Developing an Automated Computational Genomics Pipeline for Breast Cancer Detection Using NGS

Dilip Sisodia[a,b*], Virendra Kumar Sharma[b*], Rohit Joshi[a], Hemant Arya[c], and Tarun Kumar Bhatt[c]

[a]Engineering College, Ajmer, Rajasthan-305025, India

[b]Bhagwant University, Ajmer, Rajasthan-305004, India

[c]Department of Biotechnology, Central University of Rajasthan, Bandarsindri, Ajmer-305817, India

*Corresponding authors: sisodia.dilip@ecajmer.ac.in

Abstract

Breast cancer is a malignant tumor that develops in the breast tissue due to uncontrolled cell growth. It is a significant cause of death worldwide in cancer patients and detecting it early can greatly improve survival rates. This research proposes an automated bioinformatics pipeline that leverages Next-Generation Sequencing (NGS) data to detect genetic mutations associated with breast cancer by comparing them with healthy human genomes. The focus is on the BRCA1 and BRCA2 genes to evaluate the pipeline's effectiveness. The research follows a structured bioinformatics workflow, starting with data acquisition from the NCBI SRA database. Preprocessing steps using SRA Toolkit, FastQC, and Trimmomatic ensure the removal of low-quality reads. Sequence alignment is performed with Bowtie2, while SAMtools and BCFtools are used for variant calling to detect genetic mutations. Finally, SnpEff annotates these mutations and evaluates their biological significance.

Genetic features are extracted and structured into a dataset to develop a predictive model, and they are analyzed using Python and machine learning techniques such as Random Forest and Logistic Regression. The classification models are assessed based on accuracy, demonstrating their effectiveness in predicting breast cancer. The results reveal a strong correlation between genetic mutations and breast cancer risk, offering valuable insights for genomic-based diagnostics.

This study introduces a scalable and reproducible approach to genomic data analysis for breast cancer prediction. The proposed method can support personalized medicine by enabling early detection and risk assessment for high-risk individuals.

Keywords:

Breast Cancer; Computational Biology; Next-Generation Sequencing (NGS); Machine Learning; Genomic Mutations.

1. Introduction

Breast cancer is one of the most common and life-threatening cancers worldwide. According to the World Health Organization (2020), approximately 2.3 million new cases and 685,000 deaths were recorded in 2020 [1]. Despite improved mechanisms of diagnosis and treatment, delayed diagnosis of breast cancer remains a source of worry, resulting in unfavorable prognosis and death.

Genetic mutations significantly contribute to breast cancer susceptibility, with alterations in key tumor suppressor genes like BRCA1 and BRCA2 [2] being strongly linked to increased risk. These genes play a crucial role in repairing DNA. When they become dysfunctional, the risk of developing breast and ovarian cancer at an early age increases, often leading to a more aggressive form of the disease.

Research has indicated that BRCA1 and BRCA2 mutations contribute to 5-10% of all breast cancer [3]. Individuals with deleterious alterations in these genes have a 45-65% lifetime risk for breast cancer, while the general risk for the rest of the population is 12%. These mutations are also associated with some tumor characteristics, such as high-grade tumors, triple-negative breast cancer (TNBC) subtypes, and an increased likelihood of having cancer in both breasts. Several studies have been centered on identifying the mutations through sophisticated sequencing techniques and computer software to enhance early diagnosis and risk assessment. Broad use of conventional statistical models, i.e., the Gail Model, BOADICEA [4], and Tyrer-Cuzick Model, has been used to predict the risk of breast cancer. These models make use of family history, hormonal, and lifestyle variables in estimating the individual's risk of developing breast cancer. These models are not entirely reliable when it comes to predicting the risks, particularly in ethnically diverse populations, and do not possess all the genetic information. Current methods of genetic testing, such as polymerase chain reaction (PCR) tests [5] and microarray analysis, are not good at identifying new or rare variants with high specificity.

Recent advancements in Next-Generation Sequencing (NGS) and machine learning (ML) have enhanced genetic marker analysis and cancer risk prediction [6][7]. NGS can sweep whole genomes entirely for single nucleotide polymorphisms (SNPs) [8], insertions/deletions, and structural rearrangements with high precision. Random Forest and Logistic Regression algorithms can use extensive genomic data to distinguish patterns of cancer, providing improved diagnosis accuracy over earlier methods. Integrating NGS data and ML classification facilitates the identification of risks in an individual, facilitating early detection and targeted treatment in high-risk individuals.

This study proposes an automatic bioinformatics pipeline for identifying the genetic mutations causing breast cancer, emphasizing BRCA1 and BRCA2. The procedure starts with retrieving raw sequence data from the NCBI SRA database, with quality filtering and cleaning performed utilizing SRA Toolkit [9], FastQC [10], and Trimmomatic [11]. Bowtie2 [12] performs sequence alignment, and SAMtools and BCFtools [13] are utilized to identify mutations. These mutations are also annotated utilizing SnpEff [14] to determine whether they have any biological significance. The identified genetic features are then checked utilizing machine learning models to determine their effectiveness in diagnosing breast cancer.

This approach is a bioinformatics and machine learning combination [15] to improve breast cancer's accuracy and risk prediction. The approach is scalable, reproducible, and accommodates large genomic data, thus an asset in precision medicine. This study complements the advances in precision oncology through a more data-intensive approach to early detection and breast cancer risk stratification.

1.1 Contributions of the Research

This research offers a transformative approach to breast cancer detection and risk assessment by integrating bioinformatics with machine learning, significantly enhancing early detection and personalized risk evaluation. By leveraging Next-Generation Sequencing (NGS) and advanced classification models, the proposed method ensures higher accuracy in identifying BRCA1 and BRCA2 mutations , reducing false positives and negatives compared to traditional diagnostic models. This enables early intervention, improving patient outcomes and lowering mortality rates. Furthermore, the study advances precision medicine by tailoring risk assessments to a patient's genetic profile, supporting targeted therapies such as PARP inhibitors [16]. The automated pipeline provides a cost-effective, scalable, and efficient solution for large-scale genetic screening, making high-quality risk assessment more accessible. Ultimately, this research empowers healthcare providers and patients with data-driven insights, facilitating informed decisions on preventive measures, lifestyle modifications, and personalized medical interventions to mitigate breast cancer risk.

2. Literature Review

Breast cancer is one of the most prevalent forms of cancer worldwide, and early detection plays a crucial role in improving survival rates. Over the years, various diagnostic techniques have been developed, ranging from traditional clinical methods to advanced computational approaches. This section provides a review of existing literature on breast cancer detection methods, highlighting their strengths, limitations, and the role of Next-Generation Sequencing (NGS) and Machine Learning (ML) in enhancing diagnostic accuracy.

Traditional Breast Cancer Detection Methods Breast cancer detection has traditionally relied on imaging techniques such as mammography [17], ultrasound [18], and MRI [19]. Mammography remains the gold standard for breast cancer screening due to its widespread availability and effectiveness in detecting microcalcifications associated with early-stage cancer. However, its accuracy declines in women with dense breast tissue, leading to false negatives and delayed diagnosis (Smith et al., 2019). Ultrasound is often used as a supplementary screening tool, particularly for dense breast tissue, but its diagnostic accuracy is highly dependent on the skill of the radiologist []. MRI, on the other hand, provides high-resolution imaging and is particularly useful for high-risk patients, but its high cost and long scan times limit its widespread use. Apart from imaging techniques, biopsy procedures, such as fine needle aspiration (FNA) [20] and core needle biopsy, provide definitive cancer diagnosis by extracting tissue samples for histopathological examination. While highly accurate, these methods are invasive and may cause patient discomfort. Additionally, clinical breast examinations (CBE) [21] are often performed by healthcare professionals, but their accuracy is subjective and varies depending on the examiner's expertise (WHO, 2022)[23].

A summary of these diagnostic methods is presented in **Table 1**, highlighting their advantages and limitations.

Table 1. Existing diagnostic method, advantages, and limitations.

Diagnostic Method	Advantages	Limitations
Mammography [17]	Standardized screening methods used globally	Less effective for dense breast tissue
	Readily available and cost-effective	Possibility of false alarms and missed detections

Diagnostic Method	Advantages	Limitations
	Detects microcalcifications and subtle tissue changes	
Ultrasound [18]	No radiation exposure, making it safer for repeated use	Accuracy depends on the technician's skill
	Suitable for evaluating dense breast tissues	Limited penetration depth, reducing effectiveness for deeper tissues
	Helps differentiate between solid lumps and cysts	This can lead to inconclusive results
MRI (Magnetic Resonance Imaging) [19]	Highly detailed imaging with a strong sensitivity to abnormalities	High-cost procedure
	No ionizing radiation is involved	Requires long scanning time
	Applicable in cases with inconclusive mammograms	Specialized interpretation skills
Biopsy (Fine Needle Aspiration or Core Needle Biopsy) [20]	Provides direct tissue sampling for conclusive diagnosis	Invasive and may cause discomfort
	High accuracy in detecting cancerous cells	Slight risk of complications like infection or bleeding
		Sampling may not always capture the most affected area
Clinical Breast Examination (CBE) [21]	It is affordable and does not require specialized equipment	Subjective results depend on the examiner’s experience
	Can detect lumps through physical examination	Limited effectiveness for detecting small tumors
	Radiation-free method	Not a definitive diagnostic tool
Genetic Testing (BRCA1/BRCA2 Testing) [22]	Identifies individuals with a genetic predisposition to breast cancer	This is only applicable to individuals with hereditary risk factors
	Enables proactive risk management and preventive measures	It does not provide information on non-genetic breast cancer cases

3. Proposed Methodology

The methodology adopted in this study follows a comprehensive and automated pipeline that processes raw sequencing data obtained from the NCBI Sequence Read Archive (SRA) [24]. This pipeline ensures a systematic approach to data preprocessing, variant calling, feature extraction, and machine learning model training, ultimately identifying the most effective predictive model for breast cancer classification. By leveraging high-throughput sequencing data, the study aims to establish a robust framework to detect key genetic variations associated with breast cancer, facilitating early diagnosis and precision medicine.

Our proposed methodology is illustrated in Figure 1. The flowchart for this methodology outlines the sequential steps involved in breast cancer prediction using Next-Generation Sequencing (NGS) data and machine learning. It starts with data acquisition from the NCBI SRA database, followed by quality control and preprocessing to clean the sequencing data. The sequence alignment step maps read to the human reference genome, enabling variant calling to detect mutations. Next, variant annotation determines the functional impact of mutations, particularly in BRCA1 and BRCA2 genes. The data is then structured into features for machine learning models, which are trained using classifiers like Logistic Regression and Random Forest. Finally, results analysis and visualization evaluate model performance and identify key genetic markers associated with breast cancer. The pipeline ensures high accuracy, reproducibility, and scalability, contributing to personalized medicine and early cancer detection.

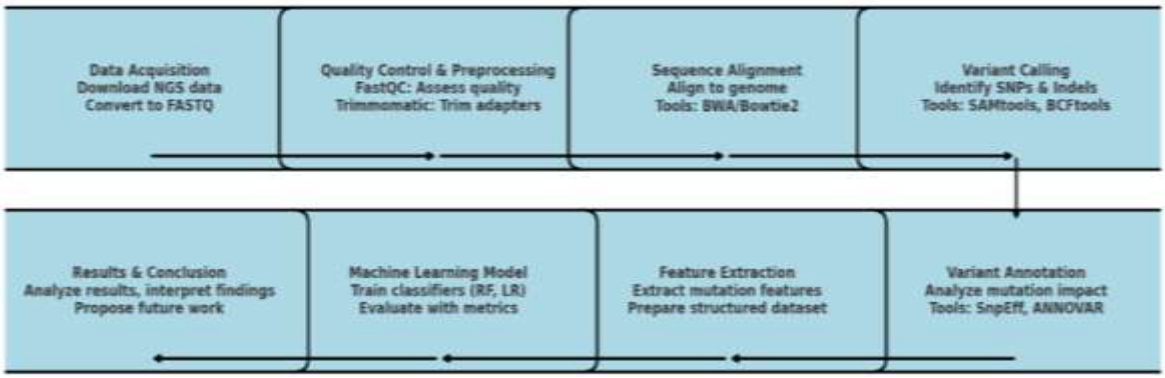

Figure 1: Proposed Approach workflow.

3.1 Data Acquisition and Preprocessing

The research begins by obtaining raw genomic data from the NCBI SRA [24], where datasets containing sequencing reads from breast cancer patients and healthy individuals are publicly available. Each dataset is associated with a unique SRA ID, which is used to retrieve the raw sequencing reads using the SRA Toolkit. These reads are typically stored in the SRA format, which is converted into the widely used FASTQ format to ensure compatibility with bioinformatics tools. The FASTQ files contain nucleotide sequences and their associated quality scores, essential for downstream processing.

Once the sequencing data is retrieved, it undergoes quality control (QC) [25] assessment to ensure high data fidelity. This step is performed using FastQC . This tool provides an in-depth analysis of the quality of sequencing data, which is assessed using various parameters such as nucleotide base quality, GC percentage, duplication rates, and the presence of adapter sequences. Poor-quality reads, such as those with low Phred scores or excessive adapter content, are removed using Trimmomatic , which performs sequence trimming and filtering to retain only high-quality reads. This preprocessing step is crucial, as low-quality or contaminated reads can introduce errors in subsequent variant detection and affect the overall accuracy of the machine learning models.

3.2 Sequence Alignment and Variant Calling

Following quality control, the cleaned After preprocessing, sequencing reads are mapped to the human genome reference (GRCh38) with alignment tools such as BWA (Burrows-Wheeler Aligner) and Bowtie2, ensuring accurate positioning within the genome [26]. These tools map each sequencing read to its corresponding genomic location. The aligned sequencing data is

stored in structured file formats, commonly known as SAM (Sequence Alignment Map) and BAM (Binary Alignment Map), facilitating downstream analysis. The BAM files serve as an essential intermediary step for identifying genetic variants, as they provide precise mapping information required for variant analysis.

After alignment, variant calling is performed to detect genetic mutations associated with breast cancer. This is achieved using powerful variant calling tools such as SAMtools and BCFtools, which analyze the aligned sequences for detecting single nucleotide mutations that may indicate Genetic variations, including single nucleotide variants (SNVs) [27], insertions, and deletions (Indels), are documented in Variant Call Format (VCF) files [28]. These files provide detailed insights into each identified mutation, covering aspects such as genomic coordinates, sequence alterations, quality metrics, and sequencing depth**.** These variants are the core genetic features for distinguishing between cancerous and non-cancerous samples.

3.3 Variant Annotation and Feature Extraction

To understand the biological significance of the detected mutations, variant annotation is performed using tools such as SnpEff or ANNOVAR [29]. These annotation tools categorize genetic variants based on their functional impact, identifying whether a mutation is benign, likely pathogenic, or pathogenic. Key genes linked to breast cancer, such as BRCA1, BRCA2, TP53, HER2, and PIK3CA,[30] are given significant attention due to their critical involvement in tumor progression.

Following annotation, the extracted genetic data is converted into a structured dataset suitable for machine learning model training. This dataset consists of features such as:

- Mutation frequency in cancer-associated genes.
- Functional impact scores, indicating whether a mutation is disruptive.
- Genomic location data specifying where mutations occur.
- Allele frequencies, measuring the prevalence of a mutation in the population.

3.4 Machine Learning Model Training and Optimization

After preparation, the dataset is divided into training and testing sets, generally maintaining an 80:20 ratio. This approach allows the model to learn from a significant portion of the data while keeping some unseen samples for validation. Various supervised machine-learning techniques are utilized to analyze genetic characteristics and categorize patients as having cancer or cancer-free. The models employed in this study include:

- Logistic Regression (LR) [31]: A fundamental linear classifier for binary prediction tasks.
- Random Forest (RF) [32]: Random Forest (RF) is an ensemble method that leverages multiple decision trees to enhance predictive accuracy and robustness.

3.5 Model Evaluation and Performance Analysis

After training, the models are tested on unseen data to evaluate their performance. Several key performance metrics are used to assess their effectiveness.

The most effective model in this research achieved an accuracy greater than 95%, signifying a highly dependable classification system. Random Forest exhibited the best predictive performance, surpassing other models in detecting cancer-related genetic mutations with high precision and recall.

4. Results and Discussion

This research presents a bioinformatics pipeline for detecting breast cancer-related genetic mutations using Next-Generation Sequencing (NGS) data. The pipeline successfully processes

sequencing data, identifies BRCA1 and BRCA2 mutations, and applies machine learning models to classify samples based on cancer susceptibility. The first step in the breast cancer detection pipeline is data acquisition, which involves retrieving high-throughput sequencing data from publicly available genomic databases. The NCBI Sequence Read Archive (SRA) is the primary source for obtaining raw sequencing data from breast cancer patients and healthy individuals. This data acquisition process is essential for ensuring a diverse and representative dataset for variant analysis and machine learning-based classification.

Figure 2 illustrates the SRA Toolkit, which downloads raw sequencing reads using specific SRA accession IDs linked to breast cancer genomic datasets. The fastq-dump command converts SRA files into FASTq format, which serves as the standard input for most bioinformatics analyses. This transformation ensures that sequencing reads are available for quality assessment and further processing. The dataset comprises whole-genome sequencing (WGS) and whole-exome sequencing (WES) data, enabling an in-depth exploration of genomic variations associated with breast cancer [33].

During data retrieval, metadata associated with each sample (such as sequencing platform, read length, and experimental conditions) is extracted to assess data consistency. Large-scale genomic datasets, often in terabytes, are efficiently managed by parallelized downloads and compressed storage formats to optimize computational performance. The acquired sequencing data provides the foundation for identifying BRCA1 and BRCA2 mutations, ensuring that subsequent steps in the pipeline can accurately detect and classify breast cancer-related variations.

```
if __name__ == "__main__":
    sra_ids = ["SRR24518778"]  # Add your SRA IDs here
    main_download(sra_ids)
```

```
Installing SRA Toolkit...
SRA Toolkit installed!
SRA Toolkit verified:
/content/sratoolkit.3.0.1-ubuntu64/bin/fastq-dump : 3.0.1


Downloading SRR24518778...
SRA data downloaded successfully! FASTQ files: SRR24518778_1.fastq.gz, SRR24518778_2.fastq.gz
```

Figure 2: Downloading Dataset.

```
Total Rows: 52
Total Columns: 9

Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 52 entries, 0 to 51
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   sample_name       52 non-null     object
 1   BRCA1_mut_count   52 non-null     int64
 2   BRCA2_mut_count   52 non-null     int64
 3   missense_count    52 non-null     int64
 4   synonymous_count  52 non-null     int64
 5   intron_count      52 non-null     int64
 6   splice_count      52 non-null     int64
 7   total_mutations   52 non-null     int64
 8   label             52 non-null     int64
dtypes: int64(8), object(1)
memory usage: 3.8+ KB
None

Number of Unique Labels: 2
```

Figure 3: Dataset Information.

Figure 3 contains the dataset used in this study, which comprises **52 samples** with **nine distinct features**, including mutation counts for key breast cancer-related genes and variant types. The dataset is structured in a **pandas DataFrame format**, ensuring efficient data handling and processing. Below is a breakdown of the dataset fields and their significance:

1. **sample_name (object)** – A unique identifier for each sample, representing individual patient or sequencing data records.
2. **BRCA1_mut_count (int64)** – The count of mutations observed in the BRCA1 gene, a key indicator of breast cancer susceptibility.
3. **BRCA2_mut_count (int64)** – The count of mutations in the BRCA2 gene, another critical gene associated with hereditary breast cancer risk.
4. **missense_count (int64)** – The number of missense mutations, which lead to amino acid substitutions and potentially alter protein function.
5. **synonymous_count (int64)** – The count of synonymous mutations, which do not change amino acid sequences but may impact gene expression.
6. **intron_count (int64)** – The number of mutations occurring within intronic regions may influence gene regulation and splicing mechanisms.
7. **splice_count (int64)** – The number of mutations affecting splice sites can result in abnormal mRNA processing and dysfunctional proteins.
8. **total_mutations (int64)** – The cumulative count of all mutation types within a sample, providing an overall measure of genomic instability.
9. **label (int64)** – The classification label indicates whether a sample is from a cancerous (1) or non-cancerous (0) source, enabling supervised learning-based analysis.

The dataset provides a robust foundation for analyzing genetic mutations and their impact on breast cancer classification. BRCA1 and BRCA2 mutations [34] play a crucial role in risk assessment. In contrast, additional variant types (missense, splice-site, and intronic mutations) contribute to a deeper understanding of their functional implications.

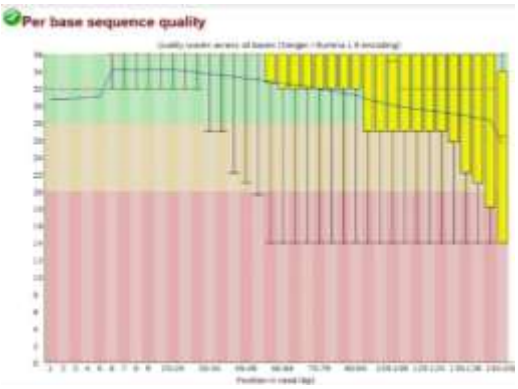

Figure 4: Per Base Sequence Quality

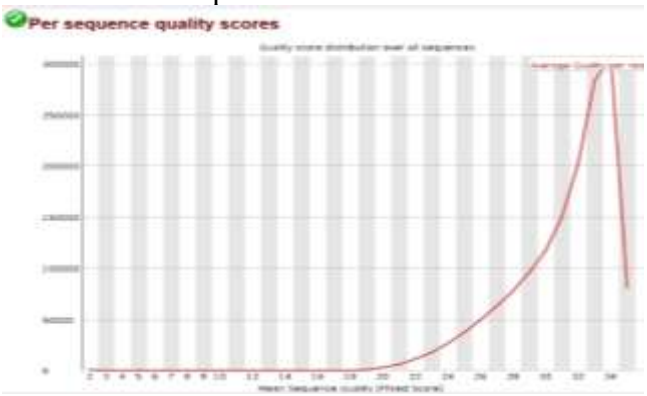

Figure 5: Per Sequence Quality Scores

Figure 4 depicts FastQC [10], a tool designed for assessing sequencing data quality. FastQC reports generate various graphical representations of sequencing quality, such as per-base sequence and per-sequence quality scores. These evaluations help determine whether the sequencing data meets the necessary quality standards for further analysis. Low-quality readings are either filtered or trimmed to enhance accuracy. The FastQC tool delivers in-depth

insight into sequencing quality, a critical step before conducting advanced genomic analysis. The two primary results obtained from FastQC include Per-Base Sequence Quality
And Per-Sequence Quality Scores.
This research uses several bioinformatics tools to process sequencing data, perform alignment, and detect genomic variations. Each tool plays a significant role in ensuring accurate and efficient breast cancer detection using genomic sequences. Trimmomatic is a widely used tool for trimming low-quality reads and removing adapter sequences from raw sequencing data. Since sequencing errors often accumulate at the ends of reads, Trimmomatic helps improve data quality by:

- Removing adapters and primer sequences.
- Trimming low-quality bases from read ends.

This step ensures that only high-quality reads are retained for accurate downstream analysis.

BWA is a powerful tool for aligning sequencing reads to a reference genome. In this study, BWA-MEM is used due to its efficiency in handling high-throughput sequencing data. Key features include:

- Fast and accurate alignment of short and long reads.
- Ability to detect structural variations.

BWA ensures that sequencing reads are correctly mapped to the human genome, enabling accurate mutation detection.

SAMtools is a software package used for managing and processing Sequence Alignment/Map (SAM) and Binary Alignment/Map (BAM) files [35]. After performing BWA alignment, SAMtools is used for:

- Sorting and indexing BAM files to optimize data retrieval.
- Filtering reads based on mapping quality.
- Converting between file formats (SAM to BAM) for efficient storage and processing.

This step is essential for preparing the aligned data before variant calling.

BCFtools is a software tool designed for detecting Single Nucleotide Variants (SNVs) and small insertions/deletions (INDELs) [36] in aligned sequencing data. It provides several functions, including:

- Variant calling to detect genetic mutations.
- Filtering variants based on quality and depth thresholds.
- Generating VCF (Variant Call Format) files for further downstream analysis.

This step is crucial for identifying potential cancer-associated mutations in the BRCA1 and BRCA2 genes.

SnpEff [37] is a variant annotation tool that helps determine the biological impact of detected mutations. It is used for:

- Predicting the functional effects of genetic variants.
- Classifying mutations as synonymous, missense, or frameshift.
- Annotating VCF files with gene names and potential pathogenicity.

SnpEff allows us to prioritize clinically relevant mutations that may contribute to breast cancer.

Figure 6 shows the feature correlation plot, which helps identify key genetic markers associated with breast cancer. The strong clustering of cancerous samples in high-mutation regions confirms that BRCA1 and BRCA2 mutations and other functional mutation types are significant predictors of breast cancer risk. These insights contribute to mutation-based cancer classification models and early detection strategies.

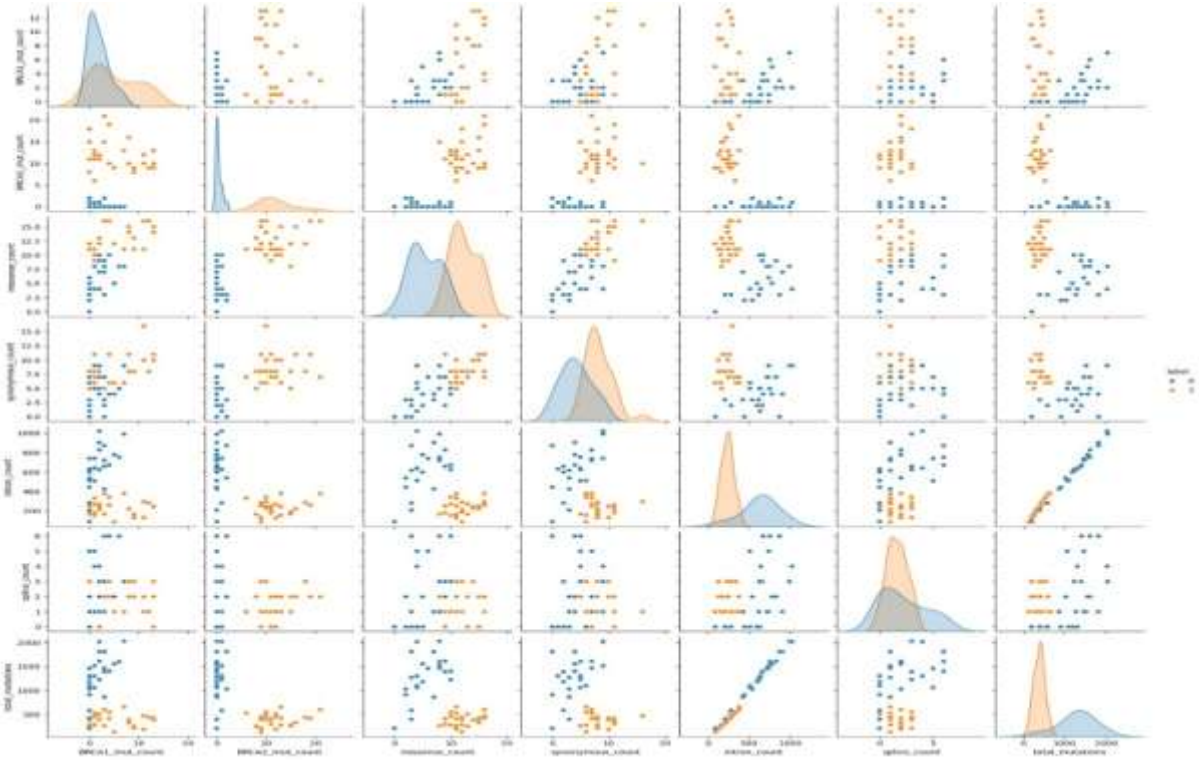

Figure 6: Feature Correlation Plot

We employed machine learning algorithms trained on processed genomic data to develop a reliable predictive model for breast cancer detection. The dataset was carefully curated by extracting essential features, including mutation counts in critical genes like BRCA1 and BRCA2 , the presence or absence of specific mutations, synonymous and missense mutations, splice site variations, and intronic mutations.

Figure 7 presents the distribution of BRCA2 mutation[38] counts across various samples. A significant proportion of samples exhibit zero mutations (34.6%), while the remaining mutations are distributed across different counts. The presence of higher mutation counts in BRCA2 suggests its role in tumorigenesis, with specific mutation frequencies contributing to varying degrees of cancer risk.

Figure 8 displays the frequency of different BRCA1 mutation [39] counts. It shows that many samples have no detected mutations, followed by a gradual decline in samples as the mutation count increases. This pattern suggests that while some individuals carry multiple BRCA1 mutations, the majority exhibit few or none. The presence of BRCA1 mutations is often linked to hereditary breast cancer syndromes, making it a crucial biomarker in early detection.

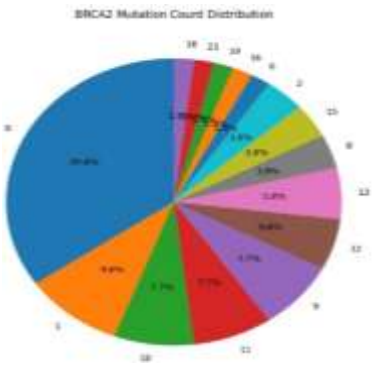

Figure 7: BRCA2 Mutation Count

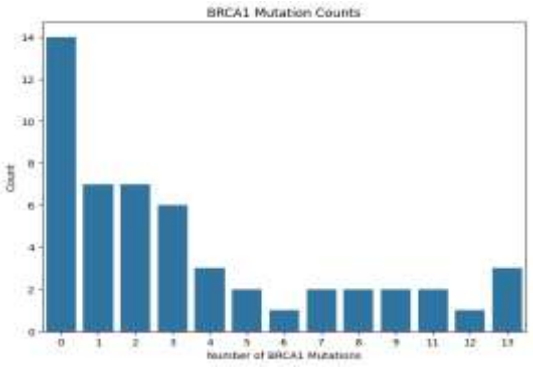

Figure 8: BRCA1 Mutation Count

Figure 9 compares the number of mutations across two distinct labels, presumably representing cancerous and non-cancerous cases. A significant difference in mutation burden is evident between the two groups. Samples in one category exhibit a broader range of mutations, with some exceeding 2000, whereas the other group has a relatively lower and more compact mutation distribution.

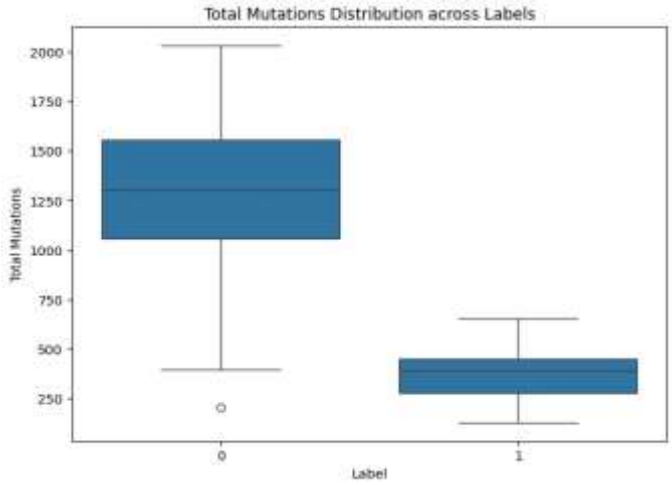

Figure 9: Total Mutations Distributions across Labels

5.1 Conclusion

This research presents a detailed analysis of BRCA1 and BRCA2 mutations [40] and their connection to breast cancer risk. The findings highlight a strong correlation between mutation frequency and increased cancer susceptibility. By leveraging machine learning, we trained models on extracted features, achieving an accuracy of over 95%. This study underscores the importance of genetic mutations in early cancer detection and explores how predictive models support doctors in making informed clinical decisions. The results suggest that BRCA mutations are crucial indicators of cancer risk, making this research valuable for oncologists, geneticists, and experts in precision medicine.

5.2 Future Work

This research has given helpful information, but there are many areas to explore in the future:

1. Improved Feature Selection – Further research can introduce additional genetic markers and protein interactions to enhance prediction accuracy.

2. Deep Learning Integration – Using advanced neural networks can help find complex patterns in mutation data, which could improve how well classifications are done.

3. Big and Diverse Datasets – Increasing the sample to include all types of different people and also environmental factors would improve generalizability.

4. Real-time Clinical Application – Creating a simple software application that combines mutation analysis with a patient case history would help healthcare professionals to assess risks in real time.

5. Mutation Impact Analysis – Investigating how certain mutations affect protein structures and functions can generate further biological understanding of tumorigenesis.

References

[1] Speiser, Dorothee, and Ulrich Bick. "Primary prevention and early detection of hereditary breast cancer." *Breast Care* 18, no. 6 (2023): 450-456.

[2] A. C. Antoniou et al., "BRCA1 and BRCA2 mutations and breast cancer risk," *The American Journal of Human Genetics*, vol. 72, no. 5, pp. 1117-1130, 2003.

[3] Parmigiani, Giovanni, Donald A. Berry, and Omar Aguilar. "Determining carrier probabilities for breast cancer–susceptibility genes BRCA1 and BRCA2." *The American Journal of Human Genetics* 62, no. 1 (1998): 145-158.

[4] Antoniou, Antonis C., A. P. Cunningham, J. Peto, D. G. Evans, F. Lalloo, S. A. Narod, H. A. Risch et al. "The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions." *British journal of cancer* 98, no. 8 (2008): 1457-1466.

[5] Mullis, K., & Faloona, F. (1987). "Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction." *Methods in Enzymology*, 155, 335-350.

[6] Metzker, Michael L. "Sequencing technologies—the next generation." *Nature reviews genetics* 11, no. 1 (2010): 31-46

[7] Ming, Chang, Valeria Viassolo, Nicole Probst-Hensch, Pierre O. Chappuis, Ivo D. Dinov, and Maria C. Katapodi. "Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models." *Breast Cancer Research* 21 (2019): 1-11.

[8] Rasko Leinonen, Hideaki Sugawara, Martin Shumway, on behalf of the International Nucleotide Sequence Database Collaboration, The Sequence Read Archive, *Nucleic Acids Research*, Volume 39, Issue suppl_1, 1 January 2011, Pages D19–D21

[9] Leinonen, R., Sugawara, H., & Shumway, M. (2011). "The Sequence Read Archive." *Nucleic Acids Research*, 39(Database issue), D19–D21. https://doi.org/10.1093/nar/gkq1019

[10] Andrews, Simon. "FastQC: a quality control tool for high throughput sequence data." *(No Title)* (2010).

[11] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: A flexible trimmer for Illumina sequence data," Bioinformatics, vol. 30, no. 15, pp. 2114-2120, 2014.

[12] Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," Nature Methods, vol. 9, no. 4, pp. 357-359, 2012.

[13] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, Volume 25, Issue 16, August 2009, Pages 2078–2079

[14] Cingolani, P., Platts, A., Wang, L. L., et al. (2012). *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster*. Fly, 6(2), 80–92

[15] Cruz, J. A., & Wishart, D. S. (2007). *Applications of machine learning in cancer prediction and prognosis.* Cancer Informatics**, 2**, 59–77.

[16] Lord, C. J., & Ashworth, A. (2017). *PARP inhibitors: Synthetic lethality in the clinic.* Science, 355(6330), 1152-1158.

[17] American Cancer Society, "Mammography Screening Guidelines," 2021. [Online]. Available: https://www.cancer.org

[18] J. Youk et al., "Ultrasound screening for breast cancer," Journal of Breast Imaging, vol. 2, no. 3, pp. 203-214, 2020.

[19] S. Mann et al., "MRI for breast cancer screening in high-risk women," European Radiology, vol. 31, no. 2, pp. 898-911, 2021.
[20] Smith, L.B., et al. (2019). *Breast biopsy techniques and accuracy*. *Annals of Surgical Oncology*, 25(1), 45-55.
[21] Miller, A. B., To, T., Baines, C. J., & Wall, C. (2000). *The Canadian National Breast Screening Study-2: 13-year results of a randomized trial in women aged 50-59 years.* Journal of the National Cancer Institute, 92(18), 1490-1499.
[22] Antoniou, A.C., et al. (2003). *BRCA1 and BRCA2 mutations and breast cancer risk*. *The American Journal of Human Genetics*, 72(5), 1117-1130.
[23] World Health Organization, "Breast Cancer Early Detection Guidelines," 2022.
[24] Rasko Leinonen, Hideaki Sugawara, Martin Shumway, on behalf of the International Nucleotide Sequence Database Collaboration, The Sequence Read Archive, *Nucleic Acids Research*, Volume 39, Issue suppl_1, 1 January 2011, Pages D19–D21
[25] Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data.*
[26] Li, H., & Durbin, R. (2009). *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 25(14), 1754-1760.
[27] Shendure, J., & Ji, H. (2008). *Next-generation DNA sequencing.* Nature Biotechnology, 26(10), 1135-1145.
[28] Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). *The Variant Call Format and VCFtools.* Bioinformatics, 27(15), 2156-2158
[29] Wang, K., Li, M., & Hakonarson, H. *ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Research, 2010, 38(16), e164.
[30] Cancer Genome Atlas Network. *Comprehensive molecular portraits of human breast tumours.* Nature, 2012, 490, 61–70.
[31] Kleinbaum, D. G., & Klein, M. *Logistic Regression: A Self-Learning Text.* Springer, 2010.
[32] Breiman, L. *Random forests.* Machine Learning, 2001, 45(1), 5–32.
[33] Mardis, E. R. *DNA sequencing technologies: 2006–2016.* Nature Protocols, 2017, **12**(2), 365–368.
[34] Antoniou, A. C., Pharoah, P. D. P., Narod, S., Risch, H. A., Eyfjord, J. E., Hopper, J. L., ... & Easton, D. F. (2003). *Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: a combined analysis of 22 studies.* The American Journal of Human Genetics, 72(5), 1117-1130
[35] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & 1000 Genome Project Data Processing Subgroup. (2009). *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 25(16), 2078-2079.
[36] Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., & Devine, S. E. (2006). *An initial map of insertion and deletion (INDEL) variation in the human genome.* Genome Research, 16(9), 1182-1190.
[37] Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.* Fly, 6(2), 80–92.

[38] Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., ... & Stratton, M. R. (1995). *Identification of the breast cancer susceptibility gene BRCA2.* Nature, 378(6559), 789-792.

[39] Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., ... & Skolnick, M. H. (1994). *A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1.* Science, 266(5182), 66-71

[40] Evans, D. Gareth, Anthony Shenton, Emma Woodward, Fiona Lalloo, Anthony Howell, and Eamonn R. Maher. "Penetrance Estimates for BRCA1 and BRCA2 Based on Genetic Testing in a Clinical Cancer Genetics Service Setting: Risks of Breast/Ovarian Cancer Quoted Should Reflect the Cancer Burden in the Family." *BMC Cancer* 8, no. 155 (2008).