3

METHODOLOGY

Cancer is a complex group of disease where the body's normal process of cell growth and division becomes disrupted, leading to the formation and proliferation of abnormal cells. These cancerous cells can infiltrate nearby tissues and organs. In certain instances, they can spread to distant regions of the human body via either the bloodstream or the lymphatic system (Kotamkar *et al*., 2021). Breast cancer is the most prevalent cancer among women globally, although it is also rarely seen in men. It originates from the lobules or the milk ducts responsible for milk production (Tiwari *et al*., 2021). While the precise breast cancer etiology remains unclear, numerous risk factors have been identified. These include age, genetic mutations like *BRCA1* and *BRCA2*, familial history of breast cancer, early onset of menstruation or delayed menopause, dense breast tissue, estrogen exposure, and a sedentary way of life (Singh and Kumar Sain, 2023).

Epidemiological studies are scientific investigations that aims to identify and understand sequences and factors that influence health and disease in populations. In the realm of breast cancer, epidemiological studies concentrate on pinpointing risk factors linked to the development of the disease (Smolarz *et al*., 2022). The insights gained from these investigations are commonly utilized to shape public health initiatives, raise awareness, and guide ongoing research efforts aimed at preventing, detecting early, and treating breast cancer (Łukasiewicz *et al*., 2021).

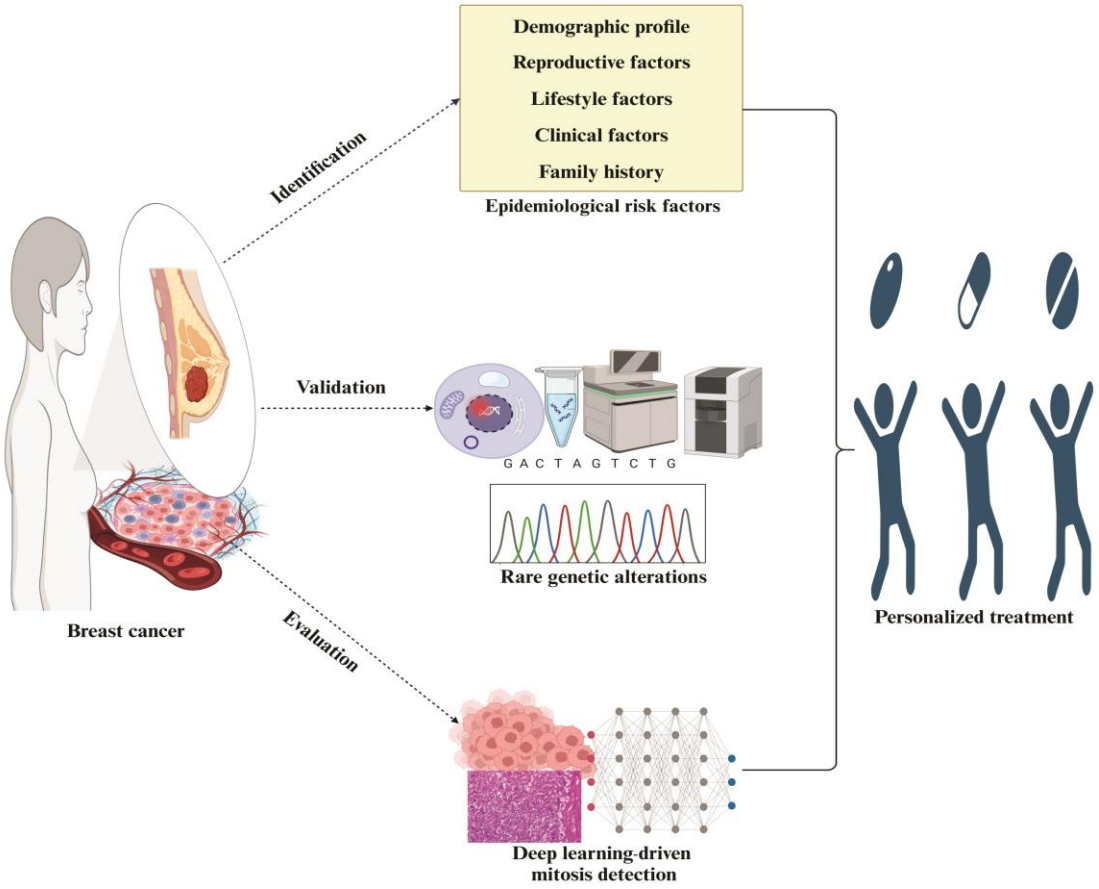
In the context of breast cancer research, exome sequencing can identify novel candidate genes that are believed to be involved in disease development and progression (Lee *et al*., 2022). These crucial genes provide insights into the epidemiological or genetic factors underlying the progression of the disease. These discoveries can pave the way for creating novel diagnostic methodologies, tailored therapies, and individualized treatment modalities in the future (Koivuluoma *et al*., 2021).

Histopathology images are critical for cancer diagnosis as well as therapy. One way to evaluate the breast cancer stages is by quantifying the mitotic score,

which measures the number of mitotic figures (cells undergoing division) in a histopathology image (Mathew *et al*., 2021). Deep learning-based mitotic score quantification in histopathology images provides an automated and objective based approach to assessing the tumor aggressiveness, which allows more accurate prognosis and treatment decision-making (Pan *et al*., 2021).

Hence, this study focused on the integrative analysis of breast cancer to identify the risk factors, rare genetic variants and proliferation of cells associated with the disease in specific-population. The outline of the proposed research study is illustrated in **Figure 8**. The figure was created with BioRender.

Figure 8: Overview of the proposed research



The objectives set for the present study were carried out by adopting the following methodology.

3.1 Layout of the study

The study was structured into four phases to accomplish the objectives laid down for the present study.

- Phase I was designed to conduct an epidemiological study to identify and understand the risk factors that may cause breast cancer. This phase involved collecting and analysing data from the hospital based cohort of breast cancer patients to identify patterns, correlations, and potential factors contributing to breast cancer development.
- In Phase II, the study aimed to explore rare genetic variants causing breast cancer by WES. Examining the genetic makeup of breast cancer patients helps to identify specific genetic mutations that might play an important role in the disease in specific-population.
- Phase III involved validating the novel genetic variants in breast cancer patients. This phase focused on confirming the presence and significance of the genetic variants.
- Lastly, the study's Phase IV focused on developing a deep learning-based automated approach for detecting mitosis in breast cancer histopathological images.

Phase I

3.2 Epidemiological study on breast cancer associated risk factors

In phase I, we attempted to collect detailed information from breast cancer patients in specific populations, including their personal details, lifestyle choices, and reproductive and medical history. The associations between these factors and the likelihood of developing breast cancer was explored.

3.2.1 Study population

The study concentrated on a cohort of primary breast cancer patients and those under treatment at the Department of Oncology, Sri Ramakrishna Hospital, Coimbatore, Tamil Nadu, between January 2021 and May 2023.

Inclusion criteria of the study included women aged $\geq$ 20 years and diagnosed with breast cancer. Exclusion criteria for the study were women who were not affected by breast cancer or any other cancers.

3.2.2 Ethical clearance

Ethical clearance for data and sample collection for the study was obtained from the Ethics Committee of Ramakrishna Hospital, Coimbatore (EC/2019/0411/CR/57) and Institutional Human Ethics Committee of Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore (AUW/IHEC/BT-22-23/FHP-02). These esteemed institutions' reviewed and approved the process and ensured the study adhered to ethical guidelines and protocols.

3.2.3 Study variables

A structured questionnaire was designed and deployed to gather data on various factors affecting breast cancer patients. Demographic characteristics, including age, education, gender, marital and employment status, were collected. Reproductive factors, such as menarche age, pregnancy history, duration of breastfeeding, age at menopause, and history of abortions, were documented. Lifestyle and occupational risk factors, including the consumption of fresh fruits and vegetables, meat, alcohol, and smoking habits, exercise patterns, water intake, sleep patterns, and exposure to radiation or chemicals, were assessed. Clinical characteristics, such as comorbidities, treatment modalities for breast cancer, side effects of cancer therapy, and previous surgical history, were collected. Additional information in case of family history of breast cancer patients was also documented.

3.2.4 Sample size and sampling method

The sample size for the study was determined using the Raosoft sample size calculator, considering a desired response ratio of 50%, a confidence level of 95%, and a margin of error of 5%. Based on these parameters, the recommended number of respondents was calculated. A representative sample of the larger population was chosen by random sampling.

3.2.5 Statistical analysis

The data analysis was performed using the Statistical Package for Social Sciences (SPSS Inc., Chicago) version 20. Descriptive analysis was employed to analyse the variables of the breast cancer patients. Mean, standard deviation and percentage were calculated to identify the average and proportion of the variables. The Pearson chi-square (χ^2) test was used to compare the frequency distribution of the variables, with a p-value of ≤ 0.05 indicating statistical significance. The range between risk variables and breast cancer was estimated using a 95% confidence interval.

The identified association between the epidemiological factors and breast cancer risk is the foundation for subsequent study phases. The insights gained from phase I improved our understanding and highlighted the need to develop targeted interventions and personalized breast cancer prevention and management approaches.

Phase II

3.3 Identification of rare variants in breast cancer patients

In phase II, an in-depth genomic analysis was done to analyse and focus on the genetic profiles of primary breast cancer patients. The genomic information was then subjected to thorough analysis using bioinformatics tools to identify rare genetic variants. The identified genes and variants were found to significantly contribute to the development and progression of breast cancer. The findings from phase II, when combined with the comprehensive epidemiological data from phase I contributed to a concrete understanding of the complex interplay between genetic, lifestyle, and environmental factors in breast cancer etiology and progression.

3.3.1 Sample collection

The tumor tissue samples from 6 primary breast cancer patients (5-Female, 1-Male), and 2 adjacent normal breast tissues were collected and stored in phosphate-buffered saline at -20°C for DNA isolation.

3.3.2 DNA isolation

DNA isolation is a crucial step, enabling the investigation of the underlying genetic factors contributing to the development and progression of tumors. DNA was

isolated from tumor tissue using QIAamp DNA Mini Kit (CA, USA). The detailed procedure is given in Appendix-I.

3.3.3 Quantification and purity of isolated DNA

Quantification and purity assessment are essential steps in DNA research and applications. DNA quantification was performed using Nanodrop (NanoDrop™ 1000 Spectrophotometer, Thermofisher). The absorbance ratio at 260 nm to 280 nm (A260/A280) gives the purity of the DNA sample.

3.3.4 DNA gel electrophoresis

The quality of DNA is essential in molecular biology and genetics research, providing valuable insights into genetic information. The quality of the isolated DNA was checked using agarose gel electrophoresis, and the method is outlined in Appendix-II.

3.3.5 Whole exome sequencing of isolated tumor DNA

Whole exome sequencing (WES) is a high-throughput DNA sequencing technique that selectively targets and sequences the genome's protein-coding regions. The exome represents the portion of the genome that encodes proteins and is responsible for most disease-causing mutations. The study examined the genetic alterations within the exome of breast cancer patients to elucidate the molecular mechanisms driving the disease. The WES was done at Neuberg Supratech, Ahmedabad. The sequencing was performed using paired-end (PE) reads with a read length of 150 base pairs. The platform used for sequencing was Illumina NovaSeq 6000. Additionally, the sequencing depth covered per sample was >8GB. The procedure is explained in Appendix-III.

3.3.6 Functional enrichment analysis of rare variant candidate genes

The identified variant candidate genes were analyzed using the FunRich software tool (http://www.funrich.org). This bioinformatics tool was pivotal in determining the functional implications of the genetic variations by investigating their involvement in biological processes, cellular components, and molecular functions. The genes and variants identified in phase II provided a molecular basis for understanding the underlying mechanisms associated with primary breast cancer. These genomic findings can offer insights into potential biomarkers, therapeutic targets, and pathways advancing precision medicine and tailored treatment approaches.

Phase III

3.4 Validating the variants of uncertain significance mutations

In phase III, we further validated the identified new variants. Characterizing the functional consequences of the novel genetic variations revealed their potential relevance in disease genesis and progression. This phase provided critical insights into the mechanisms through which specific genetic alterations contribute to the pathophysiology of breast cancer.

3.4.1 Primer designing

Since we are reporting the new variants for the first time, no specific primers were available. So, we designed the specific primers for the new variants, enabling efficient and accurate amplification for sequencing. The steps involved in primer designing include the following:

- **FASTA sequence:** The specific chromosome of interest was located, and the FASTA sequence for that chromosome was retrieved from the NCBI (National Center for Biotechnology Information) website (https://www.ncbi.nlm.nih.gov).
- **Flanking sequence:** Using the Ensembl (Genome Browser) (https://www. ensembl.org), the flanking sequence was obtained. Comparing the FASTA sequence and flanking sequence, the mutated region of interest was identified. The flanking sequence was used for primer designing.
- **PCR primers:** The Primer3Plus software (https://www.primer3plus.com) was utilized to design specific PCR primers. The appropriate primer design parameters were selected after inserting the flanking sequence acquired from Ensembl. The parameters included an optimal melting temperature of 60°C, a primer length of 20 base pairs, and a product size below 500 base pairs. These specifications were chosen to ensure efficient and specific PCR amplification of the target region.
- **Amplicon specificity:** The UCSC *in silico* PCR tool (https://genome.ucsc.edu/) was utilized to verify the specificity of the designed primers. The primer sequences were provided to ensure the amplicons were specific to the intended target.

- **Sequence similarity:** The Nucleotide-BLAST tool on the NCBI website was employed to search for sequence similarity between the designed amplicons and the human genome. The designed forward and reverse primers for the variants are given in **Table 1**.

Table 1: Primer sequence of specific variants

Variants	Forward primer	Reverse primer	Product size
MRPL13: c.380T>C	CCTTGTTCAAAAGCAAGCAGTG	GCTTTCTTACTATCGCAAGTCAC	407bp
MYBPC3: c.2816G>A	CACCTCCACTGGACACCAAG	ATCAGAGGAGTGGGCAGTGG	316bp
PTCH1: c.1889G>A	GTGTCCACTTCGTACAGGGG	CCTACACCGACACACACGAC	309bp

3.4.2 PCR amplification

Polymerase chain reaction (PCR) is a powerful tool commonly used for DNA amplification. Touchdown PCR was done to amplify specific regions of the identified novel variants from WES. The procedure is given in Appendix-IV.

3.4.3 Sanger sequencing

Sanger sequencing was mainly used for validating novel variants in breast cancer. It is a valuable tool to confirm the presence of mutations or genetic variants in the gene. The Sanger sequencing was done using Applied Biosystems 3500 Genetic Analyzer at Bharathiar University, Coimbatore. The detailed procedure is given in Appendix-V.

3.4.4 Interactome network mapping

The interactome network mapping of novel candidate genes was done using Network Analyst (https://www.networkanalyst.ca) to analyse the gene-gene interactions and to understand the functional relationships and molecular interactions among the candidate genes.

3.4.5 KEGG pathway enrichment analysis

The KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway Database (https://www.geome.jp/kegg/pathway) provides a collection of metabolic pathways,

signaling pathways, and other functional modules. It was used to understand the biological context and pathway associated with the novel candidate genes validated by Sanger sequencing. This integration of genomic data with pathway information aids in deciphering the complex regulatory networks and functional implications associated with the identified genes, ultimately advancing the knowledge of gene function and its relevance to physiological and pathological conditions.

The biological consequences of specific genetic alterations were validated in phase III, and their potential impact on breast cancer was assessed. This phase deepens the understanding of the molecular mechanisms involved. It paves the way for translating these findings into clinically relevant insights for improved diagnostics and targeted therapy within the studied population.

Phase IV

3.5 AI-based automated system for mitosis detection in breast cancer histopathological images

Understanding the mechanisms driving cell proliferation in breast cancer is fundamental for developing tailored therapeutic interventions. Through sequencing and bioinformatics studies, we have identified mutations that disrupt cell growth and division regulatory mechanisms. Elevated mitotic activity in tumors is linked to a poorer prognosis. So, in the fourth phase of our research, we employed deep learning to detect mitosis in histopathology images of breast cancer. Mitotic detection through a manual process is often a time-intensive process. In our study, we have implemented AI-based mitotic detection to address this challenge, which improves diagnostic accuracy and reduces overall time consumption.

3.5.1 Dataset

The ICPR 2014 dataset (MITOS-ATYPIA-14 grand challenge) (https://mitos-atypia-14.grand-challege.org) was used to train the model containing histology images with mitosis region. This dataset consisted of 1200 training and 496 test breast cancer histopathology images obtained from 16 distinct biopsies. Each image had dimensions of 1539 × 1376 pixels at a magnification of 40×. These images exhibit diverse factors, including tissue acquisition

procedures, staining techniques, and lighting conditions. To achieve exceptional performance on this dataset was a challenge for the study. The histopathology images were observed using the AperioScanScope XT slide scanner, and a total of 267 breast cancer histopathology images were collected from the dataset for the study.

3.5.2 Real-time histopathology images

The real-time histopathology images were obtained from the six breast cancer patients whose genetic alterations were previously studied in whole exome analysis. A total of 31 real-time breast cancer histopathology images were utilized for the study.

3.5.3 Experimental setup

The proposed method was executed utilizing the sci-kit-learn machine learning package on a desktop computer with an operating system of Windows 10.

3.5.4 Training and testing of CNN (Convolutional Neural Network)

CNN-based model has great potential to improve diagnostic accuracy and efficiency to detect mitosis in breast cancer histopathology images. CNNs are particularly effective in image classification tasks because they can extract important characteristics from images.

3.5.4.1 Data preprocessing

A total of 298 images were used for training and testing the dataset. The images with their corresponding atypia scores were scaled based on the low average and high confidence scores. The dataset was randomly classified into two sets: 80% for training and 20% for testing/validation.

3.5.4.2 Architecture of the developed CNN model

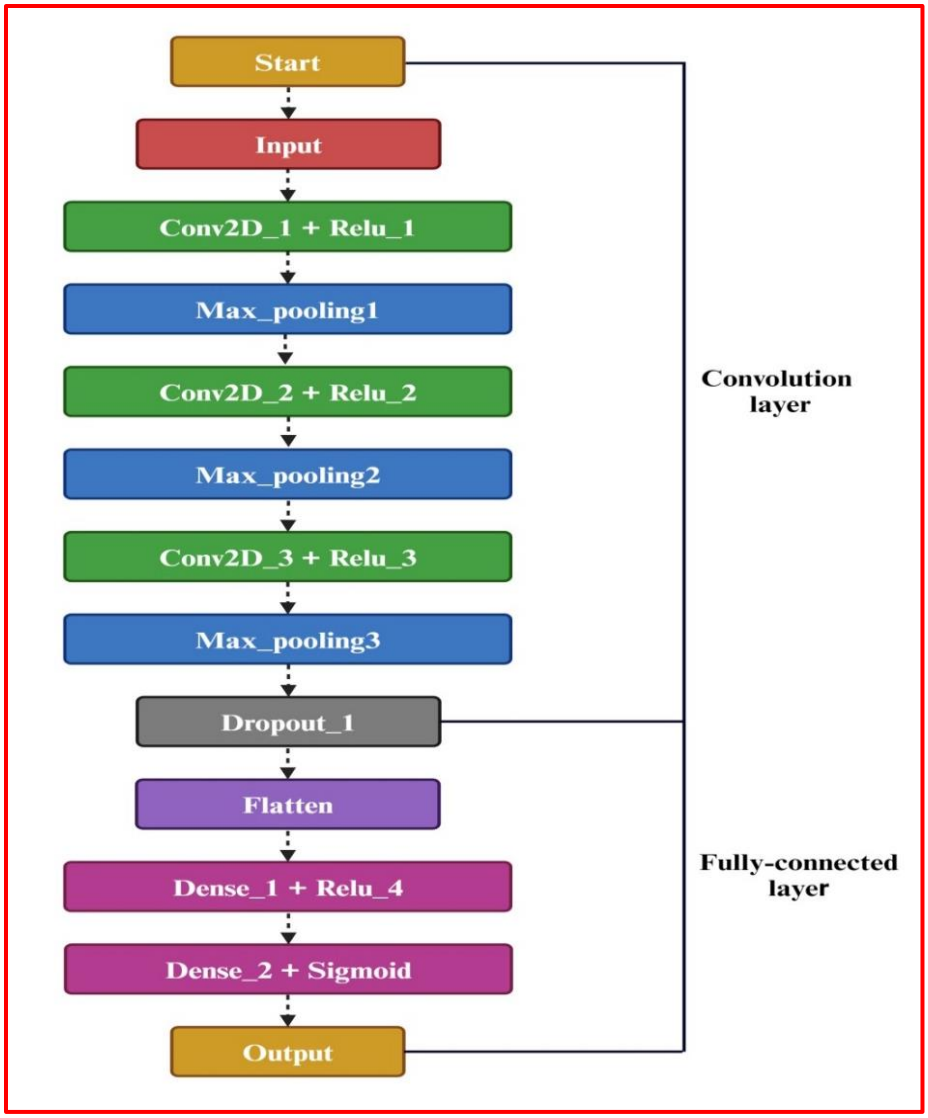
The architecture of the CNN consists of input, convolution and output layers, as detailed in **Figure 9**.

- **Input layer:** The input images were of size 240x240 with three colour channels (RGB).
- **Convolutional layers:** The convolutional layers, specifically conv2d_1, conv2d_2, and conv2d_3, extracted features from input images. These layers learned to detect

patterns and structures within the images, such as edges, textures, and more complex traits related to mitotic figures in histopathology images.

- **ReLU activation:** Following each convolutional layer, the rectified linear unit (ReLU) activation function was applied. This introduced non-linearity to the model, enabling it to learn complex relationships between the extracted features.
- **Max-pooling layers:** The max-pooling layers, max_pooling1, max_pooling2, and max_pooling3, reduced the spatial dimensions of the feature maps while retaining the vital information. This downsampling helped to reduce computational complexity and increase the model's ability to focus on significant features.

Figure 9: Flow chart of the CNN Architecture



- **Dropout layer:** Dropout was employed to mitigate the risk of overfitting, a common challenge in deep learning. During training, the dropout_1 layer randomly sets a fraction of input units to zero in each update, preventing the network from becoming overly specialized on the training data and increasing its ability to generalize to new, unseen data.
- **Flatten layer:** This layer reshaped the 3D tensor generated by the preceding layers into a 1D vector. This flattening step prepared the data for input into the fully connected layers.
- **Fully connected layers:** Two fully connected (dense) layers were added to the network architecture. The initial dense layer comprised 128 neurons utilizing ReLU activation, while the final dense layer consisted of a single neuron employing sigmoid activation for binary classification. The probability score indicated the likelihood of a sample belonging to a particular class in breast cancer diagnosis.

3.5.4.3 Parameters and training

The total number of parameters trained is 7,401,697, calculated based on the weights and biases of the model's layers. The Adam optimizer was used with a learning rate of 0.001. Adam was an optimization algorithm that adapted learning rates per parameter, leading to efficient training with 'binary_crossentropy' for loss function to detect mitosis. The model was trained for 25 epochs (complete passes through the training dataset) using the provided training data (x_train and y_train).

3.5.4.4 Performance evaluation

The evaluation of the developed method's efficacy relies on its ability to identify mitotic cells accurately. Following the competition's standards, correct detection is determined by identification within a specified proximity of the ground truth position. Specifically, an optimistic prediction is considered a true positive if it is located within 5 µm (equivalent to 20 pixels) and 8 µm (equivalent to 32 pixels) of the actual ground truth position, as outlined by the ICPR 2014 dataset.

The accuracy score was computed based on these criteria by dividing the number of accurate predictions by the total number of predictions made. This metric provides a quantifiable measure of the model's performance in correctly identifying mitotic cells within the specified proximity range, as prescribed by the competition's evaluation guidelines.

$$\text{Accuracy} = \frac{TP + TN}{P + N}$$

Where,

TP - True Positives (correctly identified mitotic cells)

TN - True Negatives (correctly identified non-mitotic cells)

P - Total Positives (total number of mitotic cells)

N - Total Negatives (total number of non-mitotic cells)

The algorithm of the developed CNN model is outlined in Appendix-VI. An integrated framework incorporating a more extensive set of histopathology images was developed using deep learning algorithms. The system utilizes deep learning algorithms, demonstrating remarkable performance in medical image analysis. This framework provides both pathologists and clinicians with an efficient and accurate visualization of breast cancer progression and facilitates informed treatment decisions in a time-effective manner.

The results obtained and the discussion and interpretation of data are presented in the next chapter.