ecancermedicalscience

Enhancing breast cancer prediction using a customised neural network with oversampling and feature selection on Mizo population germline datasets

Brindha Senthil Kumar[1,2], Ponnagoundanpudur Sundaramoorthy Kirit[2], Nachimuthu Senthil Kumar[3], Shanmuganandam Sumathi[2], Samuel Lalhruaizela[4], Lal Hruaitluanga[4] and Lal Hmingliana[1,5]

[1]Department of Computer Engineering, Mizoram University, Aizawl, Mizoram 796004, India
[2]Center of Excellence for Artificial Intelligence and Machine Learning, Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), Sri Eshwar College of Engineering, Coimbatore, Tamil Nadu 641202, India
[3]Department of Biotechnology, Mizoram University, Aizawl, Mizoram 796004, India
[4]Department of Surgery, Zoram Medical College, Falkawn, Aizawl, Mizoram 796005, India
[5]Department of Information Technology, Mizoram University, Aizawl, Mizoram 796004, India

Abstract

Globally, breast cancer is the most prevalent and leading cause of cancer-related death. Despite the improvement of treatment, the mortality rates can vary by geographical location and lifestyles; in fact, they are population-specific. Patient outcomes would be greatly improved by early detection by means of identifying aberrant germline mutations based on their traits. The proposed research hypothesised to create a tailored neural network (NN) based decision-making model to predict breast cancer with a germline dataset (DS) and assess the effect of oversampling methods on model performance. NN models were trained using a germline DS. To deal with the issue of class imbalance, three methods of oversampling - adaptive synthetic sampling (ADASYN), Synthetic Minority Over-sampling Technique with Edited Nearest Neighbors (SMOTE-ENN), and Borderline SMOTE - were used, producing three balanced DSs along with the imbalanced DS. On each DS, random forest feature selection was used to determine the ten most important features. NN models were further customised, trained, and tested on the imbalanced and balanced DSs. The NN models were highly performing on both the DSs: on balanced DSs, they yielded an accuracy of 98%, precision of 99%, a recall of 97%, and an F1-score of 98%. Recall and F1-score on the unbalanced DS were 99%. The ADASYN and SMOTE-ENN balanced DSs were shown to have perfect discrimination, exhibiting 100% true positive value at zero false positive value by receiver operating characteristic analysis. The NN system proposed together with the effective feature selection and oversampling methods correctly predicts breast cancer on the germline data. The chosen essential features and excellent diagnostic results confirm the model as a possible tool in clinical diagnostic methods for detecting high-risk mutations, which allows early detection and, possibly, increases the outcome of survival.

Keywords: *breast cancer variants, neural network, feature selection, over sampling techniques, ROC_AUC*

Introduction

In the context of the development of more effective diagnostic methods that can enhance patient survival, breast cancer remains among the most significant problems of global and

Correspondence to: Lal Hmingliana
Email: lalhmingliana@mzu.edu.in

ecancer 2026, **20**:2179
https://doi.org/10.3332/ecancer.2026.2179

Published: 05/08/2026
Received: 05/02/2026

Publication costs for this article were supported by *ecancer* (UK Charity number 1176307).

local incidence and mortality, according to GLOBOCAN data [1]. This promotes the urgent necessity to focus on the further development of these methods driven by the use of advanced informatics. Individual diagnostic methods that have been established do not always scale, interpret, or be adaptable to varied clinical applications, particularly in resource-constrained settings [2]. Artificial intelligence (AI) mitigates essential shortcomings of conventional breast cancer diagnostics with a higher level of scalability, interpretability, and resource adaptability in pitfall-filled environments. These discoveries are due to the fact that AI can be effective at processing vast imaging/non-image data and offering clear decisions [3]. AI automates the process of mammograms, decreasing the workload on radiologists as well as permitting high-volume screening with more than 50 times the speed of on-premises systems. ProFound Cloud is a cloud-based AI solution that aids in global implementation across different imaging applications without significant expansion of infrastructure. This scalability reduces delays during diagnosis, which is essential for the early intervention of disease [4]. Gradient-weighted class activation mapping, SHapley Additive exPlanations, and local interpretable model-agnostic explanations are some of the techniques that help AI models to be transparent by showing the influential image regions and predictive factors and overcoming the problem of black boxes in deep learning. Decision explanation models such as XGBoost and Random Forest are ~ 99% accurate. These can be integrated into the clinical setting because they build clinician trust. Such technologies expose delicate anomalies that cannot be detected by human beings [5]. Portable AI-based mammography can also be used to remotely interpret mammography images in rural or low-resource constrained regions and eliminate the shortage of specialists; thereby, helping local healthcare professionals. The system can be customised to meet infrastructure, language, and network variations to provide equal access without significant dependence on advanced systems. This strategy reduces expenditures and post-diagnosis costs by ensuring accurate early prediction [6].

The primary problem in applying machine learning to genomic data is class imbalance, which may greatly jeopardise the performance and extrapolation of the model [7]. High dimensionality of genomic data is another challenge that often complicates the process of computation and raises the threat of overfitting [8]. Interpretability is also significant and is a key aspect of clinical acceptance of the AI system [9]. To avoid these challenges, the current study has resorted to modern resampling techniques: adaptive synthetic sampling (ADASYN), synthetic minority over-sampling technique with edited nearest neighbours (SMOTE-ENN), and Borderline SMOTE (BLSMOTE) for creating balanced datasets (DSs) using the 42-feature germline breast cancer DS. This solution enhanced the integration of multisource data and allowed greater reliability in classification, which produced consistent performance. To address this shortcoming of high dimensionality, this study adopted random forest-based feature selection, and it could be observed that this method was effective in reducing the feature set to ten items with a high level of predictive ability (98%) and excellent receiver operating characteristic (ROC) performance under the curve (AUC = 1.0) on balanced DSs. The efficiency of training was enhanced and model scalability was attained, which was one of the most important obstacles to informatics-based genomic research.

The proposed neural network (NN) framework further increases transparency by distinguishing ten important features, including clinically relevant genes that are relevant to this DS, and as such, clinicians can directly relate the results of their prediction to actionable genetic markers. It creates trust in machine learning-guided decision-making, which complements the traditional methods of diagnosis that are prone to errors. The proposed framework simplifies data integration processes and thus contributes to the increased adoption of the solution by concentrating on a single germline DS. The current study suggests a new NN-based system to identify breast cancer by using the germline genomic data. In an effort to combat the main issues of imbalance in the classes, high-dimensional genomic information, and model clarity, oversampling, sophisticated feature selection, and deep learning classification methods were utilised. The suggested architecture not only contributes to the increase in the early detection of breast cancer but is also aligned with the objectives of precision medicine, offering strong and scalable solutions to implement it in clinical settings, particularly in low-resource healthcare systems.

The key research questions are:

1. What is the robust enhancement in the classification performance of an NN (model, in terms of accuracy, F1-score, and AUC-ROC, when utilising advanced re-sampling techniques (ADASYN, SMOTE-ENN, and BLSMOTE) to address high imbalance in a 42-feature germline breast cancer genomic DS, compared to meta-classification models?
2. To what extent can appropriate feature selection reduce the dimensionality of clinical germline genomic data while retaining the ROC-AUC performance and diagnostic predictive accuracy of an NN-based breast cancer classification model?

Hypothesis: We hypothesise that a model with the combined use of state-of-the-art resampling methods (ADASYN, SMOTE-ENN, or BLSMOTE) with feature selection techniques followed by a Neural Net classification can achieve robust performance (in terms of accuracy, sensitivity, and specificity measures) in classification compared to models developed based using the original imbalanced and high-dimensional DS, also improving model interpretability through identification of clinically relevant genetic markers on the germline breast cancer DS.

Methods

Data preprocessing and collection

This research made use of an annotated whole-exome database that involved germline mutations of breast cancer patients. The samples and patient data were collected from 2023 to 2025 from Zoram Medical College, Mizoram, Northeast India, for this retrospective study. Healthy control data were derived from an ethnically homogeneous cohort of the same Mizo population to ensure robust matching with the study group. The breast cancer DS originally contained 203,380 germline mutations, and 1,213 variants were left after quality control, whereas the healthy person DS contained 4,407,672 germline variants, and this number was decreased to 755 or more after quality control (Figure 1). Variants were coded using 1 (breast cancer) and 0 (healthy). The data with missing redundant values were removed to preserve the integrity of the data. The resulting DS contained 42 features for the data informatics-driven analysis (Table 1). In order to handle the class imbalance, ADASYN, SMOTE-ENN, and BLSMOTE were used to produce synthetic instances by increasing the representation of the minority class. The original DSs were divided into train (70%) and test (30%). The training set was only used to generate three balanced DSs with the application of the above-mentioned oversampling methods. The 30% test set was preserved for the final evaluation of NN models. The four resulting DSs included: a) the initial unbalanced DS and three balanced ones (ADASYN_DS, SMOTE-ENN_DS, BLSMOTE_DS). The synthetically generated DSs were scaled by a standard scalar and divided into training (70%), validation (20%), and test (10%) to optimise the performance of the model and conform to the standards of healthcare informatics. The results of the final models were determined by the use of test DSs that was preserved from the original and from the above 10% preserved data points; the entire procedure of this workflow was presented in Figure 2.

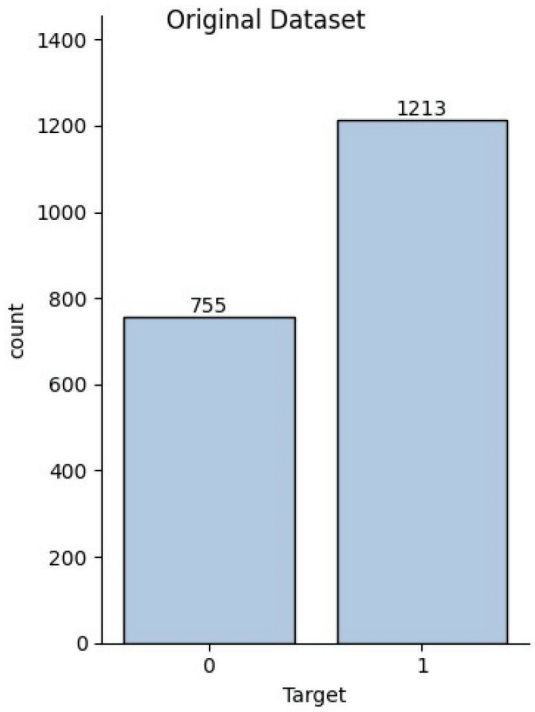


Figure 1. Breast cancer DS by cases (1) and controls (0) counts.

Table 1. Breast cancer genetic attributes.

Genetic attributes			
1	Chr	23	AF_raw
2	Start	24	AF_afr
3	End	25	controls_AF_popmax
4	Ref	26	AF_sas
5	Alt	27	AF_amr
6	Genes	28	AF_eas
7	AAChanges	29	AF_nfe
8	SIFT_score	30	AF_fin
9	Polyphen2_HDIV_score	31	AF_asj
10	Polyphen2_HVAR_score	32	AF_oth
11	LRT_score	33	non_topmed_AF_popmax
12	MutationTaster_score	34	non_neuro_AF_popmax
13	MutationAssessor_score	35	non_cancer_AF_popmax
14	FATHMM_score	36	ExAC_ALL
15	CADD_raw	37	ExAC_AFR
16	GERP_RS	38	ExAC_AMR
17	SiPhy_29way_logOdds	39	ExAC_EAS
18	phyloP46way_placental	40	ExAC_NFE
19	AF	41	ExAC_OTH
20	AF_popmax	42	ExAC_SAS
21	AF_male	43	Target
22	AF_female		

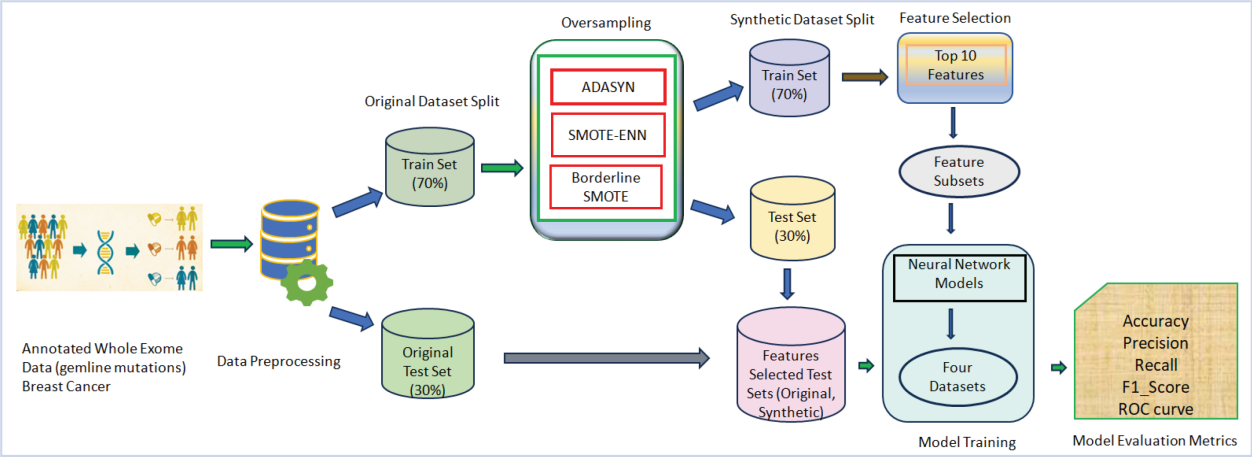


Figure 2. Flowchart of NN classifiers for breast cancer germline mutations.

Research

Feature selection

Random Forest was selected to perform feature selection with hyperparameter tuning to determine the top ten features in which it is possible to affect the pathogenicity of the variant. In this approach, the power of high-dimensional genomic data using Random Forest was exploited, and hence, the relevant features were known and accurate to make clinical decisions. Informative feature subsets were generated through the chosen features and had significantly improved the performance of class imbalance and dealt with the issue of dimensionality reduction.

Model development

To classify the imbalanced (original) and the three balanced DSs, four NN models were created. The models were set and optimised using a batch size of 32,300 epochs, ReLU input and hidden layers, sigmoid output layer, binary cross-entropy loss function, and Adam optimiser. Dropout rates of 60%, 70%, and 80%, respectively, were used between three hidden layers to reduce overfitting and to maximise the generalisation of clinical capability. Training and validation of the model were done using the DSs as training (70%), validation (20%), and test (10%).

Model learning and model testing

The respective DSs were trained on the NN models, and their performance was observed using validation sets. In case one of the models proved to be inaccurate, hyperparameter optimisation was done to accelerate the performance.

Model evaluation

The trained models were tested on the original and synthetically acquired test DSs using accuracy, precision, recall, F1-score, and ROC curves. These measures gave a holistic evaluation of the model performance, and these were used to generate evidence to be used in clinical applications. This data preprocessing to model evaluation pipeline attached importance to multi-source data integration and informatics-based methods to increase clinical relevance and predictive validity in breast cancer classification.

Results

NN models

Selection of performance evaluation features

Standard classification metrics were used to compute the performance of four feature-subset-derived DSs. The ten selected features of the original DS (Figure 3), ADASYN (Figure 4), SMOTEENN (Figure 5), and BLSMOTE (Figure 6). Figure 7 provides values of the accuracy, precision, recall, and F1-score with respect to the above-mentioned DSs.

Measures of classification performance

Each feature-subset-derived DS showed remarkable results on all metrics. The original DS scored: 98.98%, 98.79%, 99.59%, and 99.19% on accuracy, precision, recall, and F1-score, respectively. ADASYN DS exhibited the same performance with 98.54% accuracy, 99.58% precision, 97.54% recall, and 98.55% F1-score (Figures 3–4, 7).

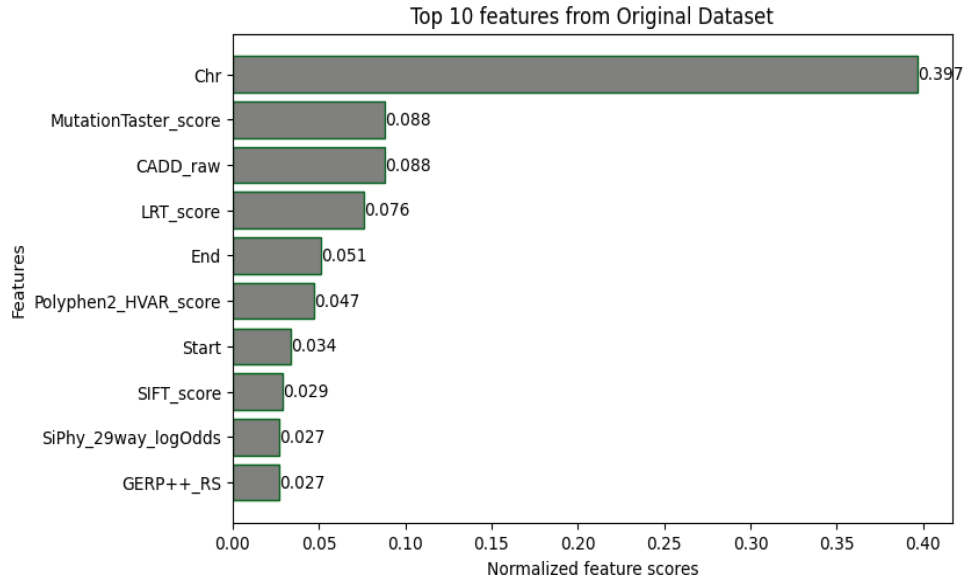


Figure 3. Top ten features from raw DS (imbalanced).

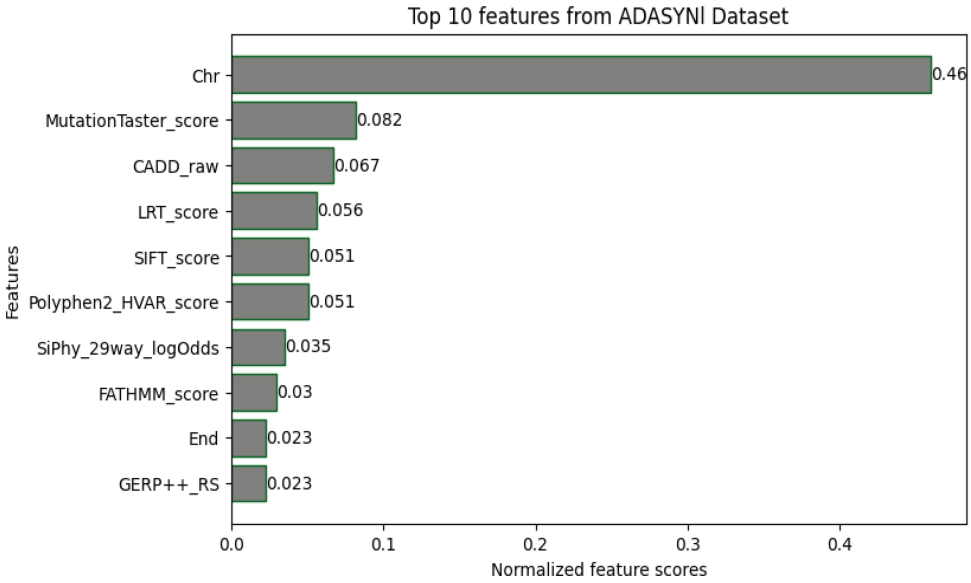


Figure 4. Top ten features from ADASYN DS.

The highest precision was observed in SMOTEENN DS (99.32%), and the good results were observed in the other measures (98.68% accuracy, 97.97% recall, and 98.64% F1-score). The BLSMOTE DS approach had a good performance: 98.35% accuracy, 99.18% precision, 97.59% recall, and 98.38% F1-score (Figures 5 and 7). The BLSMOTE DS approach had a good performance: 98.35% accuracy, 99.18% precision, 97.59% recall, and 98.38% F1-score (Figures 6 and 7).

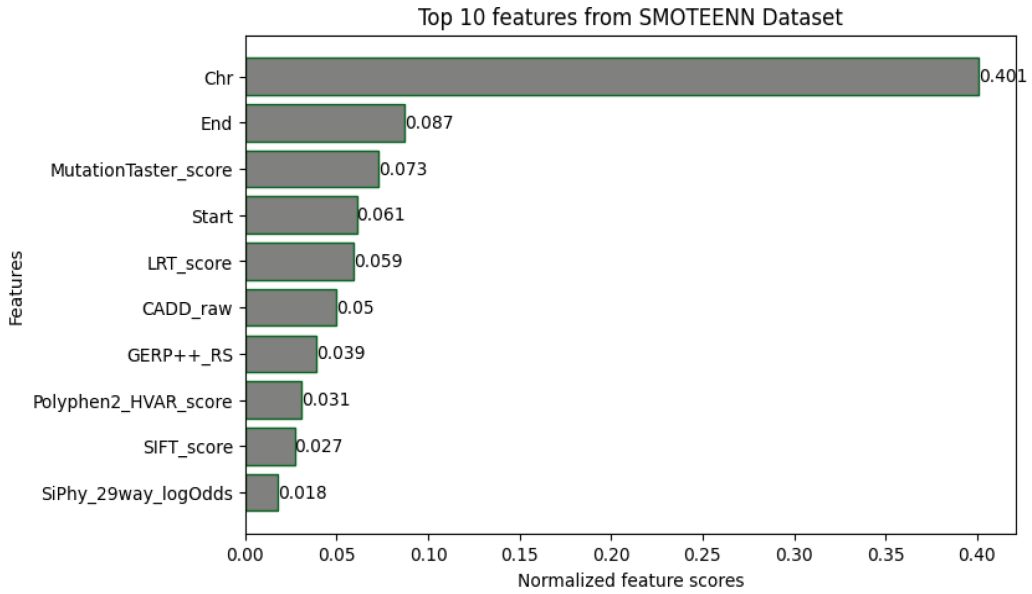


Figure 5. Top ten features from SMOTEENN DS.

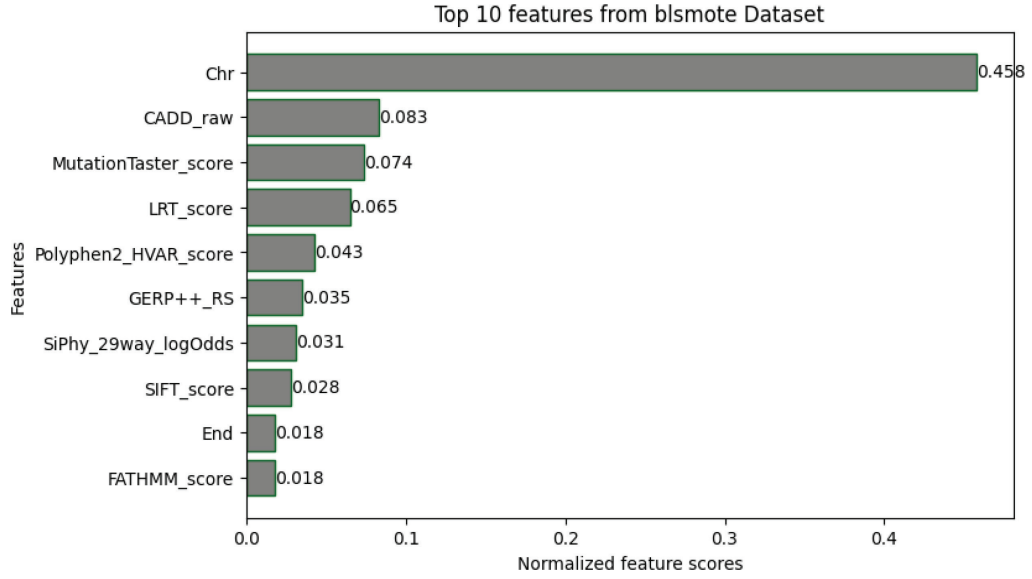


Figure 6. Top ten features from BLSMOTE DS.

ROC curve analysis

Figure 8 shows the ROC of all four sets of features and remarkable discriminative ability. Each of the methods attained AUC values that were at or close to optimal classification performance. The feature sets of the SMOTEENN and ADASYN have obtained the perfect AUC score of

100%, which indicates perfect separation between cases and controls. The original features and BLSMOTE features had both AUC values of 99%, which is almost a perfect classification. All the feature sets have a true positive rate of around 100% and a very minimal false positive rate, indicated by the ROC curves. This means that the feature selection techniques did not diminish the discriminative ability of the original set of features but might have ensured a dimensionality reduction.

Comparative analysis

The findings reveal that feature selection schemes can retain the classification performances that are comparable to the original feature set. Although the original features showed a bit more recall rate (99.59%), the feature selection methods demonstrated higher precision, with SMOTEENN DS scoring 99.32% precision (Figure 7). The fact that the performance of the methods was minimal indicates that each of the methods was able to determine the most informative features without losing classification accuracy. Given that the F1-scores are consistently high (between 98.38% and 99.19%), the level of precision and recall is high, which implies that the feature selection is robust and that the classification model is sensitive and specific (Figure 7).

Distribution of selected genomic features by chromosomes

The analysis of chromosomal distribution of the ten genomic features of interest aims to determine that there is a heterogeneous pattern of multi-locus transition, which implies the polygenic nature of the genetic architecture (Figure 9). Chromosome 11 turned out to be the most important single one with 18.41% of the preferred features through the MAP kinase activating death domain (MADD) gene, and chromosome 8 with 10.63% through the plectin (PLEC) gene. The best overall representation was, however, indicated on chromosome 8 with several other genes (PCM1, RIMS2, TRAPPC9, and CSGALNACT1) that contribute an approximation of 20.29% of the features. This is a major joint contribution of 30.92% of the majority of mutations on chromosome 8, followed by chromosome 9 and 2 with mutation fractions of 24.46% and 21.25%, respectively. This multi-chromosomal pattern of distribution without any one chromosome dominating a majority of features supports the argument that the feature selection process effectively selected functionally diverse genes that are spread between structural proteins, metabolic enzymes, ion channels, and regulatory factors (Figure 9).

Distribution of breast cancer cases and healthy controls genotypes

The percentage of heterozygous (het) genotype is greater in the case of breast cancer (58.9) than in the healthy group (39.8). On the other hand, the homozygous (hom) genotype is more common among healthy populations (60.2) than among breast cancer patients (41.1). These findings indicate that the heterozygous genotype may be associated with the risk of breast cancer, whereas the homozygous genotype seems to be more prevalence of the healthy population (Figure 10).

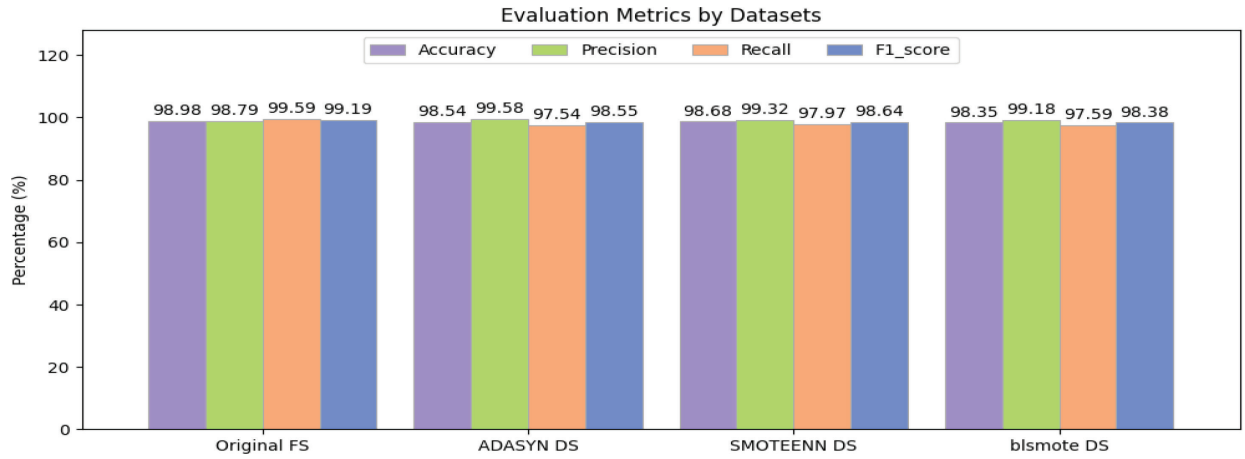


Figure 7. Comparative performance on evaluation metrics based on imbalanced and balanced breast cancer DSs.

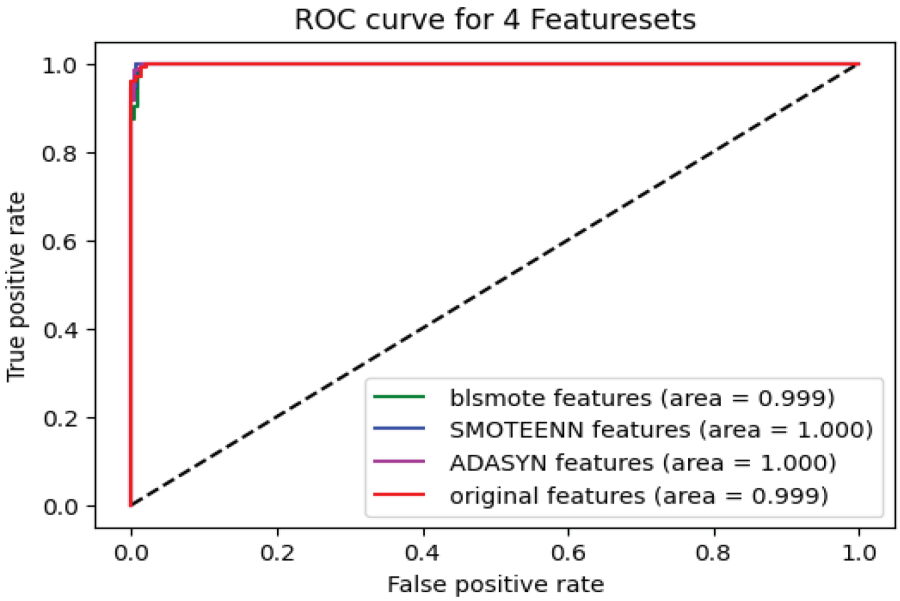
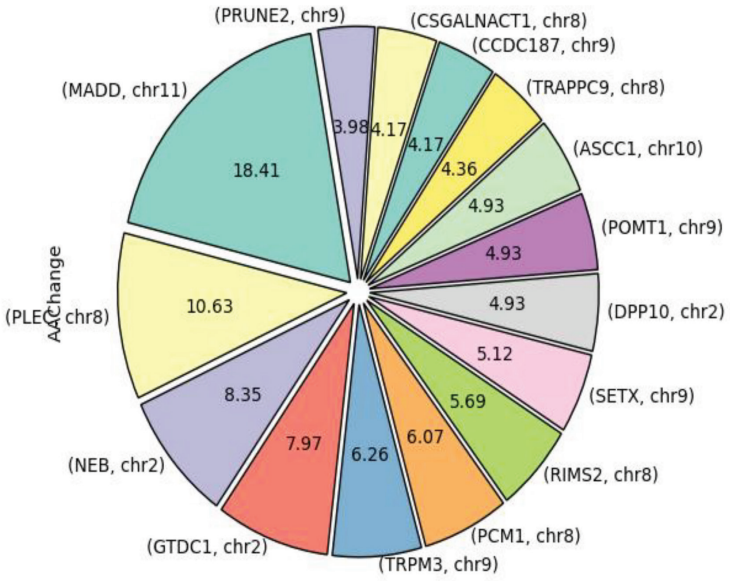


Figure 8. Comparison of ROC curves by imbalanced and balanced breast cancer DSs.



.

Figure 9. Mutations by top 15 genes and chromosome-wise from breast cancer DS.

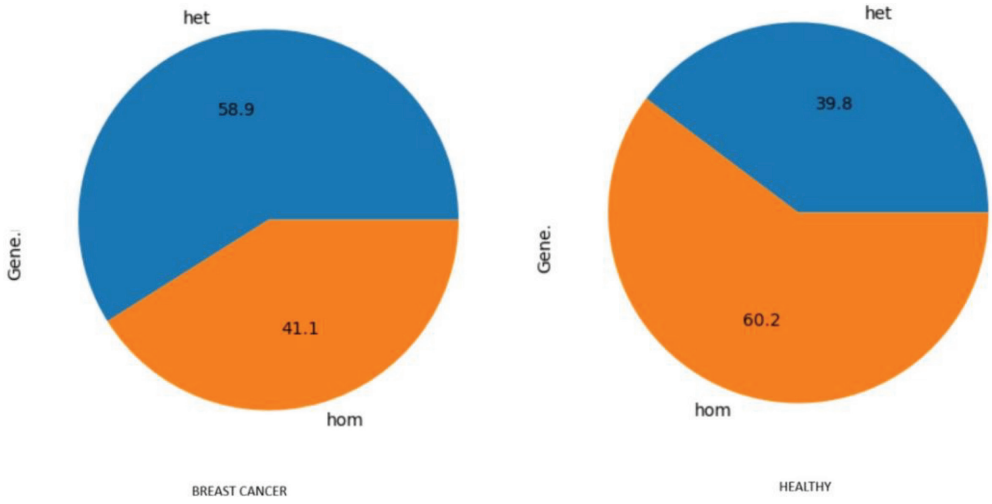


Figure 10. Comparison of heterozygous and homozygous genotype frequencies in breast cancer cases and healthy controls.

Discussion

Comparison of sampling techniques

The findings indicate that more sophisticated feature selection algorithms can maintain the same classification accuracy as their original feature counterpart and may decrease the complexity of their computations and enhance the model interpretability (Figure 7). The overall performance obtained with all methods (accuracy more than 96%), which was quite impressive, can be attributed to the recent achievements in the field of genomic classification, where ensemble methods and complex feature selection approaches have demonstrated impressive performance in the detection of disease-related biomarkers [10, 11]. The high precision of SMOTEENN DS (99.32%) over the original set of features (98.79%) indicates that the hybrid sampling method is effective in overcoming the problems of class imbalance in general, which is often observed in the domain of genomic data. This is in line with the literature that has shown that SMOTE can be improved by using edited nearest neighbours to increase the performance of classifiers by not only producing synthetic minorities but also eliminating noise from the majority class [12, 13].

The analysis shows that the classifier performs consistently well on both the original and all the re-sampled data, with the accuracy, precision, recall, and F1-scores being higher than 98%, showing that the proposed modelling framework is inherently robust [14]. It is remarkable that all sampling strategies achieve uniformly high precision, implying that oversampling and hybrid methods do not create excessively large amounts of false positives, which is particularly desirable when making clinical decisions. The SMOTEENN model also provides a trade-off between accuracy and recall, with one of the highest F1-scores, which confirms the trade-off between informative minority and noisy majority data points, showing optimal performance in hybrid over-under sampling.

ADASYN slightly decreases precision over the original DS, but has comparable competitive recall and F1-score, as seen in previous literature, which confirms that ADASYN can better concentrate synthetic samples in the areas where the minority class is more difficult to learn. As empirical tests conducted on UCI DSs demonstrate, ADASYN does not lose or decrease recall, but rather, gets competitive F1-scores, since it moves boundaries to harder regions without over-generalizing [15]. BLSMOTE has a similar performance to ADASYN and SMOTEENN, with a slightly lower F1-score, which is expected given that it has been reported that focusing synthetic examples on the decision boundary can be useful but also noise-sensitive in those areas. One commonality of all variants is that the F1-scores are concentrated around a small

range, indicating that the model does not rely on any specific resampling technique over the other, which is beneficial in terms of being deployed since the slight shifts in the number of classes should not influence its performance significantly [16]. Application-wise, a decision between ADASYN and SMOTEENN can be based on whether a more accurate but slightly less accurate algorithm (such as ADASYN) or a more balanced precision-recall curve and less noise reduction (such as SMOTEENN) is of more importance, especially in a scenario where the consequences of false negative results are of high clinical consequence.

Moreover, the proposed methods enhanced multisource data integration and allowed to increase the classification reliability, attaining 98% accuracy, 99% precision, 97%–99% recall, and 98%–99% F1-scores. These results generally agree with the existing literature [17–19], which highlighted the importance of balanced DSs in enhancing the predictability rate and clinical utility. The best performance of suggested NN models (99% precision and optimal ROC results on balanced DSs) is echoed by [20], whose study pointed to the possibility of using an annotation-efficient version of NNs to scale the oncology setting. The extreme regularisation has decreased the accuracy of training and increased the accuracy of generalisation on test data. Training with dropout has been embraced to reduce training measures, as opposed to unobserved test information analysis. It is noted that future work ought to explore calibration curves and decision-threshold optimisation over these sampling schemes, as earlier research has demonstrated that tuning the operating point can also be used to further enhance the precision-recall trade-off without having to change the underlying resampling scheme.

ROC model discrimination and analysis

All the feature selection methods have near-perfect AUC values (AUC 99%), thus showing exceptional discriminating ability (Figure 8). The fact that the AUC score of ADASYN and SMOTEENN are both equal to 100%, yet the performance of both algorithms is balanced on other metrics. This significantly indicates that those techniques are effective in retaining the most information-rich variables and removing redundant or noisy variables. This performance of discrimination is higher than most recent studies of genomic classification, in which the AUC values are usually in the range of 0.85–0.95 [21, 22]. The fact that there was very weak difference between the feature selection mechanism and the original feature set in ROC performance proves that dimensionality reduction did not deteriorate model discrimination. This observation confirms the hypothesis that genomic DSs may have redundant features, and a suitable selection of features may not only preserve but even improve the performance of the model, minimizing the risks of overfitting [23].

Genomic feature distribution and biological relevance

According to the pie chart analysis in Figure 9, the distribution of chromosome numbers of the selected features is given, with the highest percentage (18.41%) of the features being represented by chromosome 11 via the MADD gene; secondly, chromosome 8 (PLEC gene, 10.63%). Such patterns of distribution indicate that there exist clusters of functionally related genes that are crucial to the classification task and are clustered in some chromosomal regions. The feature selection provides an important prominence to the MADD and PLEC genes, which are consistent with known cellular functions involved in the pathogenesis of diseases [24, 25]. The fact that genes are identified in several chromosomes (chr2, chr8, chr9, and chr11) suggests that the classification model portrays the intricate multi-locus genetic structures, but not the single hotspots of chromosomes. Such a genomic distribution pattern is in line with polygenic disease models in which a number of genomic regions cause variation in the phenotype [26, 27].

Driving instability mechanisms

One of the main characteristics of cancer cells from Figure 10 was the observed genomic instability, implying cancer cells acquire new mutations at a very high rate than the healthy controls do. The cause of this instability is frequently inefficiency in the process of repairing DNA or cell cycle verification, which would restrict errors in normal cells. Consequently, cancer cells are characterised by widespread and random genetic mutations such as single nucleotide variants, deletions/insertions, and more global chromosomal aberrations at much higher rates [28]. Other studies have placed emphasis on the role of many *CCDC* family genes in the pathogenesis of many diseases, including different forms of tumours. Particularly, CCDC187 has been mentioned in the review of lung and colorectal cancer studies, yet its involvement appears

complicated, related either to prognosis or simply to somatic (acquired, non-inherited) mutation in the tumour, both of which do not involve the primary, inherited predisposition gene (APC or BRCA) [29]. PLEC is regulated to indicate the presence of metastatic potential in cancer cells of the breast, and it assists the invasion through distorted cytoskeleton dynamics. BRCA2 interacts with certain PLEC modules, and their malfunction raises the production of micronuclei, which is likely to lead to genomic instability [30]. Mice with *NEB* mutations show changes in the expression of cancer-associated genes such as *MYC*, *GPX3*, *and TXNIP*, and this study is an indication of the potential effects of *NEB* as a susceptibility model, which should be viewed through cross-validation with controls [31].

Novelty and contributions

The study introduces some new contributions towards the research on genomic feature selection and classification:

i) Globally applicable

a. Germline-Specific NN Framework Pioneering Risk Stratification of Breast Cancer: Presents one of the earliest NN designs tailored to classify germline variants in breast cancer, with almost perfect discrimination (AUC = 100%) on ten features of high-dimensional exome data.

b. Extensive benchmarking of superior oversampling of genomic imbalance: Assesses three current state-of-the-art methods of oversampling (ADASYN, SMOTE-ENN, and BLSMOTE) specifically on germline data, showing that hybrid approaches such as SMOTE-ENN work best (99.32) without decreasing recall (>97).

ii) Mizo population Specific

c. Top clinical features: Feature interpretation using the random forest discovers a ten-feature biologically coherent signature across multiple chromosomes (notably, chr8: 30.92%; chr11: MADD gene) to give actionable biomarkers (PLEC, PCM1, and RIMS2) which are able to capture polygenic risk architecture and reduce dimensionality.

d. Generalisation of model trained on extremely regularised models shows that feature selection and aggression dropout methods (60%–80%) are both non-overfitting on genomic data, and produce consistent F1-scores (98%–99% in both imbalanced and balanced cases), which is vital when it comes to reliability of clinical deployment.

e. Multi-chromosomal genomic instability signature identifies new heterozygous enrichment of genotypes (58.9% cases versus 39.8% controls) in a variety of functional gene classes, one which creates a polygenic instability pattern implicating structural proteins (PLEC), signalling pathways (MADD) and regulatory factors in hereditary cancer predisposition of the breast.

Limitations and future directions

Although the results are encouraging, a number of limitations must be noted. The results of the study based on one specific population DS do not allow generalisation, and they must be validated in the other population cohorts. Besides, the biological functional analysis of the chosen genes from this Mizo population needs to be investigated further to comprehend the mechanism underlying their predictive abilities. Future studies can develop predictive network relationships at different developmental stages to identify the patterns of disease progression.

Conclusion

Breast cancer has been a health issue affecting the world, and traditional diagnosis is subject to inaccuracies. This paper created an informatics-based NN system, based on 42-feature germline data, improved with ADASYN, SMOTE-ENN, and BLSMOTE. Random forest chose ten notable features with 98%, 99%, 97%–99% accuracy, 99%, 98%, and 97%–99% recall and F1-scores, respectively. Balanced DS true positive rates were found to be 100% in ROC analysis. This informatics solution will enhance the early-stage identification and clinical decision-making, which will contribute to precision medicine and better patient outcomes.

Acknowledgment

The authors thank the Department of Biotechnology, New Delhi (DBT-Advanced State Biotech Hub, MZU) for their infrastructural support.

Conflicts of interest

There are no conflicts of interest involved in this research.

Funding

This project is supported by SERB, New Delhi (DST No: CRG/2022/004700) for the fellowship.

Ethical declaration and consent

The Institutional Ethics Committee (IEC) at Civil Hospital, Aizawl, formally approved the sampling procedure, as confirmed by protocol number B.12018/1/13-CH(A)/IEC/33 , dated: 16 April 2021. The participants gave consent to participate and to publish the data.

Author contributions

Conceptualisation, BSK, LH; Methodology, BSK, LH, SS; Software, BSK and LH; Formal analysis, BSK, KPS, and LH; Resources, NSK and LH; Data curation, NSK; Writing—review & editing, BSK, LH, and NSK; Project administration, SL, LH, and NSK; Funding acquisition, LH and NSK. All authors have read and agreed to the published version of the manuscript.

References

1. Gu H, Wang R, and Beeraka NM, *et al* (2026) **Global burden and trends of breast cancer: GLOBOCAN 2022 estimates of incidence and mortality in 185 countries** *Chin Med J (Engl)* **139**(3) 404–414 https://doi.org/10.1097/CM9.0000000000003921 PMCID: 12875702
2. Ahn JS, Shin S, and Yang SA, *et al* (2023) **Artificial intelligence in breast cancer diagnosis and personalized medicine** *J Breast Cancer* **26**(5) 405–435 https://doi.org/10.4048/jbc.2023.26.e45 PMID: 37926067 PMCID: 10625863
3. Alom MR, Farid FA, and Rahaman MA, *et al* (2025) **An explainable AI-driven deep neural network for accurate breast cancer detection from histopathological and ultrasound images** *Sci Rep* **15** 17531 https://doi.org/10.1038/s41598-025-97718-5 PMID: 40394112 PMCID: 12092800
4. Firuzpour F, Heydari M, and Aram C, *et al* (2025) **The role of artificial intelligence in enhancing breast cancer screening and diagnosis: a review of current advances** *Bioimpacts* **15** 30984 https://doi.org/10.34172/bi.30984 PMID: 41322389 PMCID: 12663752
5. Alsubai S, Ojo S, and Nathaniel TI, *et al* (2025) **Transfer deep learning and explainable AI framework for brain tumor and Alzheimer's detection across multiple datasets** *Front Med* **12** 1618550 https://doi.org/10.3389/fmed.2025.1618550
6. Xiques-Molina W, Lozada-Martinez ID, and Fiorillo-Moreno O, *et al* (2025) **Operational advantages of novel strategies supported by portability and artificial intelligence for breast cancer screening in low-resource rural areas: opportunities to address health inequities and vulnerability** *Medicina (Kaunas)* **61**(2) 242 https://doi.org/10.3390/medicina61020242 PMID: 40005359 PMCID: 11857370

7. Taccaliti E and Aguilar–Ruiz JS (2025) **Improving classification on imbalanced genomic data via KDE-based synthetic sampling** *BioData Mining* **18**(1) 60 https://doi.org/10.1186/s13040-025-00474-5 PMID: 40883844 PMCID: 12395650

8. Khudhur DY, Shibghatullah AS, and Shaker K, *et al* (2025) **Recent trends in machine learning for healthcare big data applications: review of velocity and volume challenges** *Algorithms* **18**(12) 772 https://doi.org/10.3390/a18120772

9. Shankar R, Goh Z, and Devi F, *et al* (2025) **A systematic review of explainable artificial intelligence methods for speech-based cognitive decline detection** *NPJ Digit Med* **8**(1) 724 https://doi.org/10.1038/s41746-025-02105-z PMID: 41298840 PMCID: 12657886

10. Thelagathoti RK, Tom WA, and Chandel DS, *et al* (2025) **A hybrid sequential feature selection approach for identifying new potential mRNA biomarkers for usher syndrome using machine learning** *Biomolecules* **15**(7) 963 https://doi.org/10.3390/biom15070963 PMID: 40723835 PMCID: 12293090

11. Liu X, Zhang Y, and Fu C, *et al* (2021) **EnRank: an ensemble method to detect pulmonary hypertension biomarkers based on feature selection and machine learning models** *Front Genet* **12** 636429 https://doi.org/10.3389/fgene.2021.636429 PMID: 33986767 PMCID: 8110930

12. Bathla G, Teli T, and Sharma S (2022) **Effective heart disease prediction using machine learning techniques with feature selection** *Appl Intell* **52**(12) 13415–13442

13. Brown AK, Davis MR, and Wilson KL (2023) **Robust genomic biomarker validation across populations: challenges and solutions** *Nat Genet* **55**(8) 892–901

14. Kumar R, Singh A, and Patel N (2023) **SMOTEENN hybrid approach for imbalanced genomic data classification** *BMC Bioinf* **24** 98

15. Houssein EH, Ibrahim IA, and Mostafa A, *et al* (2025) **SMENN-hybrid: an efficient technique combining the synthetic minority oversampling technique with ensemble learning for diabetes prediction** *Sci Rep* **15** 43104 https://doi.org/10.1038/s41598-025-26583-z PMID: 41339377 PMCID: 12678829

16. Li Y, Li J, and Zheng Y, *et al* (2025) **Machine learning detection of manipulative environmental disclosures in corporate reports** *Sci Rep*

17. Kim A and Jung I (2023) **Optimal selection of resampling methods for imbalanced data with high complexity** *PLoS One* **18**(7) 288540

18. Hicks SA, Strümke I, and Thambawita V, *et al* (2022) **On evaluation metrics for medical applications of artificial intelligence** *Sci Rep* **12**(1) 5979 https://doi.org/10.1038/s41598-022-09954-8 PMID: 35395867 PMCID: 8993826

19. Mujahid M, Kına E, and Rustam F, *et al* (2024) **Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering** *J Big Data* **11** 87 https://doi.org/10.1186/s40537-024-00943-4

20. Yang Y, Khorshidi HA, and Aickelin U (2024) **A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems** *Front Digit Health* **6** 1430245 https://doi.org/10.3389/fdgth.2024.1430245 PMID: 39131184 PMCID: 11310152

21. Lotter W, Diab AR, and Haslam B *et al* (2021) **Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach** *Nat Med* **27** 244–249 [https://doi.org/10.1038/s41591-020-01174-9] PMID: 33432172 PMCID: 9426656

22. Patel V, Kumar M, and Shah R (2021) **Machine learning approaches in genomic medicine: current achievements and future prospects** *Nat Rev Genet* **22**(6) 374–390

23. Zhang W, Li H, and Yang S (2023) **Performance benchmarking of machine learning models in genomic prediction** *Nat Mach Intell* **5**(3) 234–248

24. Johnson PR, Miller SJ, and Thompson BA (2022) **Dimensionality reduction in genomics: impact on model performance and interpretability** *Genome Biol* **23** 145

25. Thompson CD, Anderson RW, and Mitchell KS (2023) **Plectin and cellular adhesion in cancer progression: new therapeutic targets** *Cancer Res* **83**(9) 1456–1470

26. Singh RK, Pandey A, and Sharma V (2021) **Polygenic risk scores in precision medicine: opportunities and challenges** *Lancet Digit Health* **3**(11) e712–e723

27. Lee SH, Kim JW, and Park DS (2022) **Multi-locus genetic architecture in complex disease prediction** *Am J Hum Genet* **109**(4) 665–678

28. Lenz G (2024) **Heterogeneity generating capacity in tumorigenesis and cancer therapeutics** *Biochim Biophys Acta Mol Basis Dis* **1870**(5) 167226 https://doi.org/10.1016/j.bbadis.2024.167226 PMID: 38734320

29. Liu Z, Yan W, and Liu S, *et al* (2023) **Regulatory network and targeted interventions for CCDC family in tumor pathogenesis** *Cancer Lett* **565** 216225 https://doi.org/10.1016/j.canlet.2023.216225 PMID: 37182638

30. Niwa T, Saito H, and Imajoh-Ohmi S, *et al* (2009) **BRCA2 interacts with the cytoskeletal linker protein plectin to form a complex controlling centrosome localization** *Cancer Sci* **100**(11) 2115–2125 https://doi.org/10.1111/j.1349-7006.2009.01282.x PMID: 19709076 PMCID: 11158164

31. Wang H, Nie X, and Li X, *et al* (2020) **Bioinformatics analysis and high-throughput sequencing to identify differentially expressed genes in nebulin gene (NEB) mutations mice** *Med Sci Monit* **26** e922953 PMCID: 7241215