Journal of Research High School

Volume 2021 (2) 50 - 60
ISSN 2688-3651 [online]

Convolutional Neural Network for Predicting Genetic Risks of Breast Cancer

Siva Bubby[1*]

[1] BASIS Scottsdale, Scottsdale, Arizona, USA

Received May 26, 2021; Revised August 10, 2021; Accepted, August 16, 2021

Abstract

Breast cancer is a complex disease with a growing global prevalence whose genetic causes remain largely unexplored. The rise of next generation sequencing has significantly augmented genetic studies in identifying breast cancer-associated mutations, the most common of which are single nucleotide polymorphisms (SNPs). While SNPs offer insights into the genetic causes of breast cancer, they do not explain its biological underpinnings nor do they provide a context within which to judge sequence-based interactions between SNPs linked to the disease. Convolutional Neural Networks (CNNs) demonstrate higher performance in hierarchical and abstract feature learning for image classification compared to other deep learning methods. This study proposes a deep learning model, named SPRiNGS, to classify a sample's genetic breast cancer risk by analyzing the sequence contexts surrounding its SNP composition. Firstly, Monte Carlo simulations were implemented to generate a sample cohort and corresponding Polygenic Risk Scores (PRS). Secondly, each sample's sequence composition matrix was resized to highlight significant semantic patterns across sequences. Thirdly, a two-dimensional CNN was constructed for feature learning and classification. This research demonstrated the validity of its simulated cohort. Additionally, SPRiNGS elucidates the improved performance of sequence-based predictions compared to SNP-based methods. The robustness of SPRiNGS was proved by experimental variations of the number of loci considered and the sequence fragment length.

Keywords: Breast Cancer, Risk Prediction, SNPs, Genomics

1. Introduction

Breast cancer is a complex disease with several subtypes and is influenced by myriad genetic and environmental factors. In 2020, the CDC categorized this disease among the "Top 10 Cancers" for its alarmingly increasing global prevalence and mortality rates (CDC, 2020). Current estimates suggest that approximately every 1 in 8 women could develop breast cancer in their lifetime ("U.S. Breast Cancer Statistics," 2020). Advancements in next generation sequencing technologies have greatly augmented nucleotide-based biomarker identification for various diseases, including breast cancer, primarily through genome-wide association studies (GWAS), which analyze the distribution of genomic mutations across cases and controls for a specified phenotype (Quezada, et al., 2017). These variants are mostly single nucleotide polymorphisms (SNPs), genetic mutations occurring at singular positions across a genome, and offer tremendous opportunity in precision medicine. While SNPs yield insights into

* Corresponding Author
sivabubby@live.com

Advisor: Dr. Michael Caplan
michael.caplan@pcds.org

the genetic causes of breast cancer, they alone do not explain its mechanisms nor do they provide a context to judge the underlying semantic interactions between SNPs. One strategy to overcome these issues is to incorporate the local DNA fragment surrounding each SNP into existing risk prediction models. By extracting DNA sequences around GWAS breast cancer mutations, this research hypothesizes that incorporating the local DNA sequence context around statistically significant mutations into disease prediction algorithms will outperform existing SNP-based methods in classifying genetic risks of breast cancer. This study proposes Sequence-based Polygenic Risk Network for GWAS SNPs (SPRiNGS), a novel computational method for classifying sequence-based breast cancer risks.

2. Materials and Methods

2.1 Data Collection and Preprocessing

This research utilized the SNPs available in the 'Breast Carcinoma' dataset from the online *GWAS Catalog* (Buniello, et al., 2019). To extract all SNPs with explicit corresponding genomic loci, the dataset was filtered to remove all entries missing both a risk allele and locus. Subsequently, for all entries without either an SNP or genomic coordinate, the missing information was manually extracted from *SNPedia* (Cariaso & Lennon, 2012) or *dbSNP* (Sherry, et al., 2001), thereby ensuring that each SNP had a corresponding locus. The remaining mutations were then filtered by statistical significance using the common GWAS *p*-value threshold $\alpha = 5 \times 10^{-8}$ (Fadista, et al., 2016). Finally, the relative strength of each allele β was extrapolated to accurately characterize the individual effects of each locus on a sample's breast cancer risk.

2.2 Sequence Extraction

The genomic location of each SNP was expanded into symmetric ranges about the risk allele. BEDTools, a powerful toolset for genomic arithmetic (Quinlan & Hall, 2010), then converted each of these ranges into DNA sequences using the *GRCh38* reference genome (Schneider, et al., 2017). A reference genome is "a digital nucleic acid sequence database assembled by scientists as a representative example of the set of genes in one idealized individual organism of a species" ("Reference genome," 2020). Extraction from a reference genome indicated that the emergent sequences were not associated with any disease (including breast cancer) because they lacked the risk-associated alleles (and were thus classified as "healthy"). To generate the breast cancer-associated sequences, risk alleles were substituted into the sequences surrounding their corresponding loci. This process generated one healthy and one risk-associated sequence for each statistically significant SNP location.

2.3 Monte Carlo Simulation

Previous genome-wide prediction studies primarily perform genomic sequencing on a sample cohort of cases and controls to extract their own specific SNP collection prior to risk analysis. This means that researchers would analyze the SNP composition of individuals who had a particular disease and those who did not. These data are then stored in databases for other researchers to use. However, these repositories often limit public access due to medical privacy and other ethical constraints. After searching popular databases including *cBioPortal* (Cerami, et al., 2012), *GWASkb* (Kuleshov, et al., 2019), and *GWAS Central* (Beck, et al., 2020) for SNP-based breast cancer case-control samples, it was found that the accessible data only described a sample's mutated gene composition. However, since one gene can contain multiple SNPs, this type of data would not be suitable for this experiment. To circumvent this, Monte Carlo methods were implemented to emulate real-world conditions when generating a representative random sample of breast cancer risk scores. Monte Carlo methods rely on stochasticity to predict a deterministic result (Adekitan, 2014). For this study, each sample's risk score was the deterministic result (see 'Polygenic Risk Score Calculation') and the sample's sequence composition represented the stochastic component.

For this sample to be truly representative, all possible risk scores needed to be accounted for.

Across all samples, the Monte Carlo simulation randomly selected between healthy and risk sequences for each SNP location with varying probabilistic frequencies. This method ensured a random sample which accounted for all risk scores for the number of statistically significant SNPs.

The sequence data for all samples were stored in a 3D array whose height represents the total number of SNPs per sample s, width is the sequence length L (see 'Sequence Encoding'), and depth represents the total number of samples N (Figure 1). The sample size was fixed at N = 20,000 due to computational constraints. The other two variables were evaluated in a sensitivity analysis to measure how they impact the model's performance. Specifically, the number of SNPs is strictly determined by statistical significance. Hence, variations in s were created by fluctuating the p-value threshold α. The threshold was tightened ($\alpha = 5 \times 10^{-9}$) and relaxed ($\alpha = 5 \times 10^{-7}$) to obtain differing numbers of mutated loci while maintaining sufficient amounts of data required to train a machine learning algorithm.

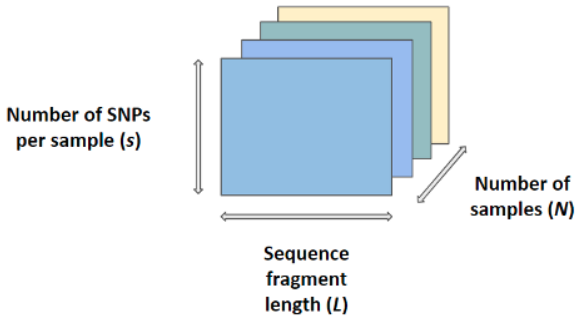

Figure 1. *Three-dimensional data structure of Monte Carlo sample population.* This simulation encodes each sample as an $s \times L$ matrix which describes the sequence composition of a particular individual.

2.4 Sequence Encoding

One-hot encoding is a popular technique to convert nucleotide sequences into binary sequences. For any given genomic sequence of length l, the length of the corresponding one-hot encoded sequence l_{ohe} is $4l$ due to the four nucleotide possibilities (A, C, G, T) at each location along the nucleotide sequence. However, l_{ohe} explodes as l increases, making one-hot encoding impractical for large sequences. To avoid high-dimensional data, genomic sequences were converted to vectors of unique numbers between [1, 4], thereby preserving sequence length and explicitly differentiating alleles. This process encodes each sample as an $s \times L$ matrix characterizing its sequence composition. Matrices were resized to dimensions 28 × 28 to bring similar semantic patterns closer together and highlight important genomic features across samples. Moreover, making these matrices denser decreases the computational storage and time required for analysis.

2.5 Polygenic Risk Score Calculation

Each sample's breast cancer risk was calculated using a weighted Polygenic Risk Score (PRS) based on its genetic sequence composition, as seen in Equation (1), where i was an integer within [1, N] representing the sample index, j represented the sequence number between [1, s], s_{ij} was the sequence classification of 'healthy' or 'breast cancer' as 0 or 1 respectively, and β_{ij} represented the sequence's relative strength based on its particular risk allele.

$$risk_i = \sum_{j=1}^{s} s_{ij} \times \beta_{ij} \quad (1)$$

While this risk calculation method assesses the cumulative impact of a sample's sequence composition, it does not account for the occurrence-based or sequence-based interactions between SNPs. Therefore, machine learning is implemented to account for these complex underpinnings (see 'Convolutional Neural Network').

To categorize samples 'healthy' or 'at risk of breast cancer,' all risk scores were normalized between [0, 1]. Samples whose normalized risk scores were above the 50th percentile were classified as 'at risk of breast cancer,' and the remainder were classified as 'healthy.' This method yielded a population with 50% breast cancer prevalence, which ensures an equal number of cases and controls in the dataset to avoid a biased training procedure for the model.

2.6 Convolutional Neural Network

Convolutional Neural Network (CNN) is a subset of deep learning largely popularized for image classification because it adaptively learns and generalizes hierarchical spatial features (Indolia, et al., 2018). CNNs split their learning processes into "building blocks, such as convolution layers, pooling layers, and fully connected layers" to extract features from multidimensional data (Yamashita, et al., 2018). In convolution layers, the model iterates over samples using copious filters and stride windows of specified dimension to learn patterns among the input data. The pooling layers reduce the dimensionality of the model's feature matrices as it continues to learn. Finally, the fully connected network makes predictions with the vector representations of the extracted features. These models specifically capture explicit and implicit patterns within data using fewer hyperparameters compared to other deep learning methods.

This research applied a two-dimensional CNN to classify breast cancer risks based on genomic sequence patterns around risk-associated loci. Binary cross entropy (Deng, 2012) was used as the loss function and the Adam algorithm (Kingma & Ba, 2015) was applied for optimization. Of the 20,000 samples in this study's dataset, 70% were used to train the model and 30% were used to evaluate its performance. All hyperparameters were tuned using random grid search, a data analytics technique which trains copious models using randomly created hyperparameter combinations within user-specified ranges to extract the settings which yield the highest predictive accuracy. The model constructed in this study, named Sequence-based Polygenic Risk Network for GWAS SNPs (SPRiNGS), was implemented in Google Colab using the Keras platform in R because it allowed for free large-scale computation on a virtual machine.

2.7 Performance Evaluation

This experiment evaluated SPRiNGS with two metrics: Area Under the Receiver Operating Characteristics Curve (AUC) and normalized Matthews Correlation Coefficient (nMCC). In binary classification problems, ROC Curves depict a model's robustness by plotting the true positive classification rate against the false positive classification rate and calculating the area beneath the graph. Significant AUC values can range from [0.5, 1], where values tending toward 0.5 indicate poor classification ability while values closer to 1 indicate greater model performance.

Additionally, Matthews Correlation Coefficient (MCC) measures the model's statistical accuracy in the context of its confusion matrix. This metric was used to assess the alignment between predicted risk classifications and actual risk classifications, ranging from [-1, 1]. nMCC is calculated by rescaling MCC values between [0, 1], where 0 indicates total misalignment, 0.5 indicates random alignment (as if the model was guessing), and 1 indicates total alignment amongst predicted and actual classes. Since AUC and nMCC depend on a model's decision threshold, the threshold yielding the greatest accuracy was chosen prior to evaluation.

All code used for this study can be found here: https://github.com/sivab468/SPRiNGS.

3. Results

3.1 Simulation Validation

This research implemented Monte Carlo simulations to generate 20,000 samples whose breast cancer risks were dependent on their sequence composition. The validity of the simulation is examined based on whether this method extracts causal breast cancer variants, follows a normal PRS distribution, and associates higher PRS with greater chances of developing breast cancer. These criteria were found in "A Guide to Performing Polygenic Risk Score Analyses" (Choi, et al., 2020).

To assess whether causal breast cancer SNPs were extracted, all SNPs were graphed in a quantile-quantile (Q-Q) plot such that more significant SNPs would appear higher on the graph (Figure 2). Statistical significance was measured using -log(*p*-value) for ease of comparison. The black dots represent the observed data points and the blue line represents a normal distribution. While SNPs of lower significance somewhat align to the normal

distribution, the upward tail on the right indicates that the simulation used in this study extracted causal breast cancer variants amongst other mutations in the dataset.

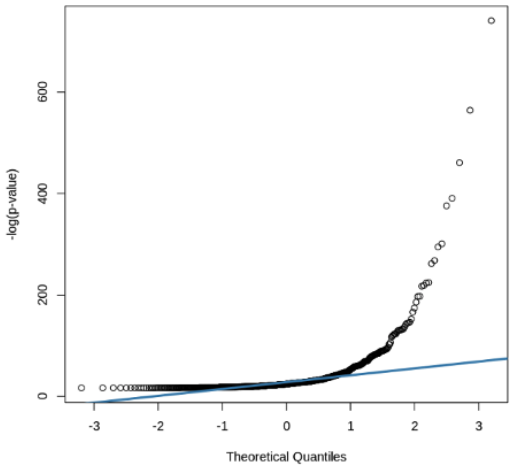

Figure 2. *Quantile-Quantile Plot of SNP Statistical Significance.* This graph depicts all SNPs in the dataset by their -log(*p*-value) such that more significant SNPs would appear higher on the graph. All points above the blue line represent causal breast cancer variants.

Next, this study analyzed the impact of PRS calculations on the breast cancer risk score distribution. Ideally, since risk scores are calculated as linear combinations of independent variables (SNPs), then standard PRS distributions should be normal (Choi, et al., 2020). To confirm the normality of the PRS distribution, a Q-Q plot of the breast cancer risk scores was created with the theoretical probability segments of a normal distribution on the x-axis and PRS on the y-axis (Figure 3). The black dots represent the observed risk scores and the red line represents a normal distribution. The more the data points align with the red line, the closer the distribution is to normal. This study quantifies the alignment between the PRS distribution and normal distribution using an R-squared value between [0, 1]. With R-squared = 0.99, this plot demonstrates this population's risk scores follow a near exact normal distribution.

The final measurement to validate this Monte Carlo simulation was to assess how breast cancer probabilities vary with PRS. Theoretically, population strata with higher PRS are more likely to develop breast cancer since they have more SNPs (and therefore have a higher odds ratio for the disease). This research divided the population into 20 equal subgroups, calculated each sample's probability of developing breast cancer for each subgroup, and converted all probabilities into odds ratios of developing breast cancer. To better visualize the data, the log(Odds Ratios) were calculated (denoted as Ψ) and plotted against PRS quantiles (Figure 4). Each point represents the mean Ψ for the population subgroup. The blue bars indicate the 95% confidence interval for each mean Ψ. The overall upward trend verifies that samples with higher PRS are more likely to develop breast cancer.

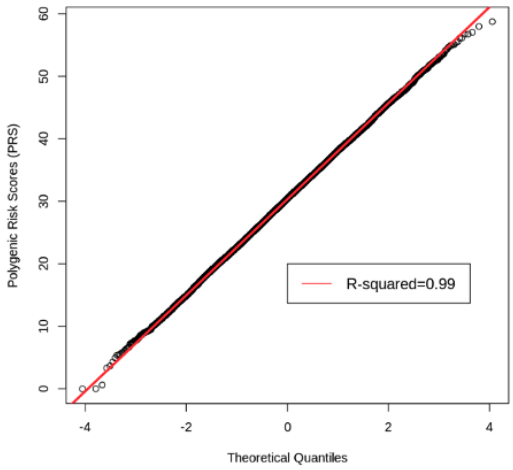

Figure 3. *Quantile-Quantile Plot of Breast Cancer Polygenic Risk Scores for Simulated Population*. This graph demonstrates the normality of the PRS distribution.

The three aforementioned figures confirm the validity of the Monte Carlo simulation presented in this experiment by extracting causal variants, observing a normal PRS distribution, and verifying the positive correlation between PRS and chances of developing breast cancer among the sample cohort.

3.2 Hyperparameters

This study consists of two hyperparameter categories: model-related and data-related. Model-related settings entail the various

hyperparameter combinations in SPRiNGS used during training and evaluation. Optimizing the model employed random grid search to sample 5% of all possible hyperparameter combinations and tune SPRiNGS with lower computational cost. The settings which minimized the loss value on the testing set were considered as the optimal hyperparameters. If more than one combination achieved the same minimum loss value, the combination which maximized the AUC value for the testing set was selected as optimal. All hyperparameters explored in SPRiNGS are summarized in Table 1 with the optimal settings bolded. All hyperparameters not explicitly mentioned remained at their default values.

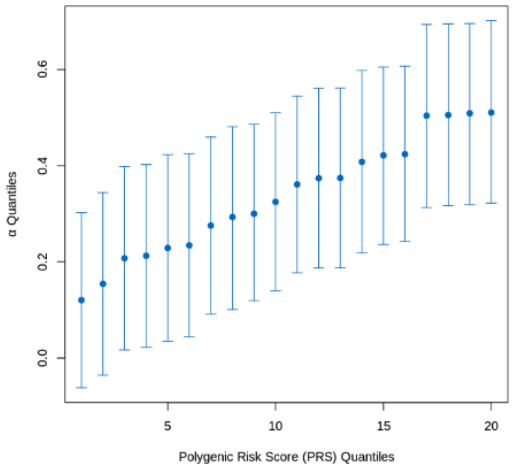

Figure 4. *Quantile-Quantile Plot of log(Odds Ratios) for Developing Breast Cancer across Stratified Polygenic Risk Scores.* This plot demonstrates the positive correlation between breast cancer PRS and the log(Odds Ratio) of developing breast cancer (Ψ).

Previous SNP-based disease prediction studies demonstrate that including more polymorphisms in PRS calculations typically yield more robust results; however, this data characteristic has not been explored in sequence-based predictions. Moreover, the impact of the DNA context length on disease prediction remains unknown. Therefore, this research treated these values as data-related hyperparameters and discusses their impact below (see 'Sensitivity Analysis').

Table 1. *SPRiNGS Hyperparameter Combinations and Optimal Settings (bolded).* This table summarizes all settings considered when designing and tuning SPRiNGS. The settings which minimized the loss value on the testing set were selected as optimal (bolded).

Hyperparameter	Option
Convolution Filters 1	**32**, 64
Kernel 1 Length	3, 4, **5**
Kernel 1 Width	**3**, 4, 5
Convolution Filters 2	16, 32, **64**
Kernel 2 Length	**3**, 4
Kernel 2 Width	**3**, 4
Dense Nodes 1	**16**, 32
Dense Nodes 2	4, 8, **16**
Optimizer	**Adam**, RMSprop

3.3 SPRiNGS Performance Evaluation

To evaluate the impact of incorporating genomic sequence contexts into SNP-based predictions, this study developed two control models which only analyze the SNP composition across the sample cohort: a classical machine learning model (SVM_RBF) and a deep learning model (1D CNN). SPRiNGS, on the other hand, analyzed the sequence composition of the simulated population. The ROC Curves for each model were plotted to compare the robustness of their predictions (Figure 5). SPRiNGS achieved the highest AUC at 0.91, while SVM_RBF and 1D CNN followed at 0.88 and 0.84, respectively. This elucidates that the sequence-based model developed in this study extracted underlying semantic patterns and significant SNP-SNP interactions associated with breast cancer better than traditional SNP-based methods.

This study also measures the nMCC to discern the holistic statistical accuracy of each model in the context of its confusion matrix. nMCC values were calculated and plotted for their corresponding model (Figure 6). All models achieved significant nMCC scores, indicating that each applied its learned patterns to effective sample classification. However, SPRiNGS achieved the highest nMCC at 0.82, while SVM_RBF and 1D CNN followed at 0.78 and 0.73, respectively. This demonstrates that incorporating the local DNA fragments into breast cancer predictions

improves their overall predictive accuracy compared to SNP-based methods.

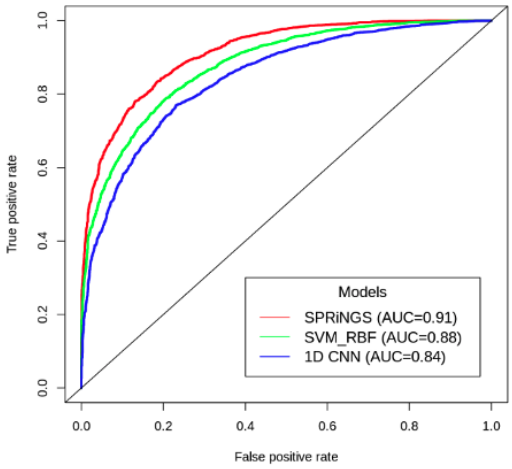

Figure 5. *ROC Curves for Breast Cancer Classification Models*. This graph demonstrates the robustness of breast cancer classifications achieved by SPRiNGS (red), SVM_RBF (green), and 1D CNN (blue) using AUC.

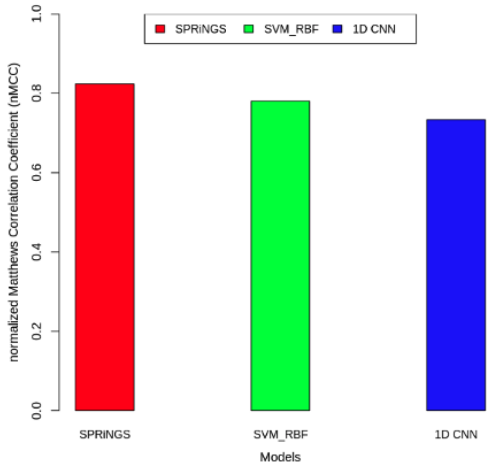

Figure 6. *nMCC Comparisons across Breast Cancer Classification Models.* This bar graph depicts the statistical accuracy of breast cancer classifications achieved by SPRiNGS (red), SVM_RBF (green), and 1D CNN (blue) using nMCC.

3.4 Sensitivity Analysis

This research conducted a sensitivity analysis to explore the impact of data-related characteristics (the number of SNPs s and the sequence length L) on SPRiNGS' performance. Changing the number of loci considered allows the model to discern genomic patterns across more sequences. To vary s, the p-value threshold was relaxed, yielding 904 SNPs ($\alpha = 5 \times 10^{-7}$). Additionally, the threshold was tightened, yielding 547 SNPs (to $\alpha = 5 \times 10^{-9}$). AUC was measured for each variation and plotted (Figure 7). Considering fewer SNPs displayed a slightly lower AUC; however, including more SNPs caused a significant decrease in AUC, likely attributable to overfitting.

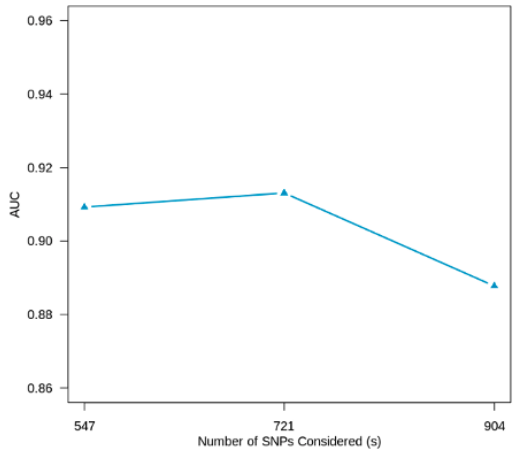

Figure 7. *Impact of number of SNPs (s) on SPRiNGS performance.* This line graph displays the impact of the number of SNPs considered (which was varied by the p-value threshold α) on SPRiNGS AUC score.

Moreover, this study examined the effect of sequence length L on the model's performance using short DNA sequence fragments. Increasing L would allow the model to identify more encoding patterns within sequences to distinguish between case and control samples. This analysis varied L between 13, 37, and 61 nucleotides. Sequences shorter than 13 nucleotides were not explored to ensure minimal genomic encoding motifs were captured during model training. Sequence lengths larger than 61 nucleotides were not explored due to computational storage constraints. The AUC was plotted to understand the impact of each L variation on model performance (Figure 8). While showing fairly consistent performance overall, the fragment length L = 37 displayed the highest AUC. The overall trend of the graph shows no significant correlation between model performance and sequence length.

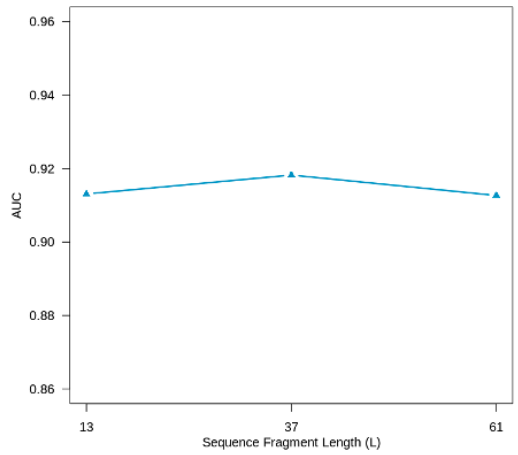

Figure 8. *Impact of sequence fragment length (L) on SPRiNGS performance.* This line graph displays the impact of the DNA sequence fragment length on SPRiNGS AUC score and shows that no significant correlation was observed between the two.

4. Discussion

This research proposes SPRiNGS, a two-dimensional convolutional neural network which analyzes genomic sequence patterns around statistically significant SNPs to classify breast cancer risks. After validating the Monte Carlo simulation used to generate a sample population, the results indicate that sequence-based breast cancer predictions outperformed SNP-based methods. Specifically, incorporating the sequence contexts of breast cancer-associated SNPs improved feature extraction and robustness (measured by AUC) and classification accuracy (measured by nMCC) compared to traditional SNP-based algorithms (such as SVM_RBF and 1D CNN). SPRiNGS analyzed the occurrence patterns and genomic interactions among significant mutations while the control models only analyzed the former. Since SPRiNGS associated higher polygenic risk scores with greater chances of developing breast cancer (as seen in Figure 4), the model discerned how SNP occurrence patterns impact genetic breast cancer risks. Moreover, incorporating DNA sequence contexts around each SNP allowed SPRiNGS to understand various encoded semantic patterns across mutations, which improved robustness and accuracy compared to SNP-based methods. These findings confirm the hypothesis that analyzing the local genomic fragments around statistically significant mutations will improve the quality of breast cancer predictions compared to SNP-based methods.

This study also explores the impact of data-related hyperparameters on SPRiNGS' performance. While previous genome-wide prediction studies have demonstrated that including more SNPs for PRS calculations improves predictive accuracy (Antoniou, et al., 2018; Hajiloo, et al., 2013; Lee, et al., 2019; Cecile, et al., 2019), this research observed the inverse. Decreasing the number of SNPs elicited a slight decrease in AUC; however, increasing the number of SNPs noticeably lowered the model's performance. Since tightening the *p*-value threshold yielded fewer SNPs, SPRiNGS learned more influential semantic patterns across more statistically significant mutations. Inversely, since relaxing the threshold yielded more SNPs, less significant mutations likely contributed to noise around otherwise important semantic features during matrix resizing. Adding more mutations to consider caused the model to overfit, thereby achieving a lower AUC.

The other data-related characteristic explored was the sequence length *L*. As previously mentioned, SPRiNGS (which performed sequence-based classifications) outperformed SVM_RBF and 1D CNN (which performed SNP-based classifications), indicating that analyzing sequence contexts improves classification robustness and accuracy. The sensitivity analysis isolated the sequence length to measure its impact on model performance. SPRiNGS performed relatively consistently across all tested fragment lengths, but observed the highest AUC when $L = 37$ nucleotides. This sequence length likely captured significant sequence motifs which elicited better predictions. While decreasing *L* did not capture these semantic features, increasing *L* incorporated more information as noise when the matrices were resized, thereby leading to slightly lower performance. Although no significant correlation was observed between *L* and model performance, this study demonstrates the importance of analyzing the local DNA sequence contexts around statistically significant SNPs in genome-wide breast cancer prediction studies.

4.1 Limitations

This experiment contained plenty of limitations to consider. First, the Keras platform used to implement SPRiNGS prevents users from seeing the algorithm's inner workings, meaning that more research is required to understand which particular genomic features helped SPRiNGS outperform the SNP-based control models. Second, breast cancer has several molecular subtypes which this model did not account for; SPRiNGS simply determined whether a sample was healthy or at risk of breast cancer, but not which type of breast cancer. One method to overcome this limitation is by using a larger sample size with subgroups for each molecular subtype. Third, there are other genetic mutations associated with breast cancer besides SNPs. Including other mutation types (insertion, deletion, and genetic amplification) would provide a more holistic mutation-based prediction. Fourth, while the simulation in this study met the previous validation criteria, it does not account for genetic variations among global populations. This means sequence-based SNP interactions could fluctuate depending on geographic regions, causing changes in SNP statistical significance and allele frequencies. Exploring other simulations to understand how SPRiNGS' performance varies based on population-specific parameters would offer greater insights into the global variations of genetic breast cancer risks caused by dynamic sequence-based SNP interactions.

4.2 Future Work

The results of this study inspire further inquiry into the biological application of genomic sequence interactions in disease prediction. The model developed here, SPRiNGS, can be generalized to predict other common complex diseases, including cardiovascular, immune, and respiratory traits alongside other cancer types. Also, as mentioned earlier, SPRiNGS can be used to subtype various diseases based on their genomic and biological features. After analyzing the semantic patterns associated with copious genetic diseases, future research could explore the biological implications of these sequence-based interactions in protein binding, transcription factor motif identification, and gene regulation. In clinical settings, medical professionals can use SPRiNGS to recommend therapeutic action depending on a patient's risk classification for a certain disease. This study has identified a link between sequence contexts and breast cancer predictions; future research should explore how these semantic interactions impact an individual's biological susceptibility to this disease and others.

5. Conclusion

Ultimately, this research confirms the hypothesis that analyzing the local DNA sequence around statistically significant mutations will improve the quality of breast cancer risk classifications compared to SNP-based methods. SPRiNGS outperformed the SNP-based control models in both robustness and statistical accuracy. Moreover, an exploration of data-related characteristics revealed that including more SNPs decreased predictive accuracy (contrary to previous literature) and that the sequence fragment length has no significant correlation with model performance. While more research is required to understand the genomic features responsible for SPRiNGS' improved performance, this model holds exciting potential for precision medicine. Hopefully, as our knowledge of high-throughput sequencing and disease prediction mechanisms grow, so too will our ability to promote human health and longevity.

References

Adekitan, A. I. (2014). *Monte Carlo Simulation*. University of Idaban.

Antoniou, A. C., et al. (2019). Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *American Journal of Human Genetics*, *104*(1), 21-34. doi:10.1016/j.ajhg.2018.11.002

Beck, T., et al. (2020). GWAS Central: a comprehensive resource for the discovery and comparison of genotype and phenotype data from genome-wide association studies. *Nucleic Acids Research*, *48*(D1), D40-D933. Retrieved from https://doi.org/10.1093/nar/gkz895

Buniello, A., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, Vol. 47 (Database issue): D1005-D1012.

Cariaso, M., & Lennon, G. (2012). SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Research*, 40 (Database issue), D1308–D1312. https://doi.org/10.1093/nar/gkr798

CDC. (2020). *Breast Cancer Statistics*. https://www.cdc.gov/cancer/breast/statistics/index.htm

Cecile, A., et al. (2019). Polygenic risk scores that predict common diseases using millions of single nucleotide polymorphisms: Is More, better? *Clinical Chemistry, 65*(5), 609-611. doi:10.1373/clinchem.2018.296103

Cerami, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, *2*(5), 401-404. doi:10.1158/2159-8290.CD-12-0095

Choi, S. W., et al. (2020). Tutorial: A guide to performing polygenic risk score analyses. *Nature Protocols, 15*(9), 2759-2772. doi:10.1038/s41596-020-0353-1

Deng, L. (2012). The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning. *Technometrics*, *48*(1), 147-148. https://doi.org/10.1198/tech.2006.s353

Fadista, J., et al. (2016). The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur J Hum Genet*, 24, 1202–1205. https://doi.org/10.1038/ejhg.2015.269

Hajiloo, M., et al. (2013). Breast cancer prediction using genome wide single nucleotide polymorphism data. *BMC Bioinformatics*, 14, S3. https://doi.org/10.1186/1471-2105-14-S13-S3

Indolia, S., et al. (2018). Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach. *Procedia Computer Science*, *132*, 679-688. https://doi.org/10.1016/j.procs.2018.05.069

Kingma, D. P., & Ba, J. L. (Eds.). (2015). *Adam: A Method for Stochastic Optimization.* ICLR. https://arxiv.org/pdf/1412.6980.pdf

Kuleshov, V., et al. (2019). A machine-compiled database of genome-wide association studies. *Nat Commun*, 10, 3341. https://doi.org/10.1038/s41467-019-11026-x

Lee, A., et al. (2019). BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet Med*, 21, 1708–1718. https://doi.org/10.1038/s41436-018-0406-9

Quezada, H., et al. (2017). Omics-based biomarkers: Current status and potential use in the clinic. *Boletín Médico Del Hospital Infantil De México (English Edition), 74*(3), 219-226. doi:10.1016/j.bmhime.2017.11.030

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841-842. https://doi.org/10.1093/bioinformatics/btq033

Reference genome. (2020, Dec 25). Wikipedia. Retrieved January 27, 2021, from https://en.wikipedia.org/wiki/Reference_genome

Schneider, V. A., et al. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, *2017*(27), 849-864. doi:10.1101/gr.213611.116

Sherry, S. T., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1), 308-311. https://doi.org/10.1093/nar/29.1.308

Wen, J., et al. (2019). A classification model for lncRNA and mRNA based on k-mers and a convolutional neural network. *BMC Bioinformatics*, 20, 469. https://doi.org/10.1186/s12859-019-3039-3

Wu, T., et al. (2021). DeepDist: real-value inter-residue distance prediction with deep residual convolutional network. *BMC Bioinformatics*, 22, 30. https://doi.org/10.1186/s12859-021-03960-9

U.S. Breast Cancer Statistics. (2021, February 4). BREASTCANCER.ORG. Retrieved April 14, 2021, from https://www.breastcancer.org/symptoms/understand_bc/statistics#:~:text=About%201%20in%208%20U.S.,(in%20situ)%20breast%20cancer

Yamashita, R., et al. (2018). Convolutional neural networks: an overview and application in radiology. *Insights Imaging*, 9, 611–629. https://doi.org/10.1007/s13244-018-0639-9