Early Detection of Breast Cancer Based on HER-2 DNA Genomic Sequence

S. G. Shaila(✉), Vijayalaxmi Inamdar, Ganapati Bhat, K. Hithyshi, and Arya Suresh

Dayananda Sagar University, Bangalore, Karnataka 560068, India
{shaila-cse,ganapati.bhat}@dsu.edu.in

Abstract. Breast cancer has become the most frequently occurring cancer that has put the lives of women at risk globally. Breast cancer is mostly caused by environmental causes such as genetic mutations, lifestyle etc. Cancerous cells may have alterations in the DNA sequence. HER2 (Human Epidermal Growth Factor Receptor 2) gene mutations are found in 20–30% of breast cancer cases. This gene is a proto-oncogene having tyrosine kinase activity, which makes proteins that are responsible for healthy growth and division of cells. Mutation in this particular gene produces multiple copies of a protein that makes the breast cell grow uncontrollably. Identifying such mutations is a challenging task. The most accurate methods of DNA analysis and finding are frequently found with the many sorts of procedures or methodologies employed for analysis. This research study will discuss how early detection and diagnosis of breast cancer helps when mortality is taken into account. Patients should undergo focused therapy, which will boost their chances of recovery. The paper aims to identify the mutation caused by the HER2 gene and extract the position of the mutation in the gene. For better results, protein sequence analysis is done and the position of mutation in the protein sequence is identified. This can be beneficial to give targeted treatments and to diminish the recurrence of breast cancer in a patient.

Keywords: Breast Cancer · DNA Sequence · Gene · HER2 · Classification

1 Introduction

Cancer has become one of the main reasons for mortality globally. Breast cancer is the second most life-threatening type of cancer in women. It is the most common malignant tumor that causes death in women. Breast cancer can be caused by a variety of factors, including lifestyle, family history, genetic mutations in genes such as HER2 (Human Epidermal Growth Factor Receptor 2), BRCA1 (Breast Cancer gene 1), and so on. Some of the symptoms may include lumps in the breast, epidermal tissue dimples, breast shape changes, fluid oozing through nipples, sunken or inverted nipples. Although there is no effective approach for preventing breast cancer, predicting and diagnosing breast cancer in the early stages can be considered beneficial, taking mortality into consideration. Doing this will improve the chances of recovery as the patients can receive targeted treatments. In almost 30% of breast cancer cases, it can be seen that there is a mutation

S. Tamane et al. (Eds.): ICAMIDA 2022, ACSR 105, pp. 448–455, 2023.
https://doi.org/10.2991/978-94-6463-136-4_38

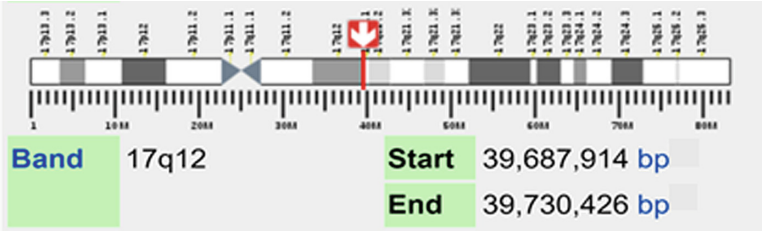

Fig. 1. Position of HER2 gene in chromosome 17 (https://en.wikipedia.org/wiki/HER2)

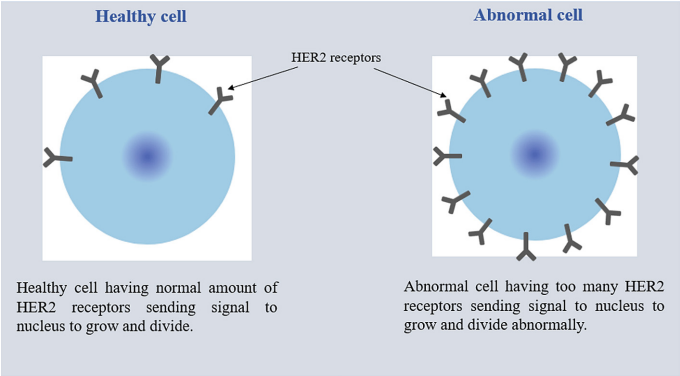

Fig. 2. Overexpressed HER2 receptors

in the HER2 gene, which is located at the long arm of chromosome 17 (17 q12) in the DNA sequence. This is represented in Fig. 1.

HER2 belongs to the family of human Epidermal Growth Factor Receptors (EGFR) and is a proto-oncogene having tyrosine kinase activity. This is depicted in Fig 2 below. It makes the HER2 protein, which regulates growth, division, and survival of healthy breast cells. However, when this gene is amplified, it produces multiple copies of itself, which in turn makes too many copies of the protein that makes the breast cells divide and grow uncontrollably. Such a type of cancer where breast cancer cells carry amplified genes that produce overexpressed proteins is called HER2 positive breast cancer. This type of cancer tends to grow faster, spread, and have a higher chance of recurrence than HER2-negative breast cancer.

When the HER2 mutation gene is identified in patients, targeted treatments can be given at an early stage. It is also important to prevent the recurrence of breast cancer. In this work, an approach is proposed to predict the presence of breast cancer by identifying the presence of a mutation in the HER2 gene and protein sequence and extracting the position of the mutation.

2 Literature Survey

Breast cancer is a common type of cancer seen in females of all age groups and is caused for various reasons like age, lifestyle, genes, etc. This section covers various reviews written by many authors. The authors in [1] have analysed breast cancer data collected from The Cancer Genome Atlas Network (TCGA) to identify two potential genes that can be used to differentiate between triple-negative breast cancer and non-triple-negative breast cancer. In the paper [2], the authors proposed a method named Genome Deep Learning (GDL) to understand the relationship between genomic variations and traits and to identify 12 different types of cancer. They have used Whole Exon Sequencing (WES) mutation files of 6083 samples from 12 different cancer types to train 12 specific models to differentiate certain types of cancer. In their work [3], the authors document the changes that have contributed to a better assessment of HER2 status and discuss the impacts of HER2 mutation. This study explains the clinical significance of HER2 mutations, especially the introduction of the efficacy of specific anti-HER2 agents in future categories of HER2 low breast cancer, especially in the breast cancer subgroup. The authors in [4] discuss the methodology to detect red and green signals in fluorescent In-Situ Hybridization (FISH) images. In this study, the authors presented an accurate cell nucleus segmentation method and signal detection methods for detecting red and green spots in segmented cells. In the work [5], the authors stated that this finding prompted them to look into epigenetic factors like DNA methylation associated with gene expression in order to discover epigenetic biomarkers for subtypes of breast cancer. Using differential analysis, they could identify a set of upregulated and downregulated genes for each subtype. In the paper [6], the authors focused on changes to the ASCO/CAP guidelines for interpreting the status of HER2 in breast malignancies, as well as certain pitfalls to avoid. Treatment decisions and responses may be influenced by changes in HER2 status. Therefore, HER2 status should be reassessed in post-NAC samples and metastatic lesions. The authors [7] presented a new method for classifying breast cancer patients based on subtypes and survival rates using The Cancer Genome Atlas (TCGA) genomics data. They have extracted unimodal and multimodal features from every input, trained several machine learning models, performed model averaging, and made final predictions. In the paper [8], the authors stated that while HER2 testing is well established to guide appropriate breast cancer treatment, the results of immunohistochemistry (IHC) and fluorescence in situ hybridization (FISH) were in a small number of cases. The high prevalence of ambiguous double HER2 equivocal cases and the disparity between IHC and alternative FISH tests suggest that alternative FISH tests using RAI1 and TP53 probes are needed for clear classification. In [9], the authors proposed the HER2 deep neural network (Her2Net) based on deep learning to detect, segment, and classify cell membranes and nuclei from HER2-stained breast cancer images. The authors in [10] studied the characteristics of mutated plasma ctDNA samples collected from advanced breast cancer patients by applying a deep tissue sequencing method to detect somatic mutations to analyze the interconnection of treatment history and clinical features with genomic variations.

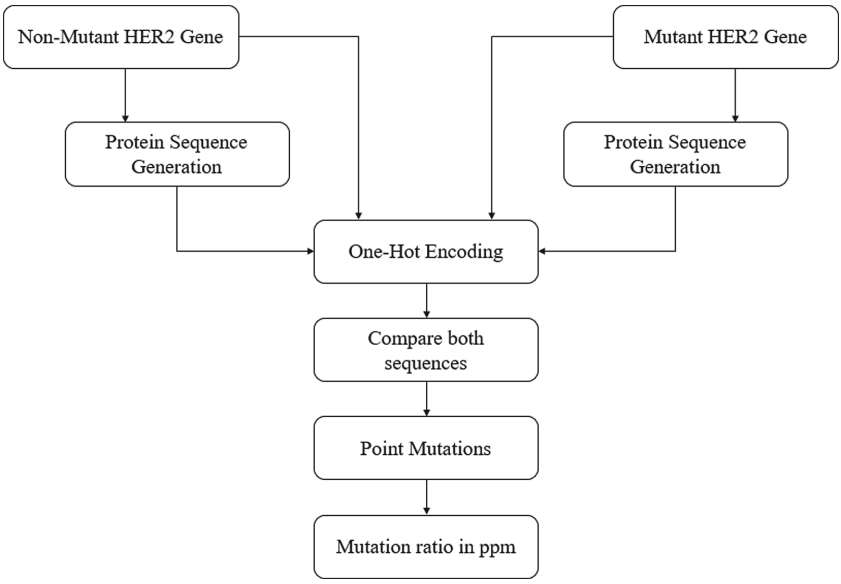

Fig. 3. Proposed Model for identifying mutation in HER2.

3 Proposed Work

The proposed model for this work is represented in Fig. 3. The dataset for the model is collected from the National Center for Biotechnology Information (NCBI).

This data is in 'fasta' sequence format and is preprocessed to remove unwanted data. The sequence obtained is further encoded using One-Hot Encoding. The sequence is compared with another sequence which is also preprocessed and encoded. This comparison is done to identify the changes in both the sequences. The changes in the sequence assume that there is a mutation in the gene which is classified as cancerous. On the other hand, no change in the sequence means no mutation, which is classified as non-cancerous.

3.1 Data-Set Description

The dataset collected is the HER2 gene sequence of human DNA represented in fasta sequence format and collected from the National Center for Biotechnology Information (NCBI). The details of the dataset are represented in Table 1. The HER2/ERBB2 gene belongs to the tyrosine kinase family of Epidermal Growth Factor Receptor (EGFR). This protein does not have its own ligand-binding domain and therefore cannot bind to growth factors. However, it binds tightly to other members of the ligand-binding EGFR family, stabilizing ligand binding by forming a heterodimer that facilitates activation of downstream signaling pathways via kinases. Studies have shown that amplification and/or overexpression of this gene can be seen in a variety of cancers, including breast and ovarian cancer.

Table 1. Details of the NCBI Dataset

Dataset	Length of sequence
Normal HER2 Gene-sequence	41147
Mutated HER2 Gene sequence	41147

3.2 Pre-processing and Encoding

The approach collects the data and preprocesses it to remove the noisy data and redundancy in the dataset. The collected dataset is in fasta sequence format of the HER2 gene sequence. Furthermore, the sequence is subjected to encoding to identify the gene sequence mutation. The approach considers two HER2 sequences—the normal HER2 sequence and the mutated gene sequence. Both the sequences are encoded. A binary vector is obtained for both the sequences and a further comparison is done to check the mutation for the sample sequence.

3.3 Gene and Protein Sequence Encoding

To process genomic sequences, it is important to convert categorical sequences into binary sequences to build a matrix input training model. One-Hot Encoding is one of the sequential coding methods. It is the process of transforming categorical data variables into a form that can be provided to machine learning algorithms or deep learning algorithms to improve the accuracy of model prediction and classification. It is used for processing DNA sequences, as it is required to convert an array of strings into a numerical sequence to build the matrix input model. The dataset collected from NCBI is preprocessed and encoded using the One-Hot Encoding scheme. The mutated data is also encoded using the One-Hot Encoding scheme. DNA has four bases in the sequence, namely, adenine (A), thymine (T), cytosine (C), and guanine (G). These bases are encoded as a vector using the On-hot encoding technique. 'A' is encoded as (0,0,0,1); 'C' as (0,0,1,0); 'G' as (0,1,0,0); and 'T' as (1,0,0,0). After encoding the sequences, the resultant sequences are compared to check for the mutation. If the sample sequence is not the same as the normal sequence, then it is mutated, which means that the cell containing this gene sequence is cancerous. However, when handling gene sequences, which are large and consume a lot of memory, execution time is also longer. To overcome this, protein sequences can be considered. Each protein sequence is made up of twenty amino acids, and each amino acid is made up of three codons. This makes the sequence more complex and can minimize the occurrence of silent mutations as not all mutations in the gene sequence might result in amino acid changes. A one-hot encoding method is applied to get the resultant binary sequence after converting a gene to a protein sequence. The mutation percentage is calculated in terms of parts per million. It can be obtained by computing the number of mutations in a gene sequence against the length of sequence as shown in Eq. (1).

$$Mutation(\%) = \frac{mutation\ count}{length(HERE\ 2\ gene)} * 1000000 \tag{1}$$

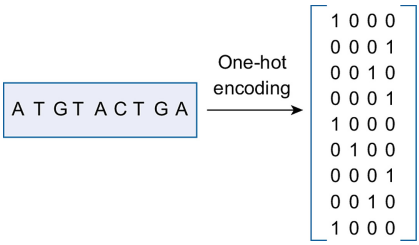

Fig. 4. Example of a DNA sequence encode using One-hot Encoding technique.

Table 2. Performance evaluation of proposed approach with other approaches

Approach	Method used	Mutation in Percentage
Yi et al.	Tissue sequence analysis	0.3%
Proposed approach	Gene sequence analysis	0.015%
Proposed approach	Protein sequence analysis	0.022%

Using the above equation, the mutation ratio is calculated for both the gene sequence and the protein sequence.

4 Experimental Results

The dataset collected from NCBI was in "fasta" sequence. This sequence was preprocessed to extract only the gene sequence, which was of length 41147. The obtained sequence is then encoded using the One-Hot Encoding scheme. Here, the orthogonal rule of feature size 4N is used to encode the bases of gene sequence 'A', 'C', 'G', and 'T' as (0,0,0,1), (0,0,1,0), (0,0,0,0), and (1,0,0,0) respectively. This is depicted in Fig. 4 below. Both normal HER2 and sample HER2 gene sequences are encoded. When comparing these two sequences, it is observed that the sample sequence is not equal to the normal sequence, which implies that there is a mutation in the sample gene, which in-turn means that the patient carries a cancer gene. Also, the comparison of the encoded sequence gives the exact position of the mutation in the sample gene. However, there can be silent mutations that need to be avoided. This can be done by analysing protein sequences, which are made up of twenty amino acids, making them complex and more accurate to find point mutations.

The mutation percentage is computed and measured in terms of parts-per-million (ppm). In this experiment, the mutation percentage obtained by analysing the genetic code was 147.903 ppm. The position of the mutation in the sequence is found in 5, 56, 71, 425, 1372, 2042 positions. The mutation percentage obtained for analysing protein sequences is 221.877 ppm and the position of point mutations is found in 23, 457, and 680. Table 2 below shows the comparison of point mutations in the approach [10] by Yi et al. and the approach proposed in our paper. This is depicted in Table 2 below.

5 Conclusion

Breast cancer has become a life-threatening disease for women all over the world, especially those between 40 and 60 years of age. About 20%-30% of the breast cancer cases have a mutation in the HER2 gene. The purpose of this project is to detect and identify the mutation in the HER2 sequence in order to detect the presence of breast cancer. In this paper, gene sequencing and protein sequencing techniques are employed to identify changes in the HER2 gene and protein. In the proposed approach, the One-Hot encoding method is used to detect the mutation in the HER2 gene and HER2 protein sequence. This method can be applied to the BRCA1 and BRCA2 genes. Detecting mutations in these genes can be beneficial for giving early and targeted treatment and also avoiding the re-occurrence of breast cancer. This can also be useful to avoid the spread of cancer cells to other organs in the body.

References

1. Kothari, C., Osseni, M. A., Agbo, L., Ouellette, G., Déraspe, M., Laviolette, F., ... & Durocher, F. (2020). Machine learning analysis identifies genes differentiating triple negative breast cancers. *Scientific reports*, *10*(1), 1-15..
2. Sun, Y., Zhu, S., Ma, K., Liu, W., Yue, Y., Hu, G., ... & Chen, W. (2019). Identification of 12 cancer types through genome deep learning. *Scientific reports*, *9*(1), 1-9.
3. Marchiò, C., Annaratone, L., Marques, A., Casorzo, L., Berrino, E., & Sapino, A. (2021, July). Evolving concepts in HER2 evaluation in breast cancer: Heterogeneity, HER2-low carcinomas and beyond. In *Seminars in cancer biology* (Vol. 72, pp. 123–135). Academic Press..
4. Çetin, Ş. B., Khameneh, F. D., Serteli, E. A., Çayır, S., Hatipoğlu, G., Kamasak, M., .. & Özsoy, G. (2018, May). Automated cell segmentation and spot detection in fluorescence in situ hybridization staining to assess HER2 status in breast cancer. In *2018 26th Signal Processing and Communications Applications Conference (SIU)* (pp. 1–4). IEEE..
5. Saha, I., Rakshit, S., Wlasnowolski, M., & Plewczynski, D. (2019, October). Identification of Epigenetic Biomarkers With the Use of Gene Expression and DNA Methylation for Breast Cancer Subtypes. In *Tencon 2019–2019 Ieee Region 10 Conference (Tencon)* (pp. 417–422). IEEE.
6. Ahn, S., Woo, J. W., Lee, K., & Park, S. Y. (2020). HER2 status in breast cancer: changes in guidelines and complicating factors for interpretation. *Journal of pathology and translational medicine*, *54*(1), 34-44..
7. Karim, M. R., Wicaksono, G., Costa, I. G., Decker, S., & Beyan, O. (2019). Prognostically relevant subtypes and survival prediction for breast cancer based on multimodal genomics data. *IEEE Access*, *7*, 133850-133864..
8. Agersborg, S., Mixon, C., Nguyen, T., Aithal, S., Sudarsanam, S., Blocker, F., ... & Albitar, M. (2018). Immunohistochemistry and alternative FISH testing in breast cancer with HER2 equivocal amplification. *Breast Cancer Research and Treatment*, *170*(2), 321-328.
9. Saha, M., & Chakraborty, C. (2018). Her2Net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation. *IEEE Transactions on Image Processing*, *27*(5), 2189-2200.
10. Yi, Z., Ma, F., Li, C., Chen, R., Yuan, L., Sun, X., ... & Xu, B. (2017). Landscape of somatic mutations in different subtypes of advanced breast cancer with circulating tumor DNA analysis. *Scientific reports*, *7*(1), 1-8.
11. National Center for Biotechnology Information (NCBI), https://www.ncbi.nlm.nih.gov/.