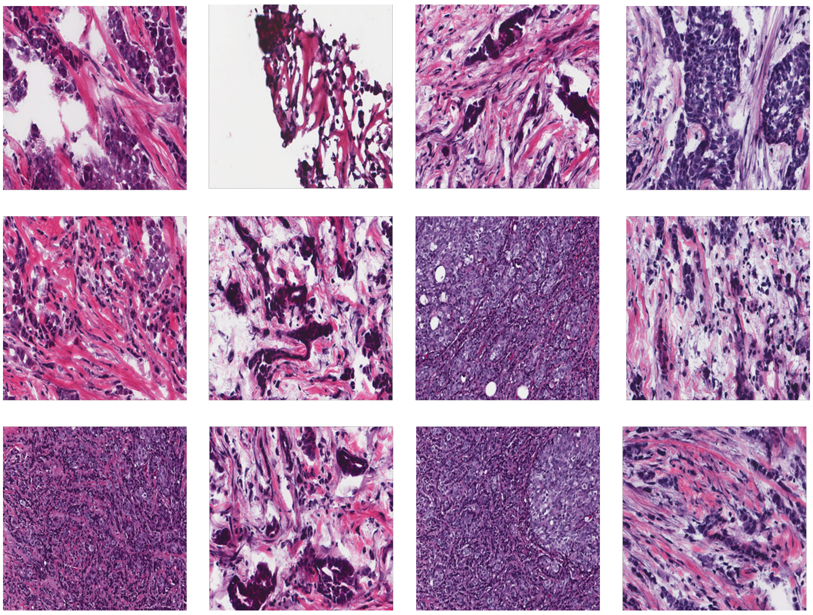


Figure 4: Example of generated patches from SVS images

Table 2
Dataset Statistics for BRCA Mutation Classification

Mutation Status	Number of SVS Images	Number of Patches
BRCA1	53	38,849
BRCA2	38	25,526
No BRCA Mutation	109	56,000
Total	200	120375

4.3. Used Metrics and Loss Functions

In this study, we use a variety of metrics and loss functions to evaluate and optimize our deep learning models for breast cancer detection and diagnosis. This includes the VQ-VAE model, which employs a specialized loss function. Below, we outline the metrics and loss functions used:

4.4. Loss Functions:

- **Binary Cross-Entropy Loss:** Applied for binary classification tasks, such as distinguishing between cancerous and normal patches. Measures the performance of a classification model with output probabilities between 0 and 1. The formula is:

$$\text{Loss}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right] \tag{2}$$

  where y_i denotes the ground truth label and $\hat{y}_i$ is the predicted probability.
- **VQ-VAE Loss Function:** The VQ-VAE model utilizes a specialized loss function that includes three key components:
  - **Reconstruction Loss:** Measures how well the decoder reconstructs the input from the quantized representation. It ensures that the reconstructed image is similar to the original input image. The formula is:

$$\text{Loss}_{\text{Recon}} = \frac{1}{N}\sum_{i=1}^{N}\|x_i - \hat{x}_i\|^2 \tag{3}$$

    where x_i is the original image and $\hat{x}_i$ is the reconstructed image.
  - **Codebook Loss:** Encourages the codebook vectors to move closer to the encoder output, ensuring that the quantization process effectively captures the data's structure. This component helps to learn a better representation by minimizing the distance between the encoder output and codebook vectors. The formula is:

$$\text{Loss}_{\text{Codebook}} = \frac{1}{N}\sum_{i=1}^{N}\|z_i - e_{q(z_i)}\|^2 \tag{4}$$

    where z_i is the continuous latent vector and $e_{q(z_i)}$ is the quantized vector.
  - **Commitment Loss:** Penalizes the encoder for not committing to a specific codebook vector, promoting stability in the learned representations. This component helps to stabilize the learning process and ensure that the encoder uses the codebook vectors effectively. The formula is:

$$\text{Loss}_{\text{Commitment}} = \beta\frac{1}{N}\sum_{i=1}^{N}\|z_i - e_{q(z_i)}\|^2 \tag{5}$$

    where β is a hyperparameter that controls the weight of the commitment loss term.

4.5. Evaluation Metrics:

- **Accuracy:** Measures the proportion of correctly classified patches out of the total number of patches. The formula is as follows:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{6}$$

- **Precision and Recall:** Precision measures the proportion of true positive predictions among all positive predictions, while Recall measures the proportion of true positive predictions among all actual positives. These metrics are defined as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{7}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{8}$$

- **F1 Score:** The harmonic mean of Precision and Recall, providing a balanced measure of model performance. The formula is:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

- **Area Under the Receiver Operating Characteristic Curve (AUC-ROC):** Measures the model's ability to distinguish between classes across all classification thresholds. A higher AUC value indicates better model performance. The AUC is calculated as the integral of the Receiver Operating Characteristic (ROC) curve:

$$AUC = \int_0^1 Precision\, d(Recall) \tag{10}$$

5. Results and discussion

The discussion section is dedicated to analyzing and interpreting the results obtained from our experiments:

5.1. Clustering Results

The clustering process using the DBSCAN algorithm [18] (Density-Based Spatial Clustering of Applications with Noise) aimed to categorize patches into specific groups based on the presence or absence of BRCA mutations. The parameters for DBSCAN were set with eps $= 8.7$ and min_samples $= 180000$, guiding the clustering of the latent space representations.

Table 3
Results of DBSCAN Clustering

Cluster	Total Number of Images	With BRCA Mutation	No BRCA Mutation
Cluster 1	91104	17467	73637
Cluster 2	32420	3611	25661

The goal of this clustering was to classify patches according to their status of BRCA mutation. The clustering process identified two distinct categories of clusters. The first cluster contains patches from both images with BRCA mutations and images without these mutations. The second cluster, however, exclusively contains patches from images identified with BRCA mutations. This clustering approach enables a more focused separation, supporting targeted analysis and model training based on the presence or absence of the BRCA mutation.

5.2. BRCA Patch Classification

In this section, we present the results of classifying BRCA patches using three different deep learning models: VGG16[19], ResNet [20], EfficientNet [21], and Inception V3 [22]. The classification task involves distinguishing between patches with BRCA mutations and those without. The data set used for the classification of BRCA mutation consists of histopathological image patches, divided into three subsets: training, validation, and testing. Table 4 summarizes the distribution of labels across the training, validation, and test sets.

The training set comprises a total of 57,152 samples, with 48,176 labeled as **NO_BRCA** and 8,976 as **BRCA**. The validation set includes 11,431 samples, of which 9,639 are labeled as **NO_BRCA** and 1,792 as **BRCA**. Finally, the test set contains 14,288 samples, with 12,044 labeled as **NO_BRCA** and 2,244 as **BRCA**.

The classification results for the detection of BRCA mutation in patches using four different models are presented in Table 5, which outlines key metrics such as precision, AUC, precision, recall and F1

Table 4
Dataset Summary for BRCA Mutation Classification

Dataset	Total Samples	NO_BRCA	BRCA
Training	64,384	55,408	8,976
Validation	11,431	9,639	1,792
Test	44,555	42,011	2,500

score for both the **BRCA** and **No Mutation** classes. These metrics provide a comprehensive evaluation of each model's performance, highlighting their ability to distinguish between patients with BRCA mutations and those without.

Table 5
BRCA Patch Classification Results

Model	Accuracy	AUC	Precision		Recall		F1-Score	
			BRCA	No Mutation	BRCA	No Mutation	BRCA	No Mutation
EfficientNet	98.81%	97.01%	98%	99%	94%	100%	96%	99%
VGG16	98.81%	97.17%	98%	99%	95%	100%	96%	99%
ResNet	98.71%	97.11%	97%	99%	95%	99%	96%	99%
Inception V3	**98.94%**	**97.36%**	**98%**	**99%**	**95%**	**100%**	**97%**	**99%**

All models exhibited exceptional performance with accuracy that exceeded 98%. **Inception V3** achieved the highest accuracy at **98.94%**, closely followed by EfficientNet and VGG16, both at **98.81%**, while ResNet achieved **98.71%**. The Area Under the Curve (AUC) further supports these findings, with all models surpassing the **97%** threshold, led by **Inception V3** at **97.36%**.

Detailed precision, recall, and F1-score reveal that **Inception V3** consistently outperformed the other models in all metrics. For the **BRCA** class, Inception V3 achieved a precision of **98%**, a recall of **95%**, and an F1 score of **97%**. For the **No Mutation** class, Inception V3 reached near-perfect performance, with a recall of **100%**, precision of **99%**, and an F1-score of **99%**. These results highlight Inception V3's balanced sensitivity (recall) and precision across both classes, making it a reliable model for BRCA mutation detection.

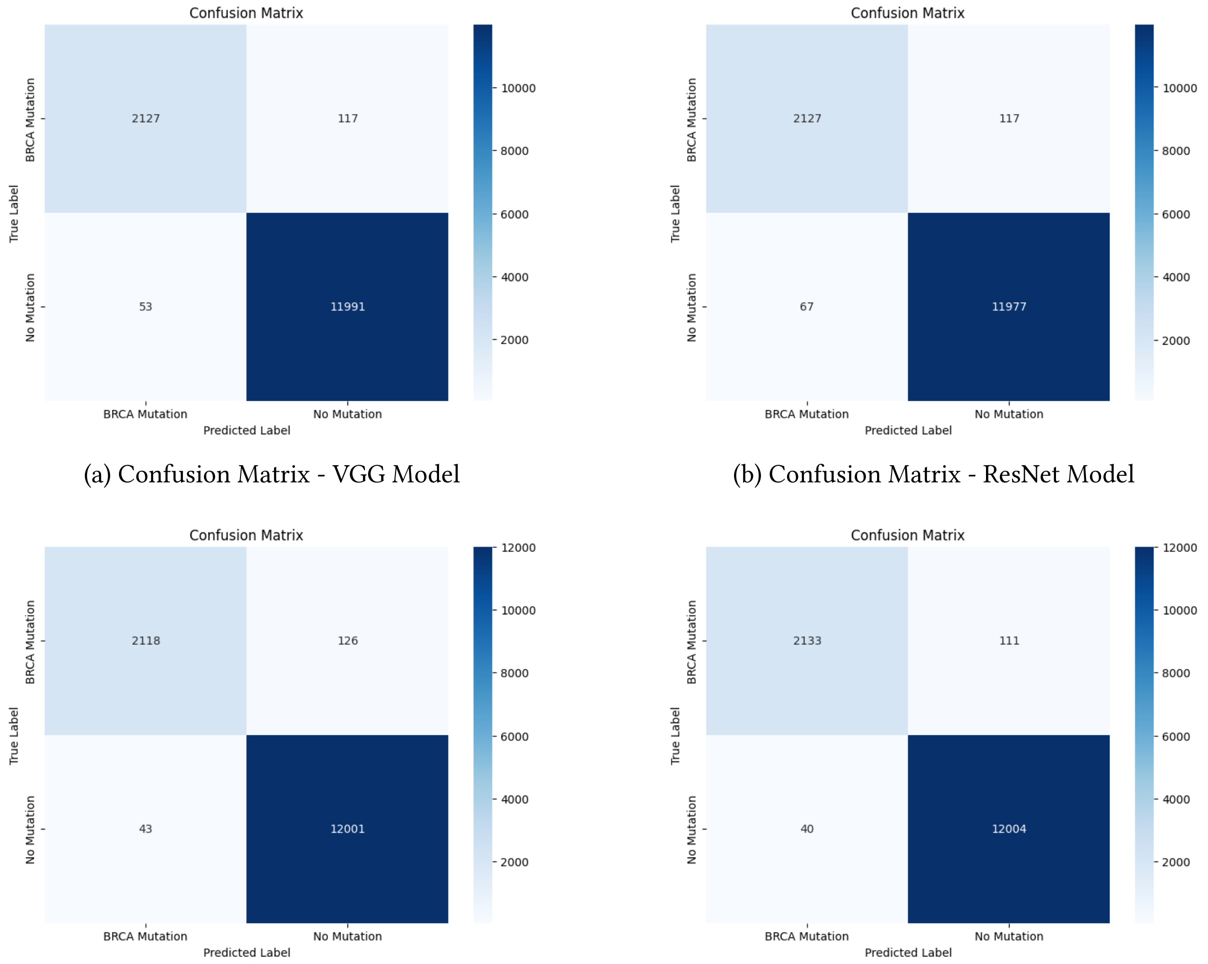
Although all models show strong performance, **Inception V3** stands out with the best overall metrics in accuracy, AUC, and F1 score. EfficientNet and VGG16 shared similar results, achieving an accuracy of **98.81%** and maintaining high precision and recall for both classes. ResNet, although slightly lower in performance compared to the other models, still achieved competitive results with a precision of **97%** for the BRCA class and high recall values.

The consistently high performance of all models underscores the effectiveness of deep learning architectures for histopathological image classification. However, the slight edge of **Inception V3** in both AUC and F1-score suggests that its architecture may be better suited for extracting subtle features in histopathological images related to BRCA mutations, possibly due to its ability to capture multi-scale features.

To further analyze the classification performance, confusion matrices for the four models—EfficientNet, VGG16, ResNet, and Inception V3—are illustrated in Figure 5, showing the distribution of true positives, false positives, true negatives, and false negatives for both the BRCA and No Mutation classes.

5.3. SVS Image Classification Results

The classification performance of our model for detecting BRCA mutations in histopathological SVS images is summarized in Table 5. The model achieved strong metrics for both the **BRCA Mutation** and **No Mutation** classes. Specifically, precision, recall, and F1-score for both classes were balanced, indicating robust classification results.



(a) Confusion Matrix - VGG Model

(b) Confusion Matrix - ResNet Model

(c) Confusion Matrix - EfficientNet Model

(d) Confusion Matrix - InceptionV3 Model

Figure 5: Comparison of Confusion Matrices for Different Models

Table 6
SVS Images Classification Results

Model	**Precision**	**Recall**	**F1-Score**
BRCA Mutation	90%	90%	90%
No Mutation	97%	97%	97%
Accuracy	95%	**AUC**	93.27%

As shown in Table 6, the model achieved an overall accuracy of 95%, with an AUC score of 93.27%. The high precision and recall for both classes demonstrate the effectiveness of our approach in detecting BRCA mutations, reducing the risk of false positives and false negatives.

5.4. Comparison with Related Work

Table 7 presents a comparison of our method with related work in the field of BRCA mutation detection from histopathological images. Our approach, combining VQVAE and DBSCAN with InceptionV3, outperformed previous studies, achieving the highest AUC of 93.27%.

Our approach offers several distinct advantages over other methods for the detection of BRCA mutations, primarily due to the integration of advanced unsupervised learning and clustering techniques. Using the VQVAE model, we efficiently encode high-dimensional histopathological images into a

Table 7
Comparison with Related Works

Reference	Dataset	Year	Methods	Metrics
Tristan Lazard et al. [11]	TCGA	2022	ResNet-18	AUC 71%
Xiaoxiao Wang et al. [10]	JSPHCM, JSCH	2021	ResNet-18	AUC 79%
Kurian et al. [13]	TCGA-BRCA	2023	SimCLR	81.34% accuracy
Valieris et al. [14]	TCGA	2020	ResNet-34	AUC 80%
Nam Nhut Phan et al. [12]	TCGA-BRCA	2021	2-Step ResNet50,101, VGG16, Xception	AUC 92%
Our Work	**TCGA-BRCA**	**2024**	**VQVAE, DBSCAN, VGG16, Resnet50, EfficientNet, InceptionV3**	**AUC 93.27%**

compact latent space, allowing the extraction of critical features while preserving key image details. Unlike conventional methods, which may struggle to capture subtle variations in tissue morphology, our model's ability to reconstruct intricate patterns enhances the detection of relevant features. Moreover, the incorporation of DBSCAN for clustering within this latent space adds a significant layer of robustness, effectively grouping similar patterns, and reducing noise. This method ensures that irrelevant or noisy data are filtered out, improving classification accuracy.

6. Conclusion

In this study, we present a pioneering deep learning framework designed to predict mutations of the BRCA gene in breast cancer utilizing histopathological images. By integrating Vector Quantized-Variational Autoencoders (VQ-VAE) for effective feature extraction and employing DBSCAN for clustering, we have established a robust model that demonstrates superior accuracy in classifying cases as BRCA mutation-positive or negative. This innovative approach surpasses conventional methods and highlights the potential of artificial intelligence to automate complex diagnostic processes within medical imaging.

In perspective, our future work will focus on the improvement of data enhancement techniques to further enhance the accuracy of the model in the detection of BRCA mutations. By generating synthetic samples that capture the variability in the expression of the BRCA mutation, our aim is to improve the robustness and generalization of our model. This will be particularly valuable for addressing imbalances in the dataset and improving the classification of rare mutation cases. In addition, our exploration will extend to investigating the roles of other genetic mutations in breast cancer.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT for grammar and spelling checks, as well as paraphrasing. After utilizing this tool, the authors reviewed and edited the content as necessary, taking full responsibility for the final publication.

References

[1] F. Touazi, D. Gaceb, M. Chirane, S. Herzallah, Two-stage approach for semantic image segmentation of breast cancer: Deep learning and mass detection in mammographic images., in: IDDM, 2023, pp. 62–76.

[2] M. Khaled, F. Touazi, D. Gaceb, Improving breast cancer diagnosis in mammograms with progressive transfer learning and ensemble deep learning, Arabian Journal for Science and Engineering (2024).

[3] F. Touazi, D. Gaceb, N. Boudissa, S. Assas, Enhancing breast mass cancer detection through hybrid vit-based image segmentation model, in: The 6th Conference on Computing Systems and Applications, Algiers, Algeria, 2024, pp. 1–10.

[4] R. A. Dar, M. Rasool, A. Assad, et al., Breast cancer detection using deep learning: Datasets, methods, and challenges ahead, Computers in biology and medicine 149 (2022) 106073.

[5] N. Petrucelli, M. B. Daly, T. Pal, Brca1-and brca2-associated hereditary breast and ovarian cancer (2022).

[6] A. Hodgson, G. Turashvili, Pathology of hereditary breast and ovarian cancer, Frontiers in Oncology 10 (2020) 531790.

[7] V. Talwar, A. Rauthan, Brca mutations: implications of genetic testing in ovarian cancer, Indian Journal of Cancer 59 (2022) S56–S67.

[8] L. F. Madrigal, M. Y. R. Garcés, F. J. J. Ruiz, Impact of non-brca genes in the indication of risk-reducing surgery in hereditary breast and ovarian cancer syndrome (hboc), Current Problems in Cancer 47 (2023) 101008.

[9] S. Zhao, C.-Y. Yan, H. Lv, J.-C. Yang, C. You, Z.-A. Li, D. Ma, Y. Xiao, J. Hu, W.-T. Yang, et al., Deep learning framework for comprehensive molecular and prognostic stratifications of triple-negative breast cancer, Fundamental Research (2022).

[10] X. Wang, C. Zou, Y. Zhang, X. Li, C. Wang, F. Ke, J. Chen, W. Wang, D. Wang, X. Xu, et al., Prediction of brca gene mutation in breast cancer based on deep learning and histopathology images, Frontiers in Genetics 12 (2021) 661109.

[11] T. Lazard, G. Bataillon, P. Naylor, T. Popova, F.-C. Bidard, D. Stoppa-Lyonnet, M.-H. Stern, E. Decencière, T. Walter, A. Vincent-Salomon, Deep learning identifies morphological patterns of homologous recombination deficiency in luminal breast cancers from whole slide images, Cell Reports Medicine 3 (2022).

[12] N. N. Phan, C.-C. Huang, L.-M. Tseng, E. Y. Chuang, Predicting breast cancer gene expression signature by applying deep convolutional neural networks from unannotated pathological images, Frontiers in oncology 11 (2021) 769447.

[13] N. C. Kurian, S. Varsha, A. Patil, S. Khade, A. Sethi, Robust semi-supervised learning for histopathology images through self-supervision guided out-of-distribution scoring, in: 2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE), IEEE, 2023, pp. 121–128.

[14] R. Valieris, L. Amaro, C. A. B. d. T. Osório, A. P. Bueno, R. A. Rosales Mitrowsky, D. M. Carraro, D. N. Nunes, E. Dias-Neto, I. T. d. Silva, Deep learning predicts underlying features on pathology images with therapeutic relevance for breast and gastric cancer, Cancers 12 (2020) 3687.

[15] A. van den Oord, O. Vinyals, K. Kavukcuoglu, Neural discrete representation learning, CoRR abs/1711.00937 (2017). URL: http://arxiv.org/abs/1711.00937. arXiv:1711.00937.

[16] A. Van Den Oord, O. Vinyals, et al., Neural discrete representation learning, Advances in neural information processing systems 30 (2017).

[17] A. Thennavan, F. Beca, Y. Xia, S. Garcia-Recio, K. Allison, L. C. Collins, M. T. Gary, Y.-Y. Chen, S. J. Schnitt, K. A. Hoadley, et al., Molecular analysis of tcga breast cancer histologic types, Cell genomics 1 (2021).

[18] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise, in: kdd, volume 96, 1996, pp. 226–231.

[19] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Conference on computer vision and pattern recognition, 2016, pp. 770–778.

[21] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.