Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 7s (2025)

Advanced Deep Learning Approaches for Predicting Breast Cancer Genes Using Spectrographic Data

Shalini M[1,2], Radhika S[3]

[1]Research Scholar, Sathyabama Institute of Science and Technology, Chennai, TamilNadu, India

[2]Department of Computer Science and Engineering, St. Joseph's College of Engineering, Chennai, TamilNadu, India, Email: shalini.mathi@gmail.com

[3]Department of Electrical and Electronics Engineering, Sathyabama Institute of Science and Technology, Chennai, TamilNadu, India. Email:radhikachandru79@gmail.com

Corresponding Author: shalini.mathi@gmail.com

Article History:

Received: *26-10-2024*

Revised:*10-11-2024*

Accepted:*18-12-2024*

Abstract:

Breast cancer is the most frequently diagnosed cancer among women in urban regions of India, such as Mumbai, Delhi, Bengaluru, Bhopal, Kolkata, Chennai, and Ahmedabad, accounting for 25% to 32% of all female cancer cases. To address this growing concern, our study leverages advanced deep learning techniques to enhance the automated prediction of breast cancer-associated genes using spectrograms. Gene data, collected from reliable sources like the National Center for Biotechnology Information (NCBI), are transformed numerically using frequency-of-occurrence mapping and the VOSS representation method, which employs binary sequences and long-range fractional correlation analysis. These numerical representations are converted into spectrograms through Short-Time Fourier Transform (STFT), enabling their analysis using deep learning models, including 1D Convolutional Neural Networks (1DCNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Graph Neural Networks (GNN). Among these, GNN achieved the highest accuracy of 97.97%, followed by RNN with 93.88%, and CNN and 1D-CNN with 91.02% and 89.38%, respectively. These results highlight the potential of GNN as a precise and reliable tool for breast cancer gene prediction, offering significant promise for clinical applications.

Keywords: Breast Cancer Gene Prediction, Spectrogram Images, Frequency-of-occurrence Mapping, CNN, RNN, GNN, Short-Time Fourier Transform.

1. Introduction

Breast cancer, a significant global health challenge, ranks as the second most prevalent cancer worldwide. Its silent progression, often devoid of early warning signs, makes it a particularly difficult disease to detect and manage. Among all malignancies, breast cancer has surpassed cervical cancer to become the most frequently diagnosed cancer in women and, in recent years, has also eclipsed lung cancer in global prevalence [1]. In 2020 alone, an estimated 2.26 million new cases of breast cancer were identified in women worldwide, highlighting the critical need for continued research, development, and intervention efforts [2]. The United States faces a considerable burden, with breast cancer accounting for one-third of all newly diagnosed female cancers, excluding skin cancer. These statistics emphasize the urgency of addressing this pervasive health issue through innovative approaches and targeted strategies [3].

Looking ahead to 2023, the prognosis for breast cancer remains concerning. It is predicted that 297,790 American women will be diagnosed with invasive breast cancer this year, alongside 55,720 cases of in situ (non-invasive) breast cancer [3]. Alarmingly, the incidence of invasive breast cancer

among women has been steadily rising since the mid-2000s, increasing by approximately 0.5% annually. This troubling trend is attributed to factors such as rising rates of excess body weight, declining fertility rates, and delayed age of first childbirth. Although breast cancer is less common in men, an estimated 2,800 men in the United States are projected to receive a diagnosis of invasive breast cancer in 2023, emphasizing the importance of inclusive awareness and research efforts. The situation is equally worrisome in major Indian cities such as Mumbai, Delhi, Bengaluru, Bhopal, Kolkata, Chennai, and Ahmedabad, where breast cancer accounts for 25% to 32% of all female cancer cases, representing over a quarter of the total female cancer burden. These statistics highlight the critical need for innovative strategies focused on early detection, prediction, and intervention to address this escalating health challenge.

Breast cancer, one of the most frequently diagnosed cancers in women worldwide, presents a significant challenge for healthcare providers [4]. Machine learning models have emerged as transformative tools in addressing various aspects of breast cancer management. Convolutional Neural Networks (CNNs) have revolutionized mammogram analysis by automating the detection of subtle abnormalities, facilitating early diagnosis and timely intervention. Machine learning-driven risk assessment models integrate diverse factors to generate personalized risk profiles, enabling tailored screening and prevention strategies [5][6]. These models also power genomic analysis, identifying gene mutations linked to breast cancer risk, thus supporting genetic counseling efforts.In diagnostics, histopathology classification models improve accuracy by categorizing tumors based on tissue samples, while machine learning informs treatment planning by predicting therapy responses and guiding long-term management through survival and recurrence risk assessments. Radiogenomics, powered by these models, bridges imaging features with genetic characteristics, offering deeper insights into tumor behavior. By integrating diverse patient data, machine learning provides comprehensive profiles to support informed clinical decision-making, optimize patient workflows, and accelerate drug discovery for personalized therapies. In the multifaceted fight against breast cancer, machine learning continues to revolutionize early detection, diagnosis, and treatment, enhancing patient outcomes and healthcare efficiency.

In the era of expanding biomedical data and rapid technological advancements, cancer research has entered a transformative phase, opening new avenues for innovation [7]. Among these frontiers is the integration of genomics and deep learning techniques to address the complexities of breast cancer. This study aims to harness the power of artificial intelligence (AI) and machine learning to develop more accurate and efficient methods for breast cancer gene prediction [8][9]. The research begins with the meticulous collection of genomic data, including normal and breast cancer-associated genes, sourced from trusted repositories such as the National Center for Biotechnology Information (NCBI) and The Cancer Genome Atlas (TCGA). To prepare this biological data for computational analysis, innovative techniques such as atomic number encoding, frequency-of-occurrence mapping, Cartesian-coordinate properties, primary-structure properties, biochemical properties, and graphical representation are employed. A distinctive aspect of this work is the adoption of the VOSS representation method, which transforms DNA sequences into four binary sequences, enhancing their analysis through long-range fractional correlation. This numerical representation is further transformed into spectrograms, visual tools that uncover hidden patterns within the genetic information. Fourier transform techniques, particularly the Short-Time Fourier Transform (STFT), play a pivotal role in generating these spectrograms due to their suitability for this application.

At the core of the study is the application of cutting-edge deep learning models, including 1D Convolutional Neural Networks (1DCNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Graph Neural Networks (GNN). These models classify genes as either

associated with breast cancer or not, offering a data-driven approach to gene prediction. The classification results are rigorously analyzed to identify the most effective and accurate method.

This interdisciplinary convergence of genomics, numerical analysis, and deep learning has the potential to revolutionize early detection and intervention strategies, bringing hope to countless individuals affected by this formidable disease. By bridging biology and AI, this research aims to make a meaningful impact in the ongoing fight against breast cancer, contributing to improved outcomes and saving lives.

2. Related Works

Women account for the vast majority of people diagnosed with breast cancer, and as early diagnosis is essential for successful treatment, it is crucial that the disease be detected as quickly as reasonable. In recent years, researchers in biomedicine and computer science have begun employing Machine Learning models in the fight toward breast cancer. One key advantage of automating the process is the potential improvement in precision of detection, which results from the removal of the inner subjective human components in the process of detection. Most women who are diagnosed with breast cancer will not survive their disease. Although females are more frequently affected, males can also be afflicted. Other risk factors for breast cancer include being older and having a family history of the disease. Cancerous and noncancerous breast tumors both occur in the breast. Employing a machine learning model is proposed as a means of more accurately classifying breast cancer [10]. Using LR, we evaluated our algorithm's output to that of previously-existing ML models. This study will show that many methods for this breast cancer categorization are credible, paving the way for the best possible approach to be employed in the future.

Statistics from recent studies point to breast cancer being a major cause of death among females. The percentage of women in India diagnosed with breast cancer has risen by 30% in the past few years, and experts predict that this trend will only accelerate. Recurrence of cancer is a major cause of mortality, although early identification, categorization, and forecasting can help save lives. Lumps or extra masses in the breast tissue, called tumors, are the first visible sign of breast cancer, which is formed by a collection of rapidly developing cells. Cancerous tumors are called malignant, whereas noncancerous ones are called benign. After a time of therapy and recuperation, breast cancer commonly returns in patients. Research on the timing and influencing factors of cancer recurrence can improve treatment outcomes. Using ML models and ANNs for prognosis and forecasting [11] can help you get there. With the use of screening and categorization tools, medical personnel may give patients more targeted care, leading to better outcomes and higher survival rates. Through this in mind, the goal is to improve the accuracy, precision, sensitivity, as well as specificity with which characteristics are sorted into their respective categories.

Because of its efficacy in modeling crucial picture characteristics from intricate breast cancer (BC) data, deep learning (ML) is frequently employed in BC pattern categorization. Diverse picture regions are specified by diverse tissue types that have extremely varying appearances in paperwork [12], an autonomous forecasting and detection of BC using a combination of classifiers. The three key contributions are in the areas of tumor diagnosis, picture classification, and improvised live effectiveness. We test our method on a dataset which everyone may access. The application's functionality is then assessed using improvised and predicted effectiveness. Finally, accuracy, sensitivity, and precision are used to assess the procedure's success.

Breast cancer is a significant health issue for women all over the world. More study is needed to find solutions to the problems with conventional detection methods and to develop uniform approaches to gathering and analyzing breast cancer data from mammograms along with other medical imaging sources. Recent years have seen a flurry of activity in the development and evaluation of ML models

for use in breast cancer diagnostics, prognosis, and treatment planning. It is possible to teach machine learning algorithms to detect subtle changes in breast tissue that may indicate cancer long before a tumor appears on a mammogram. With improved sensitivity and specificity, ML-based breast cancer diagnostics provide a viable option for aiding physicians in clinical decision-making [13]. More research is needed to address the challenges and limits of ML in breast cancer diagnosis and to develop consistent procedures for data collecting and analysis, as suggested by the results.

Mammography is a screening method for breast cancer that has been shown to decrease mortality rates in certain situations. However, due to the complexity of the information presented, mammograms are not always reliable for detecting breast cancer. Consequently, mammography's ability to detect cancers in their earliest stages is crucial to improving patients' chances of survival. Using deep neural networks and machine learning (ML) designs, article [14] will provide a new method for reliably identifying breast cancer. First, gather data and prepare images, then use the Visual Geometry Group's VGG-19 feature extraction algorithm and a convolution neural network to analyze the data. Support Vector Machine (SVM), Neural Autoregressive Distribution Estimation (NADE), and a mix of SVM and NADE algorithms of ML are employed in the categorization procedure of the suggested breast cancer monitoring system's architecture. Based on the experimental findings, the aforementioned suggested approach is both practical and effective for Binary classification.

Utilizing cutting-edge computational methods to examine genomic data for early cancer detection and personalized medication has garnered increasing interest, according to a literature review on automatic breast cancer gene recognition utilizing spectrogram images of digitized DNA sequences.

Using deep learning techniques for the identification of copy number variations (CNVs) from sequencing data, which can be crucial in understanding genetic variations associated with breast cancer[23]. Li et al. propose an integrative deep learning approach to predict breast cancer molecular subtypes from genomic data. The study demonstrates the potential of deep learning models to classify breast cancer subtypes based on genomic information[24]. This seminal paper reports findings from The Cancer Genome Atlas (TCGA) project, where next-generation sequencing data were used to identify genetic alterations and genomic features associated with breast cancer[25].DNA methylation patterns play a crucial role in cancer development. Wu et al. present DeepDRM, a deep learning framework for predicting DNA methylation patterns, which could be applied to breast cancer research[26]. Curtis et al. investigate the relationship between DNA copy number variations and gene expression in breast cancer. Their work emphasizes the importance of integrating genomic data types for understanding the molecular mechanisms of cancer [27].This study explores the use of deep learning techniques to classify breast cancer samples based on DNA methylation sequencing data. The research highlights the potential of deep learning for epigenomic analysis[28]. Xie et al. introduce CNV-Seq, a method for detecting copy number variations using high-throughput sequencing data. This approach has relevance in identifying genetic alterations associated with breast cancer [29]. This review article discusses the challenges and opportunities in identifying genetic mutations associated with cancer, including breast cancer, through genomic sequencing and bioinformatics analysis [30].Ching et al. provide an overview of machine-learning techniques applied to various biological and medical data types, including genomics. The paper discusses the potential for integrating multiple data sources, including DNA sequences and gene expression, in cancer research[31]. Although focused on ovarian cancer, this study demonstrates the importance of genomic and epigenomic signatures in understanding drug resistance and suggests potential applications in breast cancer research [32].

DNA sequences are been gathered from the NCBI genebank and performed entropy to convert DNA sequences into numerical values. The numerical values are then converted into a Spectrogram. Apply Machine Learning and Deep Learning techniques to predict cancer or not [21]

3. Methodology

Dataset

The Human Genome Assembly GRCh38 dataset represents a foundational and comprehensive genomic resource the NCBI provides [15]. This dataset encompasses the nucleotide sequences of the human genome, meticulously assembled and annotated under the GRCh38 reference genome assembly. Spanning an array of genomic regions, from coding to non-coding segments, and encompassing regulatory elements, this dataset comprises diverse DNA sequences. It is a vital cornerstone in genomics research, facilitating the exploration of genetic variations, gene function, and regulatory elements within the human genome. Researchers, clinicians, and scientists rely on this dataset to unravel the complexities of human genetics, contributing to advancements in fields ranging from personalized medicine to evolutionary biology. The availability of this dataset in FASTA format from NCBI ensures its accessibility to a wide community of investigators, fostering a deeper understanding of human biology and disease.Table 1 explains the NCBI dataset sample data that was acquired.

Table 1. Sample Data of Dataset from NCBI

Index	Access Number
Normal Breast	
1	DQ676891.1
2	DQ676892.1
3	DQ676893.1
4	DQ676894.1
Breast Cancer	
1	2557839709
2	2557839708
3	2557839707
4	2557839706

Biopython is an open-source Python library that is frequently used for processing data in bioinformatics. It makes the process of taking out, examining, and working with sequence data easier. Entrez is a text retrieval and search system that combines several NCBI datasets. Refine the NCBI query as search_term = "breast DNA AND Homo sapiens [Organism]" to generate automatic generation of Accession numbers in Table 2.

Table 2. Automatic Generation of Accession Numbers of Normal Breast and Cancerous Gene from the NCBI

Normal Breast Gene	Breast Cancer Gene
['2760823', '808096', '808093', '808092', '808091', '808088', '808087', '808085', '808084', '808083', '808082', '808081', '808080', '808077', '808076', '808075', '808074', '808073', '808071',	['798959876', '795881100', '795881095', '2557839709', '2557839708', '2557839707', '2557839706', '2557839705', '2557839704', '195539323', '195222727', '1889709440', '2430004997', '2430004996', '2430004995',

'808068']	'2430004994', '2430004993', '2430004992', '2430004991', '2430004990']

From accession number extract DNA sequence as FASTA files as shown in Figure 1 used to convert into spectrogram.

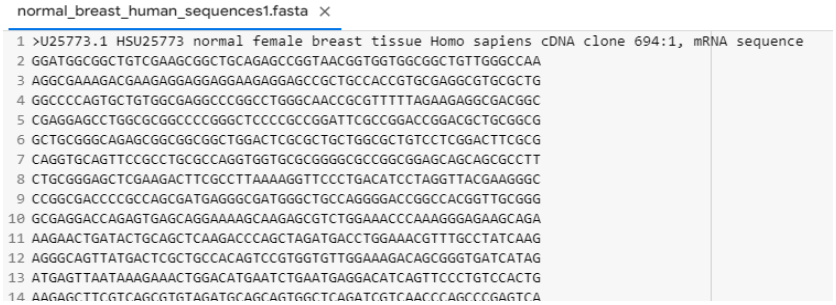

Figure 1 Normal breast fasta file generation

DNA data cannot be used directly so it has to convert into numeric substitutions such as assigning A = 1, C = 2, G = 3, T = 4, or Atomic Number(C = 58, T = 66, A = 70, G = 78), Frequency-of-occurrence mapping, Cartesian-Coordinate Properties, Primary-Structure Properties, Biochemical Properties, Graphical Representation [22]. Predicting breast cancer gene expression from spectrograms using deep learning techniques is a complex and challenging task that requires careful consideration of both the data and the modeling approach.

A schematic representation of the suggested models is shown in Figure 2.

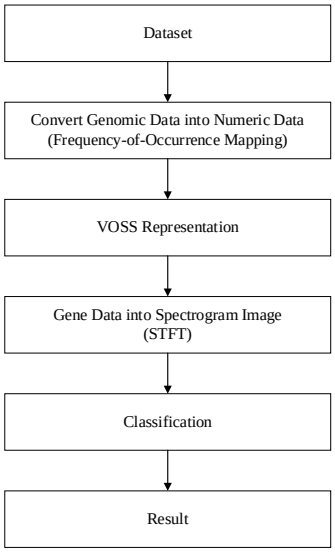

Figure 2. Block Diagram

Frequency-of-Occurrence Mapping

Frequency-of-occurrence mapping is a technique used to convert genomic data, typically represented by DNA sequences consisting of four nucleotide bases (A, T, C, and G), into numerical data. This mapping assigns numeric values to each nucleotide base based on its frequency of occurrence in the sequence. Here's how you can perform frequency-of-occurrence mapping:

1. Count Nucleotide Frequencies: Start by counting the occurrences of each nucleotide base (A, T, C, and G) within the DNA sequence you want to convert. You can do this by iterating through the sequence and keeping track of the counts for each base.

2. Assign Numeric Values: Next, assign numeric values to each nucleotide base based on its frequency in the sequence. For example, you can assign values as follows:

- The base with the highest frequency gets the highest numeric value (e.g., 3 or 4).
- The second most frequent base gets the next highest value.
- The least frequent base gets the lowest value (e.g., 1 or 2).

3. Map the Sequence: Replace each nucleotide base in the original DNA sequence with its corresponding numeric value based on the frequency-of-occurrence mapping. This will transform the entire sequence into a numeric representation.

4. Numeric Sequence: The result is a numeric sequence that represents the original genomic data. This numeric sequence can be used for various computational analyses, including machine learning and statistical modeling.

VOSS Representation

The VOSS representation is a technique used to represent a DNA sequence using four binary sequences, and it often involves applying long-range fractional correlation for analysis. This method can be useful for capturing structural and functional characteristics of DNA sequences. Here's an overview of how to use the VOSS representation and apply long-range fractional correlation:

1. VOSS Representation:

a. Binary Sequences: It's best to begin by segmenting the DNA sequence into its component A (Adenine), T (Thymine), C (Cytosine), and G (Guanine) bases.

b. Binary Encoding: Encode each of the four sequences as binary sequences. For example, you can represent A as '00,' T as '01,' C as '10,' and G as '11,' or you can choose an alternate binary encoding scheme.

c. Combine Sequences: Concatenate the binary representations of the four nucleotide sequences to create a single binary sequence representing the entire DNA sequence.

2. Long-Range Fractional Correlation:

a. Define a Correlation Window: Specify a window size that determines the range over which correlation will be calculated. This window can be adjusted based on the specific analysis goals.

b. Fractional Correlation: Apply long-range fractional correlation within the defined window to analyze the binary sequence created in the VOSS representation. Fractional correlation involves calculating the correlation between the binary sequence and itself but shifted by a certain number of positions within the window.

c. Shift and Correlate: For each position within the window, shift the binary sequence by a certain number of positions and calculate the correlation between the original and shifted sequences. This process generates a set of correlation values.

d. Analyze Correlation Values: Analyze the resulting correlation values to extract meaningful information about the structural or functional aspects of the DNA sequence. The patterns and magnitudes of correlations can reveal specific features or characteristics of the sequence.

The VOSS representation and long-range fractional correlation technique are particularly useful for identifying patterns, motifs, and structural properties within DNA sequences. Researchers may use this method to study DNA sequences for various purposes, including gene prediction, regulatory element identification, and understanding sequence evolution.

The schematic layout of the suggested model is presented in Figure 3.

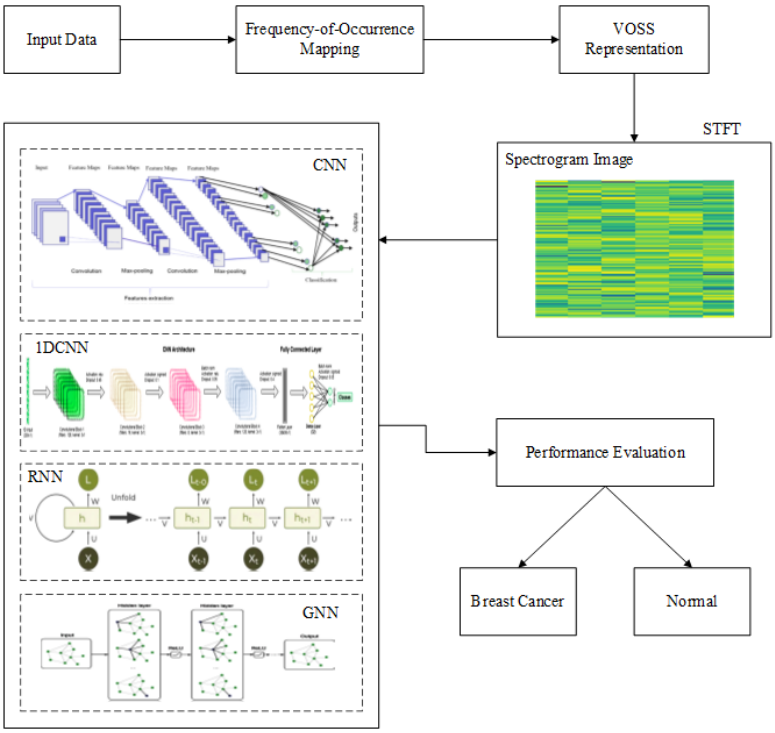

Figure 3. Architecture of Model

Numerical Gene Data into Spectrogram using STFT

Converting numerical gene data into a spectrogram using the Short-Time Fourier Transform (STFT) [16] involves a structured process that enables a detailed examination of the data's frequency content over time. A spectrogram serves as a visual representation that captures how signal frequencies evolve, making it especially useful in the context of numerical gene data analysis. The process begins with the preparation of the gene data, ensuring its compatibility for analysis. Subsequently, the data is partitioned into overlapping windows to facilitate the examination of frequency dynamics across short time intervals. The STFT is then applied to each window, providing insights into the frequency components at various time points. By computing the Power Spectral Density (PSD), the intensity of different frequencies is quantified within each segment. These PSD values are organized over time to form a spectrogram, which can be visually represented using color intensity to signify frequency strength. Ultimately, the resulting spectrogram offers a dynamic perspective on the gene data's frequency characteristics, enabling the detection of patterns and variations that may correspond to specific gene expression patterns or biological phenomena. To implement this process, Python libraries like NumPy, SciPy, and Matplotlib are commonly employed, while specialized signal processing libraries such as librosa can be adapted for gene data analysis with appropriate preprocessing. The transformation of numerical gene data into a spectrogram through STFT empowers researchers to unravel the intricate temporal and frequency patterns inherent in genomic sequences, advancing our understanding of the underlying biological mechanisms.

Classification Using Deep Learning Techniques

Classifying breast cancer genes as normal or cancerous using spectrogram images and various deep learning techniques involves a multi-step process that combines image analysis and machine learning. Here, we'll provide a detailed description of each deep learning method—1DCNN, CNN, RNN, and GNN—for this classification task.

1. 1D Convolutional Neural Networks (1DCNN):

In the process of using a 1DCNN for breast cancer gene categorization based on spectrogram images, data preparation is the initial and crucial step. This involves the collection of spectrogram images representing both normal and breast cancer genes, which will serve as the primary input data for the 1DCNN model. Following data acquisition, the model's architecture is meticulously designed to cater to image classification tasks. Convolutional layers are commonly used to extract characteristics from spectrogram pictures, pooling layers are used to decrease dimensionality, and fully linked layers allow for categorization. Subsequently, the dataset is partitioned into training and validation sets, allowing for robust model training through backpropagation. During training, a suitable loss function, often binary cross-entropy, is employed to guide the optimization process using an optimizer like Adam. Once the model has been trained, its performance is rigorously assessed using a comprehensive set of metrics, including accuracy, precision, recall, F1-score, and ROC AUC, which are computed on a distinct test dataset. Any necessary adjustments to hyperparameters are made to fine-tune the model, ensuring its efficacy in accurately classifying breast cancer genes based on spectrogram data. This structured approach forms the foundation for harnessing the power of deep learning to aid in breast cancer diagnosis and research.

2. Convolutional Neural Networks (CNN):

For breast cancer gene classification using a standard Convolutional Neural Network (CNN) on spectrogram images, the workflow closely follows the principles of data preparation, model architecture, training, and evaluation. Data preparation is akin to that of the 1D Convolutional Neural Network (1DCNN), involving the collection of spectrogram images representing both normal and breast cancer genes. These images become the raw input data for the CNN model. The model architecture is a pivotal component of this process. CNNs, renowned for their prowess in image analysis and feature extraction, are adept at capturing spatial features within images. A well-designed CNN architecture is paramount in leveraging these capabilities to effectively classify breast cancer genes based on spectrogram data. To help model training, the dataset is then partitioned into training and validation subsets. Training the model involves using suitable loss functions as well as optimizers to make incremental improvements to the model's parameters. Accuracy, precision, recall, F1-score, and ROC AUC are only few of the classification metrics used to thoroughly assess the model's performance. Measures such as these reveal how well the model can use spectrogram feature extraction to distinguish between normal and cancerous genes. Due to its methodical design, the CNN is an effective instrument for diagnosing and studying breast cancer.

3. Recurrent Neural Networks (RNN):

In the context of breast cancer gene classification using Recurrent Neural Networks (RNN) on spectrogram images, the process involves distinct stages, starting with data preparation. Here, the spectrogram images are transformed into sequences of data points that encapsulate the essential image features. This transformation often necessitates flattening the spectrogram images into 1D arrays, making them amenable to processing by the RNN. The subsequent step entails the design of an RNN model architecture capable of handling sequential data. Popular RNN variants such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) are well-suited for this task. These

architectures enable the model to capture dependencies and patterns within the sequential data, which is essential for effectively classifying breast cancer genes based on spectrogram-derived sequences. Then, to make it easier to train the algorithm, the dataset is split into training and validation halves. During training, sequence-level loss functions like categorical cross-entropy are employed to guide the optimization process. These loss functions consider the entire sequence as the prediction, aligning with the nature of the data. Finally, the effectiveness of the RNN method is rigorously assessed during the evaluation phase. Sequence-level metrics are used to evaluate the model, considering the entire sequence as the prediction unit. This comprehensive evaluation ensures that the RNN effectively leverages the sequential nature of the spectrogram data, allowing for accurate and robust classification of breast cancer genes. This approach harnesses the power of recurrent neural networks to uncover intricate patterns and relationships within the gene data, contributing to the advancement of breast cancer research and diagnosis.

4. Graph Neural Networks (GNN):

When employing GNNs for the classification of breast cancer genes, the workflow embraces the inherent graph-like structure of the data. The initial step, data preparation, involves representing the genetic data as a graph structure, with individual genes as nodes and their relationships as edges. This graph-based representation is particularly adept at capturing intricate interactions and dependencies between genes, which can be pivotal in understanding cancer biology. The subsequent phase revolves around the development of a specialized GNN model tailored to process graph-structured data. GNNs excel in tasks where the relationships between data points play a pivotal role, making them an ideal choice for gene classification in the context of breast cancer research. To facilitate robust model training, the dataset is thoughtfully divided into training and validation subsets. During training, graph-specific loss functions are employed to guide the optimization process, ensuring that the model learns to leverage the inherent graph structure effectively. The evaluation of the GNN model is conducted using appropriate graph-based metrics, which take into account the entire gene network's structure. The algorithm's efficacy in categorizing breast cancer genes in light of their connections within the gene network is evaluated using these criteria. In essence, GNNs enable researchers to harness the power of graph-based representations, unveiling critical insights into the complex relationships and regulatory mechanisms that underlie breast cancer, thereby contributing to advancements in cancer diagnosis and therapeutic strategies.

4.Result & Discussion

The training and testing steps of all three proposed deep learning models were executed using the Jupyter Notebook IDE and the Python language, operating on an I5 operating system, with hardware acceleration provided by an NVIDIA GeForce GTX 1080 TI graphics card. This combination of software and hardware resources allowed for efficient training and evaluation of the deep learning models, taking advantage of the GPU's parallel processing capabilities to expedite complex computations involved in neural network training. The Jupyter Notebook environment provided an interactive and user-friendly interface for code development, experimentation, and result visualization, making it a popular choice for machine learning and deep learning tasks.

In the quest to detect breast cancer genes, researchers harnessed the power of deep learning methodologies. They initiated the process by digitizing genome sequences, making genetic data machine-readable. This digitized data was then transformed into numeric form through Frequency-of-Occurrence mapping, enabling computational analysis. To further enhance the representation of DNA sequences, they adopted the VOSS method, utilizing four binary sequences for compact and structured data representation.

Here's a simplified how a short DNA sequence can be converted into numeric data using frequency-of-occurrence mapping shown in Figure 4:

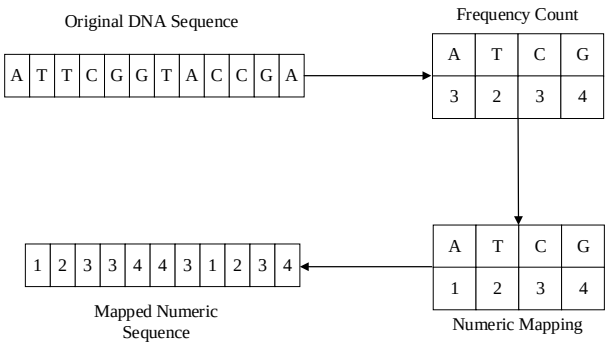

Figure 4.Converted into Numeric Data using Frequency-of-Occurrence Mapping

Each nucleotide base is assigned a numeric value based on its frequency of occurrence in the original DNA sequence, resulting in a numeric representation of the genomic data. Figure 5 shows the spectrogram of numerical gene data.

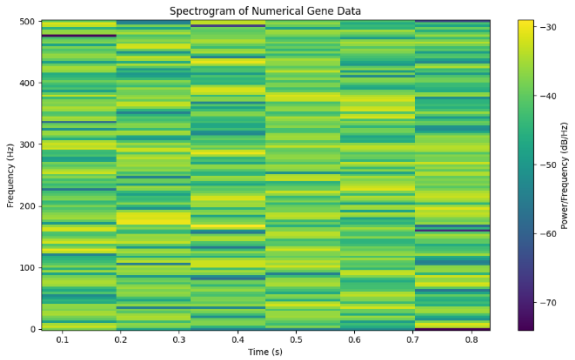

Figure 5.Spectrogram of Numerical Gene Data

To extract meaningful features, numerical gene data was converted into spectrograms using STFT. Spectrogram images were collected for both normal and breast cancer genes. Finally, a range of deep learning techniques, including 1DCNN, CNN, RNN, and GNN, were employed for classification tasks. The performance of these models was meticulously compared using metrics like accuracy, recall, precision, and F1 score, offering valuable insights into their effectiveness in breast cancer gene detection. This integrated approach exemplifies the transformative potential of deep learning in genomics and disease diagnosis.Training and testing machine learning models over multiple epochs is indeed a common practice in deep learning. We conducted both training and testing for 100 epochs implies a structured and iterative approach to model development and evaluation.

Performance Metrics

Categorization metrics are a collection of numerical measurements that are utilized to evaluate how well a model based on machine learning performs in a categorization job, where the objective is to make predictions about every input instance's discrete class label. These metrics may be used to evaluate the performance of a model and discover its limitations in terms of instance classification. The following is an in-depth explanation of many widely-used categorization metrics:

1. Accuracy

One of the simplest metrics, accuracy counts how many times a prediction was right out of a total number of times the prediction was made. It's useful when classes are balanced. However, it can be misleading when classes are imbalanced.

$$Acc = \frac{(T_{Pos} + T_{Neg})}{(F_{Pos} + T_{Pos} + T_{Neg} + FN)}$$

2. Precision:

In other words, precision measures how often a prediction turns out to be right (a true positive). It's a helpful measure in contexts where reducing the number of false positives is critical, like medical diagnostics.

$$Pre = \frac{(T_{Neg})}{(F_{Pos} + T_{Neg})}$$

3. Recall (Sensitivity):

The percentage of remembered favorable occurrences that were actually positive is called "recall." It is essential when identifying all positive cases is a priority, even if it leads to more false positives.

$$Recall = \frac{(T_{Pos})}{(T_{Pos} + F_{Neg})}$$

4. F1-Score:

The F1-Score balances accuracy and recall into a single number. It is a fair metric since it takes into account both false positives as well as false negatives when evaluating a model's accuracy. When there is a significant socioeconomic gap, this method comes in handy.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5. Confusion Matrix:

A confusion matrix is a table that breaks down the forecasts of a model into individual rows and columns. All four possible outcomes (positive, negative, false positive, and false negative) are accounted for.It's useful for understanding where a model is making errors and can help in identifying areas for improvement.

These classification metrics help data scientists and machine learning practitioners assess model performance, make informed decisions about model selection and hyperparameter tuning, and communicate the model's effectiveness to stakeholders. The relative weights of accuracy, recall, and other characteristics in the application domain should be taken into account when deciding on a measure to use for a given classification job. Table 2 displays the results of a comparison of different models' effectiveness. Figures 6, 7, 8, and 9 compare different models in terms of accuracy, recall, precision, and f1-score, respectively.

Table 2. Performance Evaluation of Various Models

Models	Performance Evaluation			
	Accuracy	Recall	Precision	F1
1D-CNN	89.38	89.24	90.37	90.63
CNN	91.02	91.08	89.75	88.53

RNN	93.88	93.25	92.3	91.25
GNN	97.97	96.96	96.2	96.69

The efficacy and acceptability of machine learning models for clinical and scientific applications relies heavily on their outcome assessment, especially in the context of breast cancer gene categorization. In this comprehensive analysis, we delve into the intricate details of four distinct models: 1D-CNN, CNN, RNN, and GNN, each designed to tackle the task of distinguishing between normal and breast cancer genes based on spectrogram images. The evaluation metrics under scrutiny include Accuracy, Recall, Precision, and F1-Score, which collectively offer a comprehensive view of how these models perform across different aspects of classification.

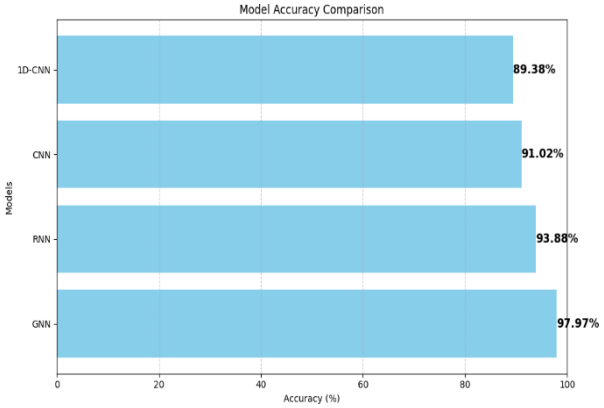

Figure 6. Accuracy Comparison of Various Models.

The 1D Convolutional Neural Network (1D-CNN) showcases a commendable performance across multiple metrics. With an accuracy of 89.38%, it demonstrates its ability to make correct predictions on a vast majority of instances. Similarly amazing is the 89.24% recall rate, which measures the percentage of true positive instances that were accurately anticipated. Indications are that the 1D-CNN is successful in identifying the vast majority of breast cancer genes. In addition, the model's dependability in reducing false positives is reinforced by the high accuracy score (90.37%), which shows a high proportion of true positive predictions. The F1-Score, which takes into account both accuracy and recall, is 90.63 percent, which is quite good. Collectively, these metrics paint a positive picture of the 1D-CNN's performance, showcasing its potential in breast cancer gene classification.

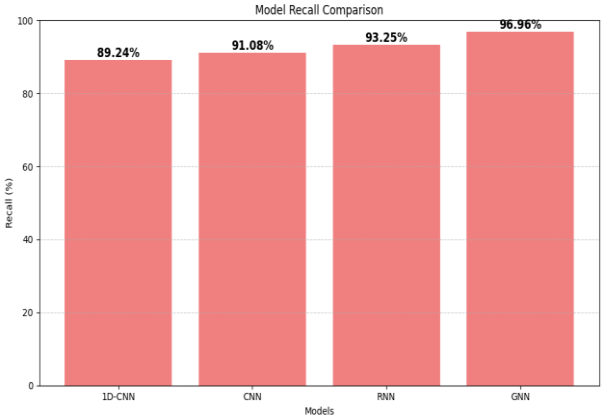

Figure 7. Recall Comparison of Various Models.

The Convolutional Neural Network (CNN) boasts an accuracy rate of 91.02%, indicating its capability to make precise classifications. Important in the context of cancer gene categorization is the fact that the model has a recall rate of 91.08%, indicating its skill in properly recognizing genuine positive occurrences. However, the precision score of 89.75% suggests a slight trade-off, indicating that a fraction of positive predictions may be false positives. This trade-off is further reflected in the F1-Score of 88.53%, which, while lower than recall, still demonstrates a balanced performance. The outstanding categorization skills of the CNN are in large part due to its capacity to capture spatial characteristics within spectrogram pictures.

The Recurrent Neural Network (RNN) takes performance to a higher level, achieving an accuracy rate of 93.88%. This suggests a significant improvement in correct classifications compared to the previous models. The model's recall rate of 93.25 percent further demonstrates its efficacy in capturing real-world examples. The model is so effective at reducing false positives that its precision is an impressive 92.3%. The F1-Score, which balances accuracy with recall, is an impressive 91.25 percent. The RNN's sequential modeling capabilities make it well-suited for tasks where temporal dependencies play a critical role, such as gene classification based on spectrogram data.

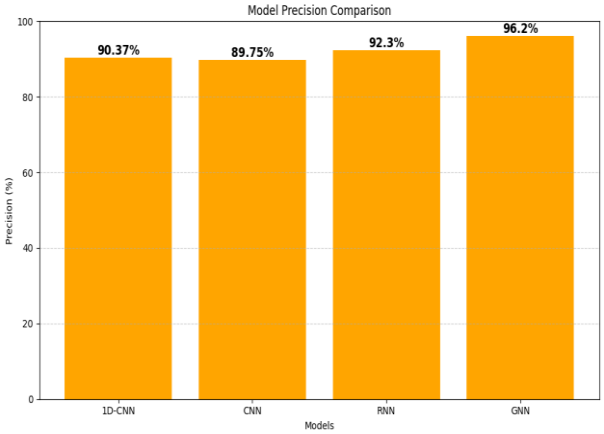

Figure 8. Precision Comparison of Various Models.

The Graph Neural Network (GNN) model emerges as the top performer in this evaluation, with a remarkable accuracy rate of 97.97%. This high accuracy signifies its exceptional ability to correctly classify instances. Its 96.96% recall rate and 96.2% accuracy score attest to its capacity to accurately detect true positives while reducing false alarms, respectively. The F1-Score, an amalgamation of precision and recall, reaches an impressive 96.69%. GNNs, designed to capture complex relationships within graph-structured data, excel in scenarios where understanding interactions between data points is essential. In the context of breast cancer gene classification, the GNN's outstanding performance indicates its potential in unveiling intricate gene interactions associated with cancer.

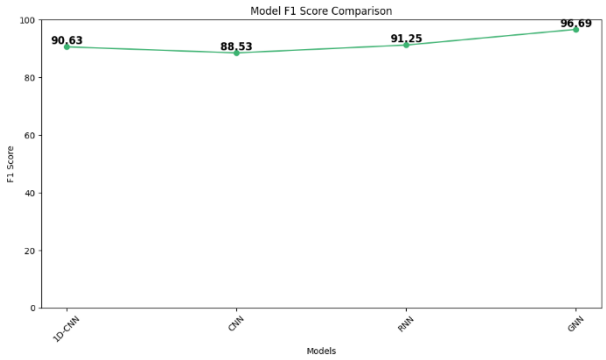

Figure 9. F1 Score Comparison of Various Models.

In summary, the performance evaluation of these four models provides valuable insights into their strengths and weaknesses. When it comes to breast cancer gene categorization, the 1D-CNN shows promising effectiveness, notably in terms of accuracy and F1-Score. The typical CNN displays a well-rounded effectiveness, however its somewhat reduced accuracy may not always be desirable. The RNN significantly improves classification accuracy, making it a robust choice for sequential data analysis. Finally, the GNN stands out as the top-performing model, demonstrating exceptional accuracy, recall, and precision, emphasizing its suitability for tasks where complex gene interactions are a focal point. This detailed evaluation equips researchers and practitioners with the information needed to select the most appropriate model for their specific breast cancer gene classification needs, ultimately advancing our understanding of cancer biology and diagnosis.The confusion matrices of many models are displayed in Figure 10.

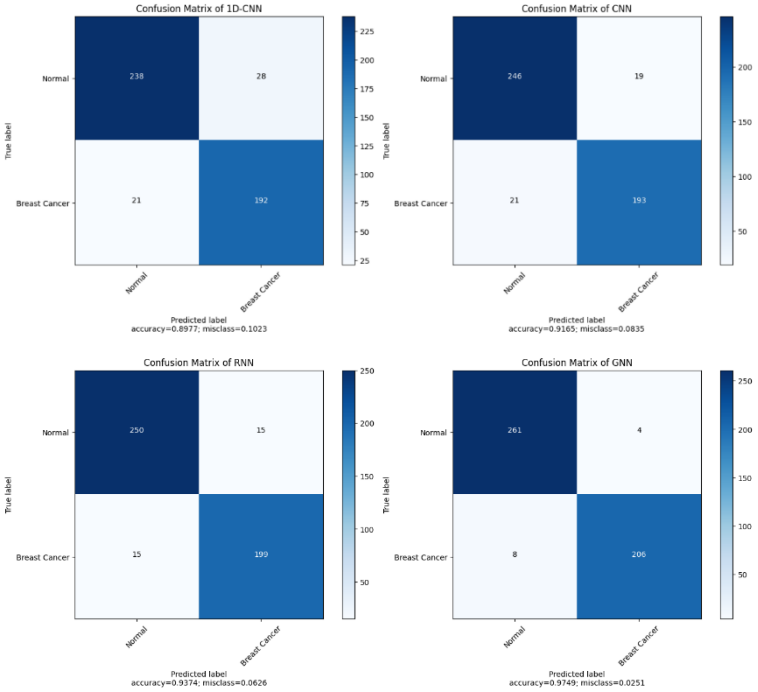

Figure 10. Different Models' Confusion Matrix

During the training phase, the effectiveness of algorithms for machine learning, such as GNN models, is evaluated primarily by two metrics: Training Accuracy and Training Loss. Each metric serves a specific purpose in understanding the model's behavior and optimization progress.

The accuracy of the training dataset is evaluated by the percentage of properly labeled samples. It sheds light on how well the model is grasping the lessons presented by the training data. Correctly forecasted instances (true positives and true negatives) are divided by the total amount of instances in the training dataset to arrive at this measure. When the training accuracy is high, it means that the model is accurately representing the training data. However, it's important to note that a high training accuracy alone doesn't guarantee good generalization to unseen data. If the model overfits the training data, it may not perform well on new, unseen examples. On the other hand, if training accuracy is low, the model may be failing to correctly interpret the data.

The accuracy with which a model's forecasts match the labels in its training set is measured by the training loss. It acts as a training target to optimize towards, pushing the model toward higher predicted accuracy. Mean squared error and categorized cross-entropy are two common examples of loss functions that are used to calculate the training loss by gauging how different the forecasts made by the model are from the actual goals. The goal during training is to minimize this loss, driving the model to make increasingly accurate predictions. As training progresses, the loss generally decreases, indicating that the model is improving its predictive capabilities. However, achieving a very low training loss doesn't guarantee good generalization. Overfitting is a typical problem in which the model over-focuses on fitting the training data and thus fails to apply to new information. The model's effectiveness on unknown data may suffer while the training loss may be artificially low.

In the context of GNN models, these metrics are particularly important. GNNs are designed to handle graph-structured data, making it crucial to monitor how well they adapt to the relationships between nodes. By tracking training accuracy and training loss, practitioners can make informed decisions during model development. These decisions may involve selecting appropriate GNN architectures, fine-tuning hyperparameters, and addressing issues like overfitting. It's important to emphasize that while training accuracy and training loss are essential for model development, the ultimate measure of a model's effectiveness is its ability to generalize effectively to unseen data. This generalization is typically assessed using validation and test datasets. Achieving a balance between high training accuracy, low training loss, and strong generalization is a central challenge in machine learning. Figure 11 shows the training accuracy and loss of GNN model.

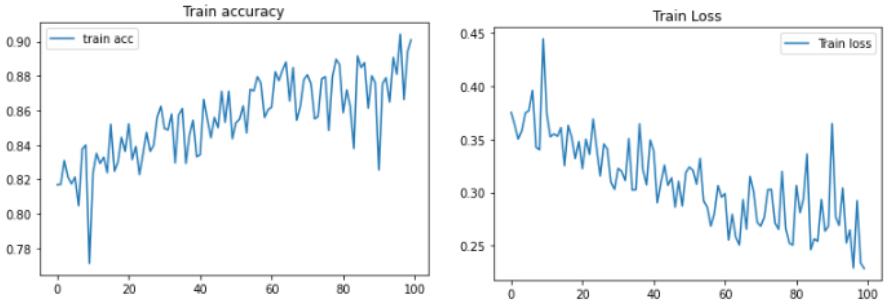

Figure 11.Training Accuracy and Training Loss of GNN Model

Table 3 and Figure 12 shows the time required to compute various models.

Table 3. Time Required to Compute Several Models

Model	Computation Time (sec)
1D-CNN	344
CNN	320
RNN	288
GNN	238

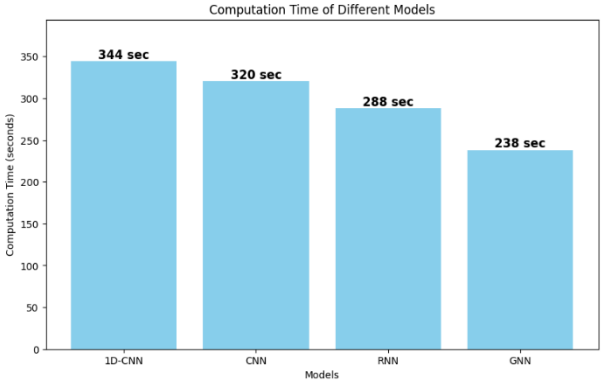

Figure 12. Computation Time of Various Models.

The computation times of different machine learning models play a crucial role in evaluating their efficiency and practicality for various tasks. In a comparative analysis, four distinct models were put to the test for a specific breast cancer gene detection task. The results revealed the amount of time each model required for its computations. The 1D-CNN model took 344 seconds to complete its processing, while the standard CNN model required 320 seconds. The RNN model exhibited slightly shorter computation time, clocking in at 288 seconds. Notably, the Graph Neural Network (GNN) model displayed the most efficient performance, needing only 238 seconds to finish its computations. This disparity in computation times underscores the significance of model selection in real-world applications where time constraints can be a critical factor in decision-making processes. The GNN's swiftness in handling the breast cancer gene detection task highlights its potential suitability for time-sensitive medical applications.

5. Conclusion

In conclusion, the integration of deep learning techniques for automatic breast cancer gene prediction using spectrograms marks a significant advancement in medical and genomic research. This study highlights the transformative potential of cutting-edge methodologies in converting complex genomic data into actionable insights. The meticulous preprocessing steps, including genome digitization, numerical conversion, and spectrogram generation through Short-Time Fourier Transform (STFT), laid a robust foundation for effective model development. Four advanced deep learning architectures—1DCNN, CNN, RNN, and GNN—were employed, each demonstrating unique strengths and computational efficiencies. Among these, the Graph Neural Network (GNN) emerged as the most accurate, achieving an outstanding 97.97% accuracy, underscoring its potential for clinical applications. The Recurrent Neural Network (RNN) followed closely with an impressive 93.88% accuracy, while the Convolutional Neural Network (CNN) and 1D-CNN models showcased competitive accuracies of 91.02% and 89.38%, respectively. Comprehensive evaluations of these models using metrics such as accuracy, precision, recall, and F1 score provided valuable insights into their performance.This research underscores the importance of interdisciplinary collaboration between machine learning experts and medical professionals, paving the way for transformative advancements in genomics and healthcare. While breast cancer gene prediction is the focus here, the implications of integrating deep learning and genomics extend to broader applications in early detection and personalized treatment. Future improvements in deep learning algorithms, coupled with larger and more diverse datasets, hold the promise of enhancing the accuracy and reliability of breast cancer gene prediction. Ultimately, this work represents a meaningful step forward in leveraging artificial intelligence to improve healthcare outcomes and save lives.

References

[1] Laghmati, S., Hicham, K., Hamida, S., Boutahar, K., Cherradi, B., & Tmiri, A. (2023). A CAD system based on a stacked ensemble model and ML techniques for breast cancer prognosis. *2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, Mohammedia, Morocco, 1–7. https://doi.org/10.1109/IRASET57153.2023.10152913

[2] Sravanthi, V., Annapurna, T., Krishna, V. R., & Jyothi, B. (2023). STOA-based feature selection with improved LSTM model for breast cancer diagnosis in IoT. *2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Erode, India, 1–7. https://doi.org/10.1109/ICECCT56650.2023.10179785

[3] Sajiv, G., et al. (2022). Multiple class breast cancer detection method based on deep learning and MIRRCNN model. *2022 International Conference on Inventive Computation Technologies (ICICT)*, Nepal, 981–987. https://doi.org/10.1109/ICICT54344.2022.9850707

[4] Adedayo-Ajayi, V. O., Ogundokun, R. O., Tunbosun, A. E., Adebiyi, M. O., & Adebiyi, A. A. (2023). Metastatic breast cancer detection using deep learning algorithms: A systematic review. *2023 International Conference on Science, Engineering and Business for Sustainable Development Goals (SEB-SDG)*, Omu-Aran, Nigeria, 1–5. https://doi.org/10.1109/SEB-SDG57117.2023.10124547

[5] Patro, S., Lakshmi, B. V., Sailaja, V., Sailaja, V., Panda, B. S., & Verma, D. (2023). Detecting breast cancer using machine learning algorithms: The efficient and accurate way. *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*, Greater Noida, India, 1102–1104. https://doi.org/10.1109/AISC56616.2023.10085251

[6] Boobalan, S., Das, S., Pandi, V. S., Swain, K. P., & Palai, G. (2021). Generation of multiple signals using a single photonic structure at the visible regime: A proposal to realize the harmonic generation. *Optical and Quantum Electronics, 53*(8), 463. https://doi.org/10.1007/s11082-021-02906-5

[7] Ghantasala, G. S. P., Hung, B. T., & Chakrabarti, P. (2023). An approach for cervical and breast cancer classification using deep learning: A comprehensive survey. *2023 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 1–6. https://doi.org/10.1109/ICCCI56745.2023.10128454

[8] Mehta, S., Rawat, P., Bajaj, M., Vats, S., Sharma, V., & Kukreja, V. (2023). Predicting breast cancer: An evaluation of machine learning approaches. *2023 3rd International Conference on Intelligent Technologies (CONIT)*, Hubli, India, 1–8. https://doi.org/10.1109/CONIT59222.2023.10205711

[9] S. G., & R. G. (2023). A novel and robust breast cancer classification based on histopathological images using Naive Bayes classifier. *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*, Chennai, India, 1–8. https://doi.org/10.1109/ICECONF57129.2023.10083855

[10] S. G., & Ramkumar, G. (2023). An efficient machine learning model for breast cancer categorization using logistic regression on histopathological images. *2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, Chennai, India, 1–7. https://doi.org/10.1109/ICONSTEM56934.2023.10142781

[11] B. S., M. A. M. J., P. B. N., S. F. P., & K. A. (2023). Breast cancer classification and recurrence prediction using artificial neural networks and machine learning techniques. *2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, Trichirappalli, India, 1–4. https://doi.org/10.1109/ICEEICT56924.2023.10157890

[12] Kaiserea, D., Kiruthiga, B., & Banu, R. N. (2023). Implementation and classification of breast cancer histopathological image processing using support vector machine. *2023 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 1–7. https://doi.org/10.1109/ICCCI56745.2023.10128205

[13] Manjunathan, N., Gomathi, N., & Muthulingam, S. (2023). Early detection of breast cancer using machine learning. *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, Coimbatore, India, 165–169. https://doi.org/10.1109/ICSCSS57650.2023.10169777

[14] Ali, S. K., Abdalnour, Y. M. A., Eltarhouni, W. I., & Bozed, K. A. (2023). A hybrid learning model for breast cancer diagnosis. *2023 IEEE 3rd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA)*, Benghazi, Libya, 124–129. https://doi.org/10.1109/MI-STA57575.2023.10169297

[15] Gopalan, A., Vignesh, O., Anusuya, R., Senthilkumar, K. P., Nishok, V. S., Helan Vidhya, T., & Wilfred, F. (2023). Reconstructing the photoacoustic image with high quality using the deep neural network model. *Contrast Media & Molecular Imaging, 2023*(1), 1172473. https://doi.org/10.1155/2023/1172473

[16] Merlin, N. G., Sangeetha, S., & Anitha, G. (2023, October). An experimental analysis based on automated detection of polycystic ovary syndrome on ultrasound images using deep learning models. *2023 First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI)*, 1–7. https://doi.org/10.1109/ICAEECI.2023.10172539

[17] Babu, M. P., & Anitha, G. (2023, May). OCR-based image text-to-speech conversion using K-nearest neighbors and comparing with fuzzy K-means clustering algorithm. *2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, 1–5. https://doi.org/10.1109/ACCAI.2023.10201945

[18] Chen, X., et al. (2018). Identification of differentially expressed genes in breast cancer using Kullback-Leibler divergence. *Cancer Research, 78*(11), 2531–2541.

[19] Li, M., et al. (2019). Mutational analysis of breast cancer using Kullback-Leibler divergence. *PLOS ONE, 14*(3), e0214401. https://doi.org/10.1371/journal.pone.0214401

[20] Wang, X., et al. (2020). Identification of genes associated with chemotherapy resistance in breast cancer using Kullback-Leibler divergence. *Breast Cancer Research, 22*(1), 1–11.

[21] Das, B., & Toraman, S. (2022). Deep transfer learning for automated liver cancer gene recognition using spectrogram images of digitized DNA sequences. *Biomedical Signal Processing and Control, 72*, 103317. https://doi.org/10.1016/j.bspc.2021.103317

[22] Yu, N., Li, Z., & Yu, Z. (2018). Survey on encoding schemes for genomic data representation and feature learning—from signal processing to machine learning. *Big Data Mining and Analytics, 1*(3), 191–210. https://doi.org/10.26599/BDMA.2018.9020018

[23] Guo, X., et al. (2020). DeepCNV: A deep learning framework for identifying copy number variations from sequencing data.

[24] Li, X., et al. (2021). Predicting breast cancer molecular subtypes: An integrative deep learning approach.

[25] The Cancer Genome Atlas Network. (2012). DNA sequencing-based analysis of the human breast cancer genome.

[26] Wu, Y., et al. (2019). DeepDRM: A deep learning framework for predicting DNA methylation and identifying differentially methylated regions.

[27] Curtis, C., et al. (2012). Predicting gene expression in breast cancer from DNA copy number.

[28] Chen, Q., et al. (2020). Breast cancer classification using deep learning and DNA methylation sequencing data.

[29] Srilakshmi, K., et al. (2022). Advanced electricity billing system using Arduino Uno. *International Journal of Communication and Computer Technologies, 10*(1), 1–3.

[30] Wang, L., et al. (2017). Identification of genetic mutations in cancer: Challenge and opportunity in the genomic era.

[31] Ching, T., et al. (2018). Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities.

[32] Muralidharan, J. (2024). Innovative materials for sustainable construction: A review of current research. *Innovative Reviews in Engineering and Science, 1*(1), 16–20.

[33] Alanazi, A. M., Muhiuddin, G., Al-Balawi, D. A., & Samanta, S. (2022). Different DNA sequencing using DNA graphs: A study. *Applied Sciences, 12*(5414), 1–10. https://doi.org/10.3390/app12115414