The EuroBiotech Journal
RESEARCH ARTICLE
Bioinformatics

Using Convolutional Neural Networks with Image-Based Representations of Amino Acid Sequences for Predicting the Effects of Genetic Variants

Gülbahar Merve ŞILBIR[1,*], Burçin KURT[2]

Abstract

Proteins are one of the fundamental molecules that regulate cellular processes in living organisms. Given the pivotal role played by protein-protein, DNA-protein, and RNA-protein interactions in a significant proportion of biological processes, variants occurring in the regions where these interactions occur have the potential to give rise to serious consequences for the phenotype. Various supervised learning techniques are employed to ascertain the correlation between protein variants and the development of a specific disease. In this study, a convolutional neural network-based prediction model is proposed to predict the pathogenicity effect of variants on the phenotype. This is achieved by converting amino acid sequences into two-dimensional images. A protein embedding method utilizing transfer learning (TAPE) was employed to generate the feature vector. The feature vector was transformed into a square-shaped, single-channel image and trained with a deep learning algorithm comprising a convolutional neural network. This study performed a binary classification (benign versus pathogenic) using missense variants in the BRCA1 protein obtained from the open-access ClinVar database as the dataset. The findings demonstrate that the developed prediction model is highly effective in predicting the pathogenicity effects of variants within the functional regions of the BRCA1 protein on phenotype. The evaluation of the model's prediction results demonstrated that variants in the benign class can be classified with 91% accuracy (93% sensitivity). Furthermore, the model demonstrated robust performance in classifying both benign and pathogenic variants, with an AUC value of 92%. These findings suggest that the developed prediction model may offer potential in classifying BRCA1 variants and assessing their potential pathogenicity. The variant effect prediction model obtained in this study shows promise and may benefit from further refinement in future research.

Keywords: Protein Representation, Convolutional Neural Network, Variant Effect Prediction, Transfer Learning

Introduction

In the human genome, 99.5% of the DNA is common to the population. The remaining 0.5% is defined as genetic variants and is known as the differences that make each human genome unique [1]. These DNA differences, about 0.5%, are linked to various traits such as skin color, height, vision, and hearing in individuals [2]. The majority of genetic variants are single nucleotide variants (SNVs). SNVs are defined as single amino acid variants (SAVs) at the protein level. Since protein-protein, DNA-protein, and RNA-protein interactions play a crucial role in many biological processes, variants occurring in these interaction regions can have serious consequences on the phenotype [3-6]. These variants result from single base changes that alter the amino acid sequence of the encoded protein. They can lead to disease by impacting protein stability, interaction, and enzyme activity [7].

The BRCA1 gene, known as Breast Cancer gene 1, is known as a tumor suppressor gene. Epidemiological studies estimate that germline mutations in BRCA1 account for approximately 5-10% of all breast cancer cases and 15-20% of ovarian cancer cases in women, making it one of the most clinically significant tumor suppressor genes [8]. This

[1] Department of Electricity and Automation/Robotics and Artificial Intelligence, Çarşıbaşı Vocational School, Trabzon University, Trabzon, Turkey
[2] Department of Biostatistics and Medical Informatics, Karadeniz Technical University, Trabzon, Turkey

***Corresponding author:**
Gülbahar Merve ŞILBIR
E-mail: gmervecakmak@trabzon.edu.tr

DOI: 10.2478/ebtj-2025-0020

gene encodes the BRCA1 protein, which contains multiple domains and is crucial in hereditary breast and ovarian cancer [8]. The BRCA1 protein plays a critical role in ensuring the integrity of the genome [9, 10]. The BRCA1 functions as a defense mechanism in the cellular response to DNA damage, DNA repair, and controlling various cellular processes such as transcription, aging, and apoptosis [8, 11, 12]. It also interacts with various proteins that are crucial for gene regulation and embryonic development [13]. Mutations in three different domains of BRCA1 are implicated in breast and ovarian cancers: the N-terminal RING domain, the coiled-coil CC domain, and the C-terminal BRCT domains [9]. The domain structure of BRCA1 is illustrated in Figure 1.

It is crucial to evaluate the impact of variants found in these functional domains and regions in BRCA1 on the human phenotype. In the literature, variants are analyzed using experimental methods to determine whether they are pathogenic or benign [14]. Nowadays, it is possible to analyze variants using computational and machine learning methods [14]. Rather than relying on expensive and time-consuming experimental methods, analyzing protein sequences with machine learning techniques can yield highly effective solutions. Prediction models can be developed using deep learning and natural language processing methods to learn the structure of the protein and amino acid sequences and describe their characteristics. The literature shows that prediction models based on the sequence

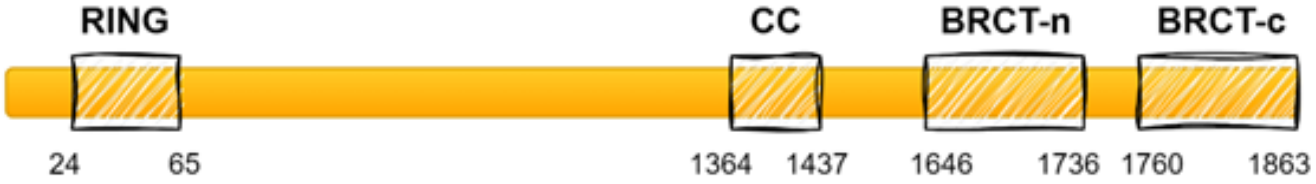


Figure 1. BRCA1 protein. Human BRCA1 contains an N-terminal RING domain (24-65aa), coiled coil domain (1364–1437aa) and two C-terminal BRCT repeats (1646-1863aa).

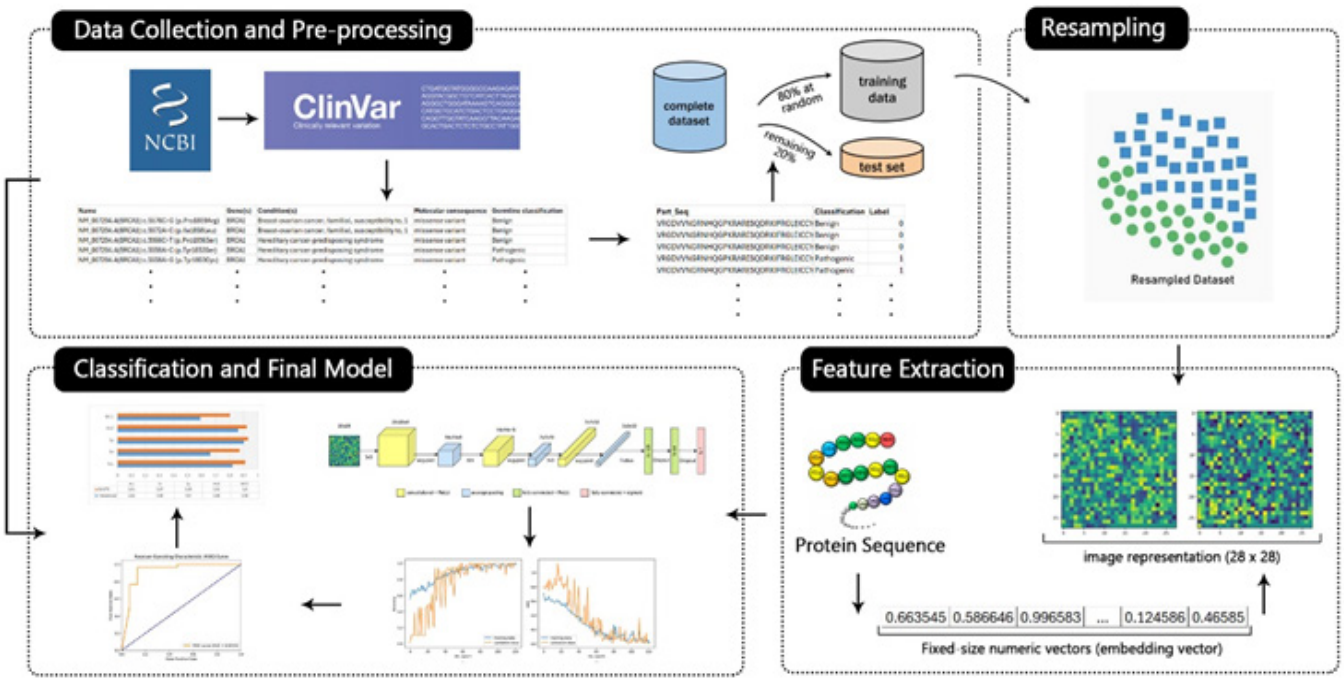


Figure 2. Schema of the study

of the protein are developed without using any structural or evolutionary information [23].

In this study, a model based on convolutional neural network is proposed to predict the pathogenic effect of variants in the functional regions of the BRCA1 protein on the human phenotype. A convolutional neural network (CNN) is a class of deep learning model that applies convolutional operations to structured data such as images and has shown excellent performance in various classification tasks [16]. The sequences in the functional regions of the BRCA1 were represented using vector representation in TAPE [15], a pre-trained natural language processing model. The BRCA1 variants have been experimentally verified and obtained from data published in open-access databases.

Materials and Methods

This study assesses the effectiveness of the prediction model in determining variant effects and its suitability for different proteins. The schema for the proposed study is given in Figure 2.

Data Collection and Pre-processing

In the study, we used the open-access database ClinVar, where variants, with experimentally validated effects on the BRCA1 protein and were labeled as benign or pathogenic, were filtered (Access Date: April 15, 2024). Consequently, a total of 532 variants for the BRCA1 protein were filtered according to the RING domain, CC domain and BRCT domain, resulting in the retrieval of 240 variants. The data were filtered according to sequence position and variant-forming amino acids, and it was observed that the same variants were included on more than one occasion in the dataset. The same variants were removed from the dataset and the dataset was given its final form. The dataset includes 224 variants, with 151 (67.5%) classified as pathogenic and 73 (32.5%) as benign. This distribution highlights an imbalance between the two classes.

Resampling

For this study, the dataset was divided into training and test sets, with 80% of the data allocated to the training set and 20% to the test set. To address the class imbalance in the training set, we employed the Synthetic Minority Over-sampling Technique (SMOTE), which generated synthetic data to balance the classes. Additionally, 20% of the training set was reserved as a validation set.

Feature Extraction

For vectorial representation of the BRCA1 protein, we used transfer learning and TAPE embedding techniques. Specifically, we employed three different methods: Long Short-Term Memory (LSTM), Dilated Residual Network (ResNet), and Transformer. These methods are based on the research by Rao et al. (2019), which focused on protein embedding using a semi-supervised approach for various protein-related tasks [15]. The pre-trained model and embedding produce a vector

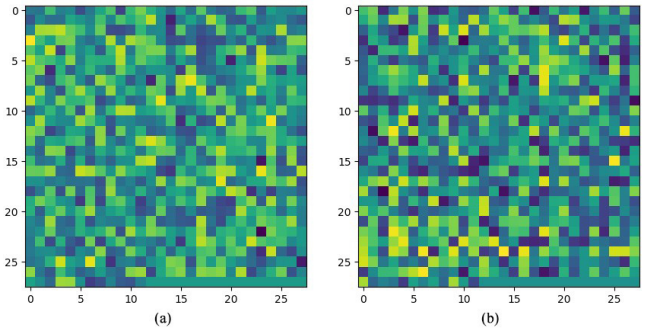


Figure 3. The heat map depicts the distribution of variants within a functional region of the BRCA1 protein. (a) Pathogenic example, (b) Benign example

representation of length 768. This vector was then reshaped and converted into an image through a padding process. Example protein images are illustrated in Figure 3.

The image size was determined directly from the dimensionality of the TAPE embedding vector. The heatmap is based on TAPE embedding values and the color scale represents relative embedding intensity. Each BRCA1 variant region (e.g., RING, CC, BRCT) was represented as a 768-dimensional vector, which was reshaped to fit a square matrix of 28×28 (784 total positions). The additional 16 positions were padded with zeros to maintain spatial consistency. This standardized input size ensures compatibility across all variants, regardless of the length of the original amino acid subsequence, as the embeddings already provide a fixed-length feature vector.

Classification and Final Model

A convolutional neural network (CNN)-based prediction model was developed to assess the deleterious effects of variants in the functional regions of the BRCA1 protein on human phenotypes. Figure 4 shows the architecture of CNN. CNNs are a deep learning method widely used in computer vision due to their effectiveness. In CNNs, mathematical operations known as convolutions are applied between layers to extract and learn features from the data. This process enables the network to learn features efficiently and speeds up the learning process by reducing the dimensionality of the data through pooling operations. The architecture of the CNN used in this study was adapted from the LeNet-5 model (16). Additionally, the prediction model was built upon the ProtConv architecture [17], as recommended in the study by Sara et al. (2021).

As seen in Figure 4, the network receives a 28x28 two-dimensional image as input. This image then passes through six consecutive convolutional and subsampling layers, each utilizing a 3x3 filter size. To maintain consistent dimensions, the padding process used in ProtConv [17] was also applied here.

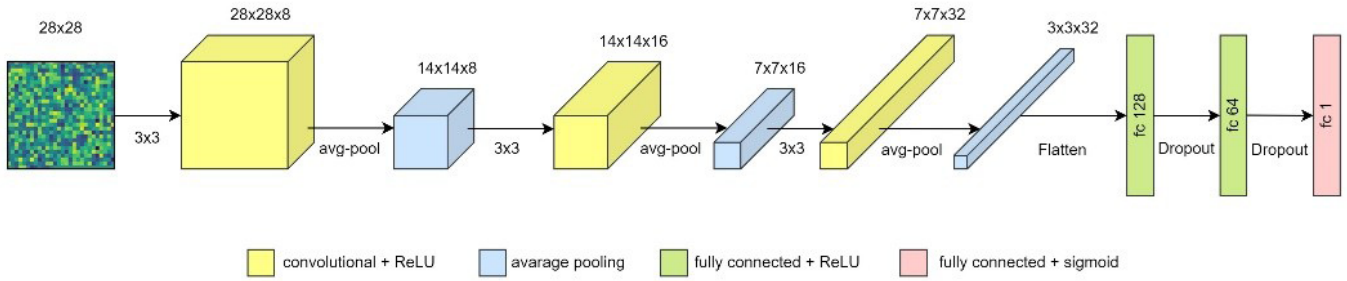


Figure 4. Architecture of the convolutional neural network

Additionally, average pooling, as employed in the LeNet-5 architecture [16], was used in this design.

In the prediction model developed for the study, the convolutional layers utilize 8, 16, and 32 filters, respectively. The Rectified Linear Unit (ReLU) activation function [18] was applied throughout the network. The output from the convolutional layers was flattened and passed through fully connected layers, where ReLU was used for the 128 and 64 neurons, while a sigmoid activation function was employed in the binary classification layer. Two dropout layers, not present in the original LeNet-5 and ProtConv architectures, were incorporated to enhance the model's performance and prevent overfitting. The binary cross-entropy function was used as the loss function in the output layer, and the ADAM optimization algorithm was applied for model optimization. The conversion of protein sequences from one-dimensional to two-dimensional vectors allows the model to effectively learn features in a two-dimensional space. In this study, the following steps were followed for hyperparameter optimization.

- Learning rate tuning was performed using three values: 0.01, 0.001, and 0.0001. As shown in Figure 8, 0.001 produced the best results (AUC = 0.92).
- Batch size was optimized via experimentation (final: batch size = 2).
- Dropout regularization layers (rates: 0.25 and 0.5) were introduced after the fully connected layers to mitigate overfitting - a modification from the base LeNet and ProtConv architectures.
- Early stopping was monitored using validation loss, and binary cross-entropy was used as the loss function.
- The optimizer used was ADAM, with default β values.

Performance Evaluation

To assess the effectiveness of the prediction model developed in this study, a range of performance metrics were utilized, including accuracy (Acc), sensitivity (Sn), specificity (Sp), the Matthews correlation coefficient (MCC), and the area under the curve (AUC). The formula for the computation using the values in the confusion matrix are given below (Equations 1,2,3,4).

$$ACC = \frac{TP+TN}{TP+FN+TN+FP} \quad (1)$$

$$Sn = \frac{TP}{TP+FN} \quad (2)$$

$$Sp = \frac{TN}{TN+FP} \quad (3)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4)$$

TN: true negatives, TP: true positive, FP: false positive, FN: false negative

Furthermore, loss and accuracy curves versus epoch were generated for the prediction convolutional neural network (CNN) model to assess the extent of overfitting. Binary cross-entropy

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_{true} * \log(y_{pred}) + (1 - y_{true}) * \log(1 - y_{pred}) \quad (5)$$

was used to calculate the loss (Equation 5).

In Equation 5, is the actual label for each sample (for binary classification, this value is 0 and 1), and is the predicted probability in model output (positive class). is the total number of samples in the dataset.

Results

Accuracy and loss values for both the training and validation datasets of the convolutional neural network-based prediction model, which was developed to assess the pathogenic effects of variants in the functional regions of the BRCA1 protein on human phenotypes, are illustrated in Figure 5. The training accuracy and loss value are the outcomes of the resampling process using SMOTE. The imbalanced class dataset is not displayed here due to the poor quality of the loss value curve.

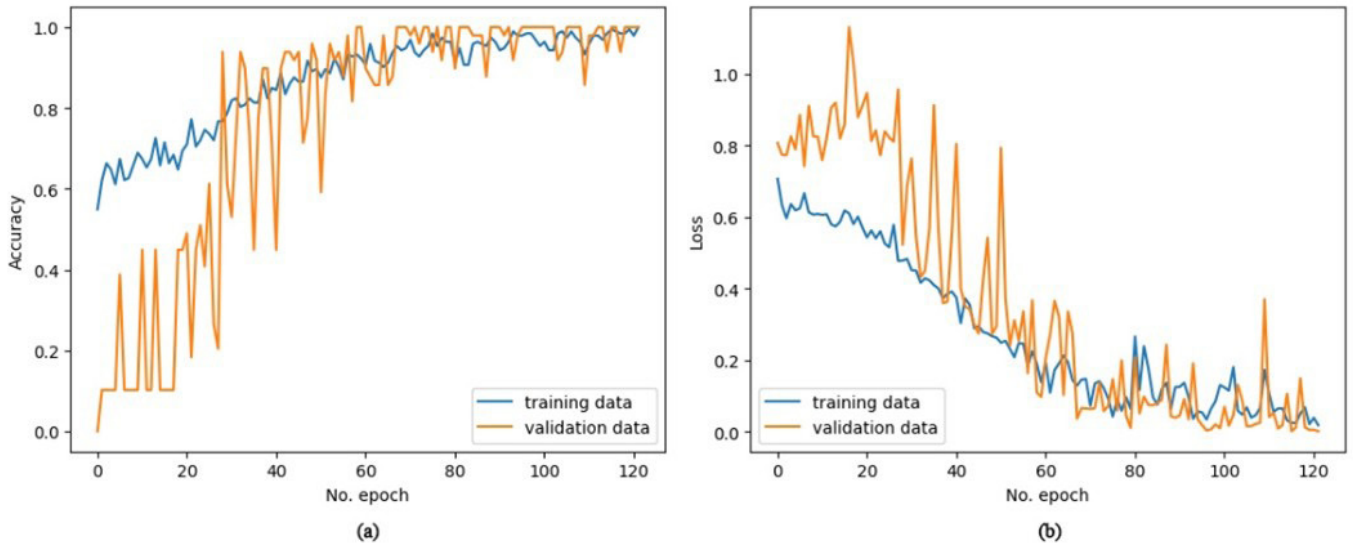
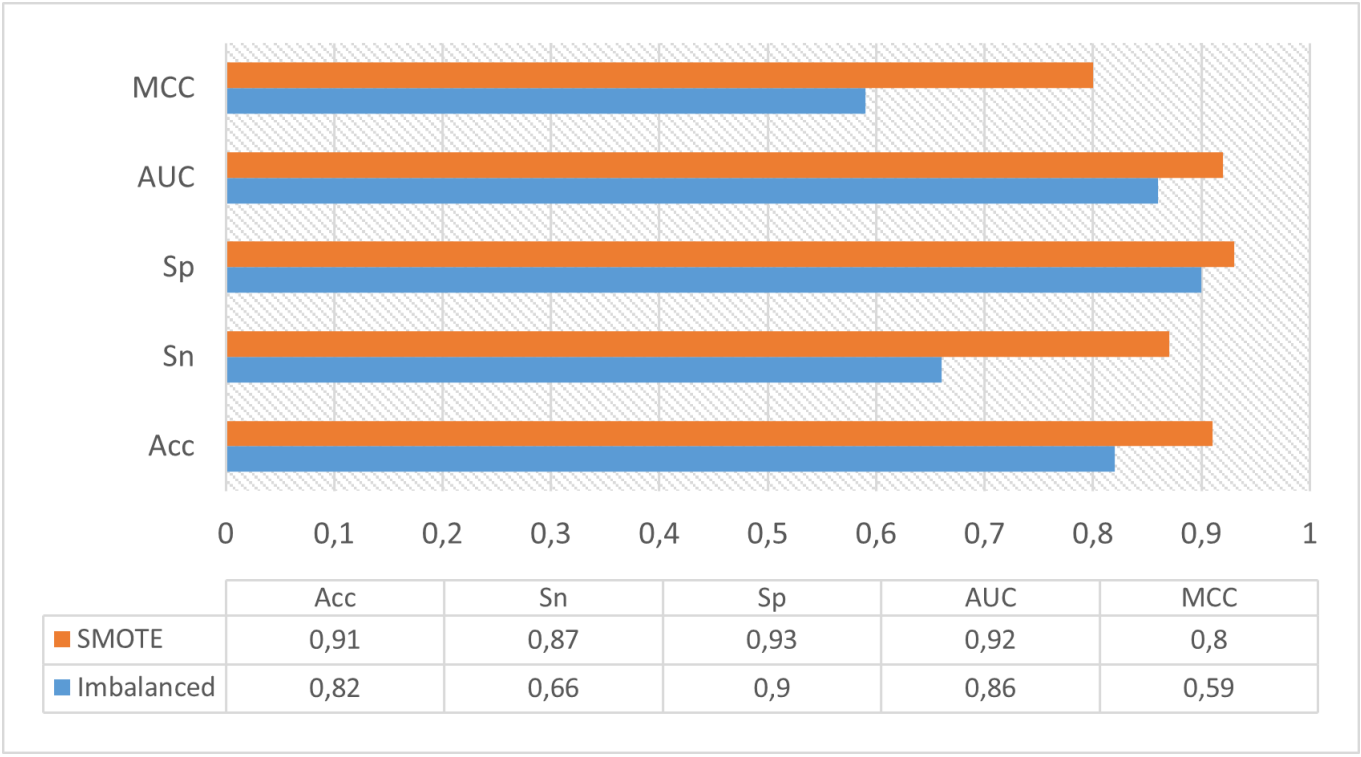


Figure 5. Graphical representation of the accuracy and loss values calculated at each epoch in the developed prediction model for the training and validation dataset, a) accuracy and b) loss values.



	Acc	Sn	Sp	AUC	MCC
SMOTE	0,91	0,87	0,93	0,92	0,8
Imbalanced	0,82	0,66	0,9	0,86	0,59

Figure 6. The results of the performance evaluation of the developed prediction model

An examination of the curves in Figure 5 reveals that the CNN-based prediction model achieved optimal accuracy and minimal loss on both the training and validation datasets after 120 epochs. The results of the analysis demonstrate that the model training was successful, with 120 epochs identified as the optimal number. Figure 5 illustrates that there is no discernible enhancement in accuracy or reduction in loss value following the completion of 120 epochs. Although the training and validation sets reached 100% accuracy before 120 epochs, the training continued to obtain the minimum loss value. The optimal batch size for training was found to be 2.

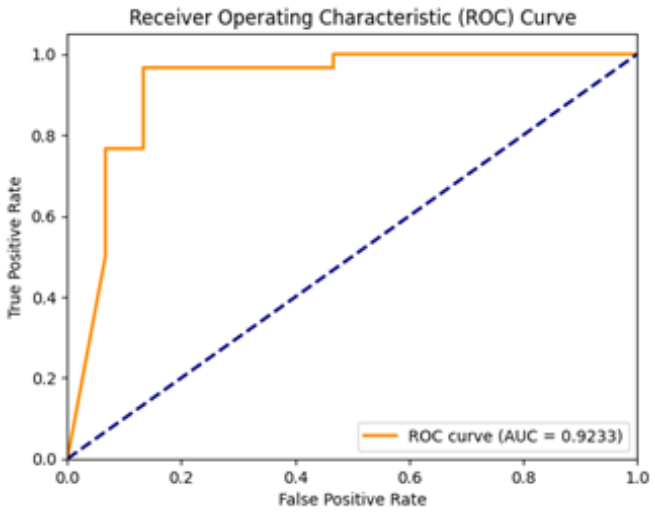


Figure 7. ROC curves for the model utilizing the SMOTE dataset

balanced class dataset. Subsequently, the dataset was resampled using the SMOTE method and employed for training the model. The model achieved an Acc of 0.91, a Sn of 0.87, a Sp of 0.93, an AUC of 0.92, and a MCC of 0.80. The results demonstrate that the second model exhibits superior predictive performance compared to the first model. Although the first model achieved a specificity of 0.96, it is evident that it is unable to accurately predict whether BRCA1 variants are pathogenic or benign. In comparison, the second model demonstrated superior performance, with an AUC of 0.92, and was more effective in predicting the pathogenicity of BRCA1 variants. The detailed receiver operating characteristic (ROC) curve and AUC for the model trained on resampled data using the SMOTE method are illustrated in Figure 7.

In the present study, the learning rate for the model developed was investigated at the following values: 0.01, 0.001, and 0.0001. The learning rate exerts a significant influence on the predictive performance of a CNN model. A relatively high learning

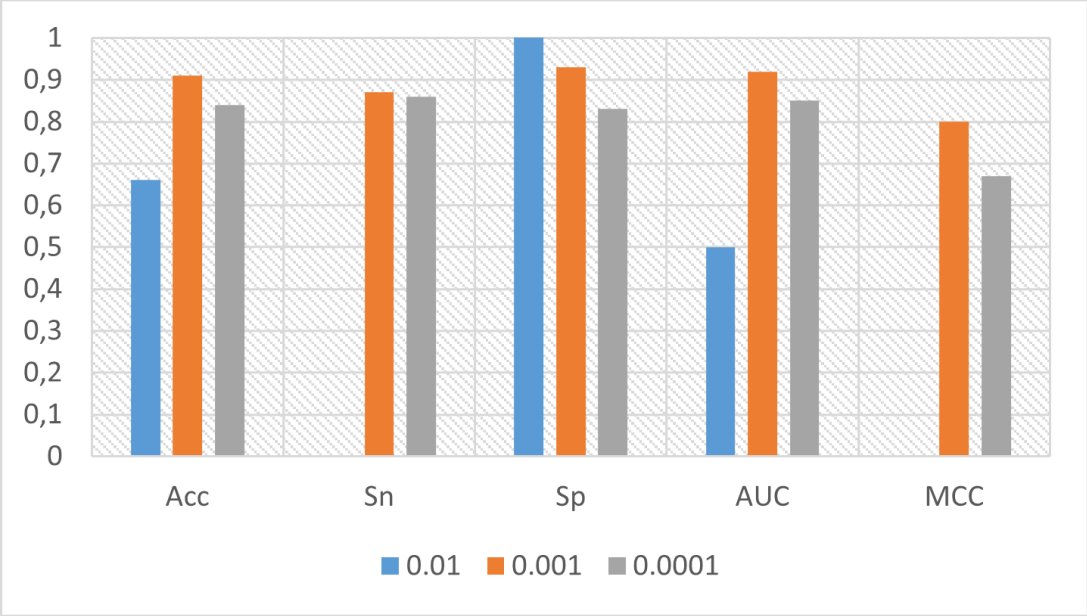


Figure 8. Evaluating the performance of the model at varying learning rates

The results of the performance evaluation for the developed pathogenicity prediction model are presented in Figure 6. Bar plots were preferred over tabular representation to allow intuitive visual comparison of performance metrics. In this figure, the results of the model developed using the training set resampled with the SMOTE method are compared with the model developed using the imbalanced class training set without resampling. The comparison is based on the following metrics: Acc, Sn, Sp, AUC, and MCC.

The initial step involved utilizing an imbalanced class dataset to train the model to facilitate the prediction of BRCA1 variants. The model demonstrated an Acc of 0.82, a Sn of 0.66, a Sp of 0.90, an AUC of 0.86, and an MCC of 0.59 on an imrate may impede the acquisition of crucial features within the dataset, whereas a low learning rate may result in the model learning for an extended period, leading to a decline in prediction accuracy. Consequently, this study aimed to examine the impact of varying learning rates on the model's predictive capabilities. The results of this investigation are presented in Figure 8, which depicts the predictive performance of the developed model at different learning rate ratios.

As illustrated in Figure 8, the model with a learning rate of 0.001 exhibits superior performance compared to models with alternative learning rate values. Accordingly, with a learning rate of 0.001, the Acc was calculated as 0.91, the Sn as 0.87, the Sp as 0.93, the AUC as 0.92, and the MCC as 0.80. Therefore,

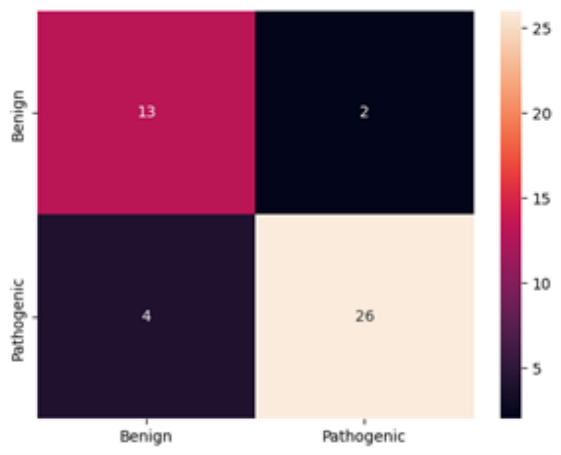


Figure 9. Confusion matrix of the final CNN-based prediction model trained on the SMOTE-resampled dataset

the learning rate of 0.001 was identified as the optimal parameter for achieving the highest performance of the model. To provide a clearer understanding of the classification performance, the confusion matrix of the final CNN model is presented in Figure 9.

Discussion

The objective of this study was to evaluate the effectiveness of a convolutional neural network (CNN)-based deep learning approach in forecasting the pathogenic implications of BRCA1 protein variants on phenotypic expression. This was achieved through the analysis of amino acid sequence data. In this study, the predictive model was developed to predict BRCA1 protein variants using variants obtained from specific regions of the BRCA1 protein. The sequence spanned the RING domain, the coiled-coil domain, and two C-terminal regions, encompassing a total of 1,863 amino acids. The TAPE embedding method, a pre-trained language model designed specifically for proteins, was employed to generate feature vectors from the variant sequence data. Subsequently, the feature vectors were transformed into square-shaped, single-channel images, which were employed to train a deep learning algorithm that integrates a convolutional neural network (CNN). The efficacy of the predictive model concerning BRCA1 variants was assessed utilizing an array of performance metrics, including Acc, Sn, Sp, AUC, and MCC. This approach furnished a comprehensive evaluation of the model's effectiveness.

While more advanced algorithms, such as wavelet, Hilbert, and Fourier transforms, could be employed to convert BRCA1 protein variant sequences into enhanced images, this study aimed to demonstrate the feasibility of using basic image conversion techniques for variant effect prediction. Thus, we utilized fundamental methods in developing the transformation techniques and prediction models. Our findings indicate that the proposed model achieves satisfactory results in predicting the pathogenic effects of BRCA1 protein variants on the phenotype.

It has been demonstrated in numerous studies that models developed using data sets with class imbalance will result in learning that is biased toward the majority class. The prevalence of the biased class results in an increase in false predictions and a tendency for the deep learning model to overfit. It has been proposed that resampling can be conducted using the SMOTE method to address this issue [19]. Nevertheless, despite the SMOTE method's efficacy in addressing class imbalance, its impact on model performance remains to be elucidated [20]. Similarly, the literature indicates that different hyperparameter values impact the success of deep learning models with hyperparameter optimization [21]. For this purpose, we employed resampling with the SMOTE method to develop the most effective models and conducted hyperparameter optimization with diverse learning rate values.

This section presents a comparative analysis of the prediction outcomes produced by the developed model. The most successful model for predicting the pathogenic effects of variants identified in the functional domains of the BRCA1 protein, based on the observed phenotype, was developed using the SMOTE representation and a learning rate value of 0.001. The results obtained from the predictive model developed in this study indicate that BRCA1 variants can be classified with an accuracy of 91% (sensitivity of 93%). Furthermore, the model demonstrated a robust capacity for accurately classifying both benign and pathogenic variants, attaining an AUC value of 92%. These results suggest that the developed predictive model has the potential to categorize BRCA1 variants and evaluate their associated pathogenicity.

In the existing literature, studies utilizing deep learning methods for the prediction of BRCA1 pathogenicity employ a diverse array of data, including DNA sequence data, clinical-pathogenic data, experimentally measured data, evolutionary conservation information, and a multitude of structural feature sets. The reported AUC values for these studies were 0.85 and 0.82, respectively [22, 23]. While the experimental data and evolutionary conservation information of the variants are important in supporting the success of the model, the AUC values obtained did not demonstrate high performance. In another study, an AUC of 93% was obtained with a prediction model developed using tertiary protein structures for variants in the BRCA1 functional region. However, an accuracy of 85% could not be achieved [24]. In this study, a highly successful prediction performance was achieved (AUC: 0.92, accuracy: 0.91, specificity: 0.93) with the convolutional neural network method developed using only BRCA1 variant sequence data. Therefore, the successful prediction performance achieved with a limited amount of data by using only BRCA1 variant sequences and the image-converted data of these sequences as data in this study demonstrates the viability of this study as a model for pathogenicity prediction of BRCA1 variants. Furthermore, the high prediction performance achieved by using basic methods in converting sequence data into images in this

study indicates that the performance of the study can be enhanced by employing more sophisticated methods.

One of the key limitations of this study is the absence of 3D structural information of the BRCA1 protein in the prediction process. While protein folding, structural stability, and domain-level interactions are known to play a significant role in variant pathogenicity, our model relies solely on sequence-derived embeddings. This simplification was intentional, to assess the predictive power of sequence-only representations in a novel image-based framework. However, integrating structural biology features such as solvent accessibility, secondary structure annotations, contact maps, or folding stability metrics (e.g., using tools like FoldX, DSSP, or AlphaFold2) may further enhance model accuracy and biological interpretability. We propose the incorporation of such structure-aware features as a future research direction.

Reviewing the literature on variant effect prediction reveals that existing studies often focus on the structural features of proteins, which can overshadow the development of prediction models based on amino acid-level features. In this study, we created a feature vector for each amino acid in the protein sequence using the transfer learning method (TAPE). This vector was then converted into an image, enabling the development of a successful pathogenicity prediction model using a convolutional neural network. In this study, we have demonstrated that variant effect prediction can be achieved using both natural language processing and image processing methods, thereby eliminating the necessity for structural property information of amino acids.

In the literature, we have not met any study that developed a prediction model using image representation for assessing the pathogenicity of BRCA1 variants. Thus, we can say that utilizing an image representation-based deep learning method to predict the impact of BRCA1 protein variants on the phenotype as either benign or pathogenic, represents a novel contribution to the literature.

The high predictive performance of the model, particularly in classifying variants within the RING, CC and BRCT domains, supports the biological relevance of these regions in BRCA1 function. These domains are known to mediate DNA repair, transcriptional regulation, and protein–protein interactions. The successful classification of variants in these regions suggests that the model may capture features associated with structural or functional disruption, although additional structural validation is needed. This study is limited by its exclusive focus on BRCA1 missense variants and its reliance on sequence-derived embeddings without incorporating evolutionary or structural information. The dataset is relatively small and imbalanced, which may lead to potential overfitting despite our use of SMOTE resampling and dropout regularization. Moreover, as no comparative analysis with existing variant effect prediction tools was conducted, the relative performance of the model remains unvalidated.

Future work will focus on extending the model to include variants from additional cancer-related genes, comparing its performance with established tools such as SIFT, PolyPhen-2, and REVEL. Furthermore, incorporation of structural biology-derived features (e.g., solvent accessibility, secondary structure, and contact maps) may enhance the biological interpretability and accuracy. Feature visualization techniques like Grad-CAM will also be explored to better understand the contribution of embedding regions to pathogenicity classification. Future research will aim to broaden the dataset across multiple genes and incorporate comparative performance analysis against established models.

Conclusion

In this study, a deep learning model based on natural language processing and image representation was developed for the classification of variants in the functional regions of the BRCA1 protein as pathogenic or benign. An imbalanced data set was subjected to a resampling method, and hyperparameter optimization was conducted to enhance the model. The results demonstrated that a deep learning model was developed that exhibited 93% sensitivity in the prediction of benign BRCA1 variants, 92% AUC, and 91% accuracy in both class predictions (pathogenic and benign).

The developed model demonstrated that variants in the functional regions of the BRCA1 protein can be predicted using natural language processing and image representation, without the use of structural features. The variant effect prediction model obtained in this study shows potential and may benefit from further refinement in future research. In future studies, model performance can be improved by using different and more complicated image representation methods, pre-training models such as ResNet, VGG, GoogleNet, transfer learning, and hyperparameter optimizations.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

No funding was received for this research project.

Authors' Contributions

GMŞ: Conceptualization, data curation, methodology, investigation, formal analysis, validation, visualization, writing (original draft). **BK:** Supervision, methodology, writing (review and editing).

Acknowledgements

Not applicable.

References

1. Mayor, S. (2007). Genome sequence of one individual is published for first time. BMJ, 335, 530-531.https://doi.org/10.1136/bmj.39335.753009.94
2. Sarkar, A., & Nandineni, M.R. (2017). Association of common genetic variants with human skin color variation in Indian populations. American Journal of Human Biology, 30, 1-11. https://doi.org/10.1002/ajhb.23068
3. Qiu, J., Nechaev, D., & Rost, B. (2020). Protein–protein and protein-nucleic acid binding residues are important for common and rare sequence variants in human. BMC Bioinformatics, 21, 1-17. https://doi.org/10.1186/s12859-020-03759-0
4. Kechavarzi, B., & Janga, S.C. (2014). Dissecting the expression landscape of RNA-binding proteins in human cancers. Genome Biology, 15, 1-16. https://doi.org/10.1186/gb-2014-15-1-r14
5. Wang, M., Zhao, X.M., Takemoto, K., Xu, H., Li, Y., Akutsu, T., & Song, J. (2012). FunSAV: Predicting the functional effect of single amino acid variants using a two-stage random forest model. PLoS ONE, 7. https://doi.org/10.1371/journal.pone.0043847
6. Lukong, K.E., Chang, K.W., Khandjian, E.W., & Richard, S. (2008). RNA-binding proteins in human genetic disease. Trends in Genetics, 24, 416-425.
7. Yates, C.M., Filippis, I., Kelley, L.A., & Sternberg, M.J. (2014). SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. Journal of Molecular Biology, 426, 2692-2701. https://doi.org/10.1016/j.jmb.2014.04.026
8. National Cancer Institute. (2020). BRCA gene mutations: Cancer risk and genetic testing. National Cancer Institute: Washington, DC, USA.
9. Clark, S.L., Rodriguez, A.M., Snyder, R.R., Hankins, G.D., & Boehning, D. (2012). Structure-function of the tumor suppressor BRCA1. Computational and Structural Biotechnology Journal, 1(1), e201204005. https://doi.org/10.5936/csbj.201204005
10. Ismail, T., Alzneika, S., Riguene, E., Al-maraghi, S., Alabdulrazzak, A., Al-Khal, N., Fetais, S., Thanassoulas, A., Al-Farsi, H., & Nomikos, M. (2024). BRCA1 and its vulnerable C-Terminal BRCT domain: structure, function, genetic mutations and links to diagnosis and treatment of breast and ovarian cancer. Pharmaceuticals, 17(3), 333. https://doi.org/10.3390/ph17030333
11. Harper, J.W., & Elledge, S.J. (2007). The DNA damage response: Ten years later. Molecular Cell, 28(5), 739-745.
12. Deng, C.X. (2006). (2006). BRCA1: cell cycle checkpoint, genetic instability, DNA damage response and cancer evolution. Nucleic Acids Research, 34(5), 1416-1426. https://doi.org/10.1093/nar/gkl010
13. Grzelak, D. (2021). Treatment options for germline BRCA-mutated metastatic pancreatic adenocarcinoma. Journal of the Advanced Practitioner in Oncology, 12(5), 488. https://doi.org/10.6004/jadpro.2021.12.5.4
14. Horne, J., & Shukla, D. (2022). Recent advances in machine learning variant effect prediction tools for protein engineering. Industrial & Engineering Chemistry Research, 61(19), 6235-6245.
15. Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., & Song Y. (2019). Evaluating protein transfer learning with TAPE. Advances in Neural Information Processing Systems, 32, 9689-9701.
16. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proc. IEEE, 86(11), 2278–2324.
17. Sara, S.T., Hasan, M.M., Ahmad, A., & Shatabda, S. (2021). Convolutional neural networks with image representation of amino acid sequences for protein function prediction. Computational Biology and Chemistry, 92, 107494. https://doi.org/10.1016/j.compbiolchem.2021.107494
18. Nair, V., & Hinton, G.E. (2010). Rectified linear units improve restricted Boltzmann machines. Proceedings of the 27th International Conference on International Conference on Machine Learning, *ICML'10*, p. 807–814.
19. Chawla, N.V, Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16,321-357. https://doi.org/10.1613/jair.953
20. Jadwal, P.K., Jain, S., Pathak, S., & Agarwal, B. (2022). Improved resampling algorithm through a modified oversampling approach based on spectral clustering and SMOTE. Microsystem Technologies, 28(12), 2669-2677. https://doi.org/10.1007/s00542-022-05287-8
21. Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. Automated Machine Learning: Methods, Systems, Challenges, 3-33.
22. Li, Y., Chen, L., Lv, J., Chen, X., Zeng, B., Chen, M., Guo, W., Lin, Y., Yu, L., Hou, J., Li, J., Zhou, P., Zhang, W., Li, S., Jin, X., Cai, W., Zhang, K., Huang, Y., Wang, C., & Fu, F. (2022). Clinical application of artificial neural network (ANN) modeling to predict BRCA1/2 germline deleterious variants in Chinese bilateral primary breast cancer patients. BMC Cancer, 22(1), 1125. https://doi.org/10.1186/s12885-022-10160-y
23. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences." In:

Proceedings of the National Academy of Sciences, p.118.
24. Li, C., Zhang, L., Zhuo, Z., Su, F., Li, H., Xu, S., Liu, Y., Zhang, Z., Xie, Y., Yu, X., Bian, L., & Xiao, F. (2023). Artificial intelligence-based recognition for variant pathogenicity of BRCA1 using alphafold2-predicted structures. Theranostics, *13*(1), 391. https://doi.org/10.7150/thno.79362