KNN for Breast Cancer Prediction utilizing Wisconsin Cancer Dataset

Preprint · November 2023

DOI: 10.13140/RG.2.2.11207.88487

4 authors, including:

Md. Sahilur Rahman
American International University-Bangladesh
3 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Munim Ahmed
American International University-Bangladesh
3 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Arjun Kumar Bose Arnob
American International University-Bangladesh
3 PUBLICATIONS **2** CITATIONS
SEE PROFILE

KNN for Breast Cancer Prediction utilizing Wisconsin Cancer Dataset

Md. Sahilur Rahman
Dept. of Computer Science
American International University-Bangladesh
20-43257-1@student.aiub.edu

Munim Ahmed
Dept. of Computer Science
American International University-Bangladesh
20-43303-1@student.aiub.edu

Arjun Kumar Bose Arnob
Dept. of Computer Science
American International University-Bangladesh
20-42156-1@student.aiub.edu

Mostaque Ahammed Niam
Dept. of Computer Science
American International University-Bangladesh
20-42140-1@student.aiub.edu

***Abstract*—Breast Cancer is one of the life-threatening diseases among females all over the world. This killer disease, however, when it can be detected in its early stages, can be a lifesaver for many. A tumor is a mass of abnormal tissue. There are two types of breast cancer tumors: those that are non-cancerous or benign and those that are cancerous, which are 'malignant'. Radiologists use the mammography images to detect the presence and absence of Breast Cancer. The field of bio-informatics leverages machine learning techniques for diagnosing Breast cancer in particular. This research work experiments with the most popularly used supervised machine learning algorithms, K-Nearest Neighbors, SVM, and logistic regression. KNN was used to classify breast cancer disease and implemented for different k-fold cross-validation and k values. Then, the obtained classification accuracies were compared with logistic regression. This work predicts Breast Cancer on the Breast Cancer Data Set (BCD) taken from the UCI Machine Learning Repository. The proposed work has achieved the best accuracy of 93.68% and the lowest accuracy of 90.35% by employing the KNN algorithm.**

***Index Terms*—KNN, breast cancer, machine learning, computer aided diagnosis**

I. Introduction

Early detection of cancer is crucial for a quick response and a better chance of recovery. Unfortunately, early detection of cancer is often difficult because there are no symptoms at the beginning. Thus, cancer remains one of the topics of health research in which many researchers have invested with the aim of creating evidence that can improve treatment, prevention and diagnostics. Moreover, there is a worldwide shortage of healthcare professionals, specially in Bangladesh where a single doctor is accountable for 1901 patients in average [1]. For breast cancer pathologists this figure is even worse.

Research in this field is a search for knowledge through surveys, studies and experiments conducted with applications to discover and interpret new knowledge in order to prevent and minimize the risk of adverse consequences. To better understand this issue, tools are still needed to help oncologists select the treatment needed to cure or prevent recurrence by reducing the harmful effects of certain treatments and their cost.

A. Logistic Regression

Logistic regression is a supervised machine learning technique used in classification tasks (for making predictions based on training data). Logistic regression uses an equation similar to linear regression, but the logistic regression result is a categorical variable while it is a value for other regression models. Binary outcomes can be predicted from the independent variables. The result of the dependent variable is discrete. Logistic regression uses a simple equation that shows the linear relationship between the independent variables. These independent variables along with their coefficients are linearly combined to form a linear equation that is used to predict the output. The equation used by the basic logistic model. The logistic function is used here to suppress the result value between 0 and 1. The logistic function can also be called a sigmoid function or a cost function. The logistic function is a shaped curve that takes the input (numeric value) and changes it to a value between 0 and 1 [2].

B. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. In the classification technique, it classifies the objects based on the k closest training examples in the feature space. The working principle behind KNN assumes that the same data points are in the same environment. It reduces the effort of building a model, adjusting a set of parameters, or making more assumptions. It captures the idea of proximity based on a mathematical formula called Euclidean distance. SVM was found effective in multiple studies related to the breast cancer analysis and diagnosis [3], [4]. However, these studies mainly utilized image based data unlike this study.

C. K-nearest Neighbor

K-nearest Neighbor is a supervised machine learning algorithm because the data passed to it is labeled. It is a non-parametric method because the classification of test data points is based on the closest training data points instead of considering the dimensions (parameters) of the dataset. It is used to solve classification and regression tasks

An object to be classified is assigned to the respective class that represents the greater number of its nearest neighbors. If k is 1, then the data point is placed in the category containing only one nearest neighbor. Given a new input data point, the distances between those points and all data points in the training dataset are calculated. Based on the distances, the training set data points with shorter distances from the test data point are considered the nearest neighbors of our test data. Finally, the test data point is placed into one of its nearest neighbor classes. Therefore, the classification of the test data point depends on the classification of its nearest neighbors. Choosing the value of K is the crucial step in implementing the KNN algorithm. The value of K is not fixed and varies for each record depending on the type of record. When the value of K is smaller, the stability of the prediction is lower. In the same way, if we increase its value, the ambiguity will be reduced, resulting in smoother borders and increasing stability. With KNN, the assignment of a new data point to a category depends entirely on the Ks value. K represents the number of closest training data points in the vicinity of a given test data point, and then the test data point is assigned to the class containing the highest number of nearest neighbors (i.e., high frequency class).

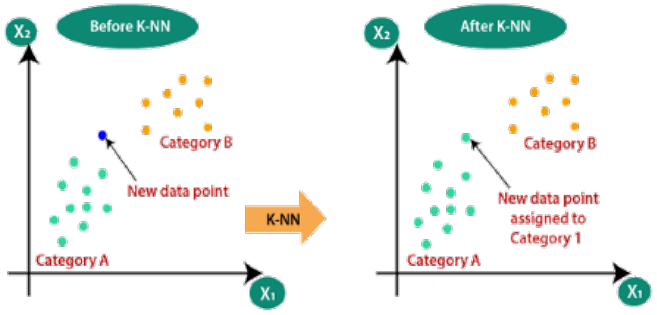

Fig. 1. K-Nearest Neighbors method.

II. Methodology

The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository is used, to distinguish malignant (cancerous) from benign (non-cancerous) samples. website where the dataset was collected in python's format. Dataset file was in csv format. There are 8 groups and 699 information into the chronological grouping of the data.

A. Dataset Description

Samples arrive periodically as Dr. Walberg reports his clinical cases. The database therefore reflects this chronological

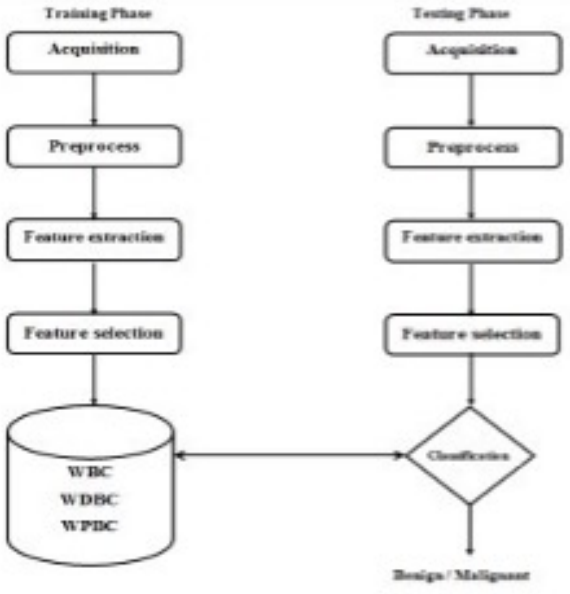

Fig. 2. Schematic diagram of proposed method.

grouping of the data. This grouping information appears immediately below, having been removed from the data itself. KNN was used to classify breast cancer disease and implemented for different k-fold cross-validation and k values. Then, the obtained classification accuracies were compared with logistic regression. Total amount of malignant and benign data in the Wisconsin dataset is shown in Table 2.

B. Feature Extraction and Selection

An important step in the breast cancer diagnostic model is feature extraction. The optimal feature set should have effective and distinctive features while mainly reducing the redundancy of the feature space to avoid the problem of the curse of dimensionality. The curse of dimensionality suggests that the sample density of the training data is too low to promise a meaningful estimate of a high-dimensional classification function with the available finite number of training data. A high number of features may lead to marginally higher accuracy but it requires significantly higher amount of time [5]. Therefore, it is important to select optimal number of features to achieve sufficient accuracy withing practically achievable time limit.

C. Wisconsin Breast Cancer Dataset

We have used Wisconsin Breast Cancer Dataset [6] for this study. The details of the attributes found in this dataset listed in table. In clump thickness, benign cells tend to be grouped in monolayers while cancer cells are often grouped in multilayers. Whereas in the uniformity of cell size/shape, the cancer cells tend to vary in size and shape. Because of this, these parameters are valuable in determining whether the cells are cancerous or not. In the case of marginal adhesion, the normal cells tend to stick together while cancer cells tend to

lose this ability. Loss of adhesion is thus a sign of malignancy. In the case of the single epithelial cell size, the size is related to the uniformity mentioned above. epithelial cells that are significant enlarged may be a malignant cell. The Bare nuclei is a term used for nuclei that is not surrounded by cytoplasm (the rest of the cell). Those are typically seen in benign tumors. Total amount of malignant and benign data in the Wisconsin dataset is shown in Figure 5.

The Bland Chromatin describes a uniform "texture" of the nucleus seen in benign cells. In cancer cells the chromatin tends to be coarser. The Normal nucleoli are small structures seen in the nucleus. In normal cells the nucleolus is usually very small if visible. In cancer cells the nucleoli become more prominent, and sometimes there are more of them. Finally, Mitoses is nuclear division plus cytokines and produce two identical daughter cells during prophase. It is the process in which the cell divides and replicates. Pathologists can determine the grade of cancer by counting the number of mitoses.

Dataset	No. of Attributes	No. of Instances	No. of Classes
Wisconsin Breast Cancer (Original)	11	699	2

Fig. 3. Description of Breast Cancer Dataset.

	Attribute	Domain
1	Sample code number	id number
2	Clump Thickness	1 - 10
3	Uniformity of Cell Size	1 - 10
4	Uniformity of Cell Shape	1 - 10
5	Marginal Adhesion	1 – 10
6	Single Epithelial Cell Size	1 – 10
7	Bare Nuclei	1 – 10
8	Bland Chromatin	1 – 10
9	Normal Nucleoli	1 – 10
10	Mitoses	1 - 10
11	Class	2 for benign, 4 for malignant

Fig. 4. Wisconsin Breast Cancer Dataset Attributes.

D. Wisconsin Diagnosis Breast Cancer (WDBC)

The details of the attributes found in WDBC dataset: ID number, Diagnosis (M = malignant, B = benign) and ten real valued features are computed for each cell nucleus: Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry and Fractal dimension [6]. These features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image [16]. When the radius of an individual nucleus is measured by averaging the length of the radial line segments defined by the centroid of the snake and the individual snake points. The total distance between consecutive snake points constitutes the nuclear perimeter. The total distance between consecutive snake points constitutes the nuclear perimeter. The area is measured by counting the number of pixels on the interior of the snake and adding one-half of the pixels on the perimeter. The perimeter and area are combined to give a measure of the compactness of the cell nuclei using the formula. Smoothness is quantified by measuring the difference between the length of a radial line and the mean length of the lines surrounding it. This is similar to the curvature energy computation in the snakes. Concavity captured by measuring the size of the indentation (concavities) in the boundary of the cell nucleus. Chords between nonadjacent snake points are drawn and measure the extent to which the actual boundary of the nucleus lies on the inside of each chord. Concave Points: This feature is Similar to concavity but counted only the number of boundary point lying on the concave regions of the boundary. In order to measure symmetry, the major axis, or longest chord through the center, is found. Then the length difference between lines perpendicular to the major axis to the nuclear boundary in both directions is measured. The fractal dimension of a nuclear boundary is approximated using the "coastline approximation" described by Mandelbrot. The perimeter of the nucleus is measured using increasingly larger "rulers". As the ruler size increases, decreasing the precision of the measurement, the observed perimeter decreases. Plotting log of observed perimeter against log of ruler size and measuring the downward slope gives (the negative of) an approximation to the fractal dimension. With all the shape features, a higher value corresponds to a less regular contour and thus to a higher probability of malignancy. The texture of the cell nucleus is measured by finding the variance of the gray scale intensities in the component pixels.

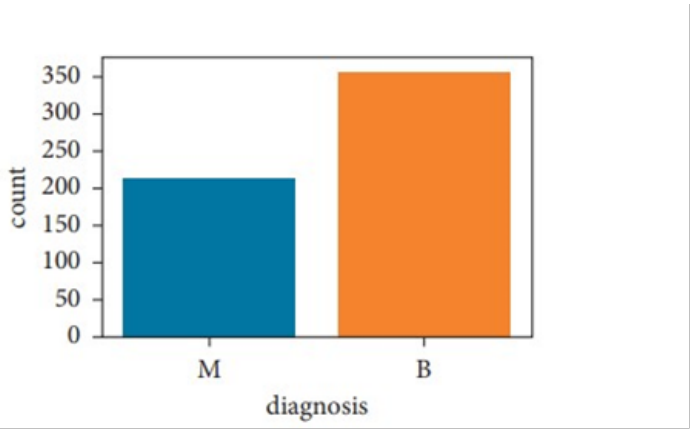

Fig. 5. Total number of malignant (M) and benign (B) data.

E. Data Understanding and Data Selection

WBCD consisted of 699 instances and 11 features. These 11 features provide precise information pertaining to the occurrence of breast cancer. Moreover, the dataset was scrutinized for unknown values, inconsistency and erroneous data.

Unknown values can have a consequential effect on the interpretations that can be derived from the data. 16 instances with missing values are present which are denoted by"?" in the Breast cancer dataset.

Each missing value is replaced with random number constant and is ignored during analysis Understanding of data distribution among dataset is very critical task and must be done effectively. Analysis of data distribution uncovers various interesting relationship and insights which can be useful in selecting best predictive feature.

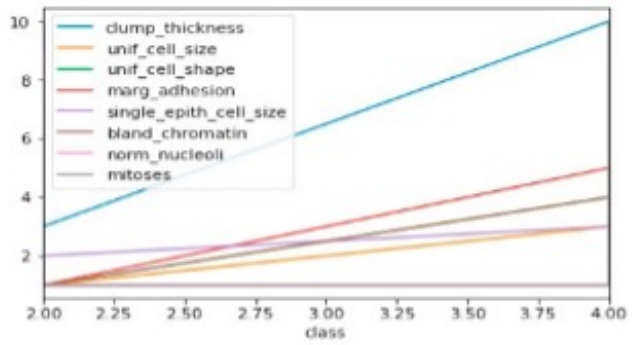

Fig. 6. Collective data variability with respect to class.

F. Training and Classification

Classifications of the data sets are done on the basis of specific properties possess by the sample variable that the capable to classify them, and each sample variable is assigned a malignant or benign class. Classification is principally done by making predictions based on known sample data that has been learned from training data. Designed algorithm is first trained on the known data labels and further uses this learning to predict the class labels for the new unknown set of data sample. The classification objective set for this study is to achieve enhanced accuracy by using LR, SVM and KNN classifiers and determine which one suits the most for diabetes classification technique.

We train the classifier with known sample data in a training dataset and check its performance by examining the test dataset, which consists of the unknown sample used to predict its class label K Neighbors Classifier is a supervised, instance-based learning classifier which learns from the labeled data samples. The pseudo code for the KNN classifier is given in Algorithm 1. K folds cross validation technique is used for training data. In this technique, the original sample is divided into k equivalent size subsamples and one subsample is used for validating the model, while the k-1 remaining subsamples are utilized as training data. After that this cross-validation process is recurring k times (called the folds), with every of the k subsamples just used one time as the validation data. It works in loop manner. In this study we set the value of k=10.

III. Result and Analysis

This segment covers all strategies and materials, as well as the dataset's depiction, block graph, stream chart, and assessment grids. We have randomly chosen the data to construct the training set. We have plotted a graph to check the error rate against k-value and We have experimented with different values of K from K = 1 to 30 while changing the training and testing size. With the KNN algorithm, the classification result of the test set fluctuates between 89.12% and 95.02%. The best performance is obtained when K is 13.

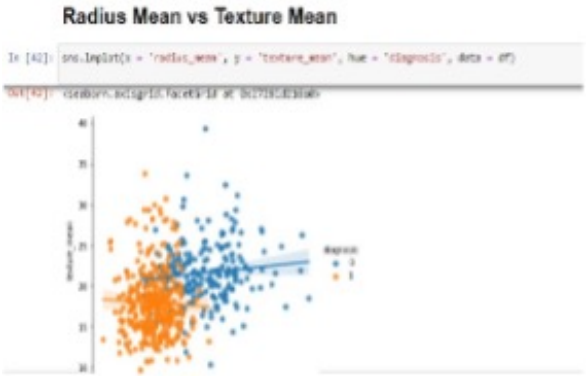

Fig. 7. Radius Mean vs Texture Mean.

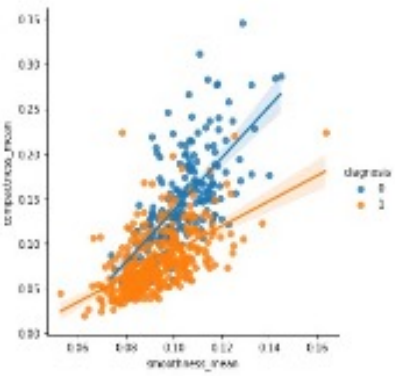

Fig. 8. Smoothness Mean vs Compactness Mean.

The reason behind the conduct of this study is to predict breast cancer using the KNN approach. We used Kaggle's Breast Cancer Wisconsin (Diagnostic) Data Set consists of 699 instances in chronological format. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Compared to the methods reported in [7]–[13] the advantages of the KNN algorithm are that the algorithm is very simple, and its implementation is very easy. Since there is no need of any training session, there is no convergence problem. In contrast, the other approaches employing neural networks may face the convergence problem, and may need long training time. New training data can also be added to the KNN algorithm without any retraining. But for the other techniques, adding new training data needs retraining because the new training data disturb the structure of the existing training set, and all the parametric or semiparametric classifiers critically depend on this structure.

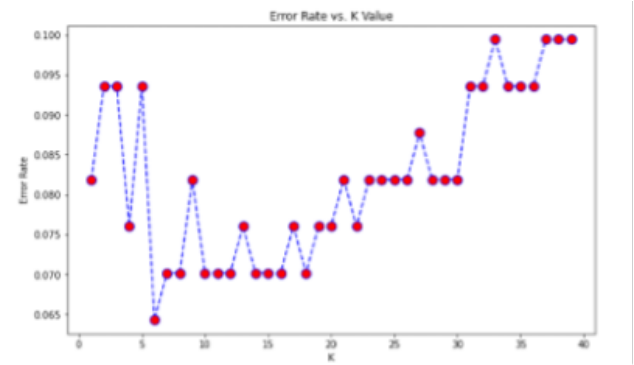

Fig. 9. Error rate vs k value.

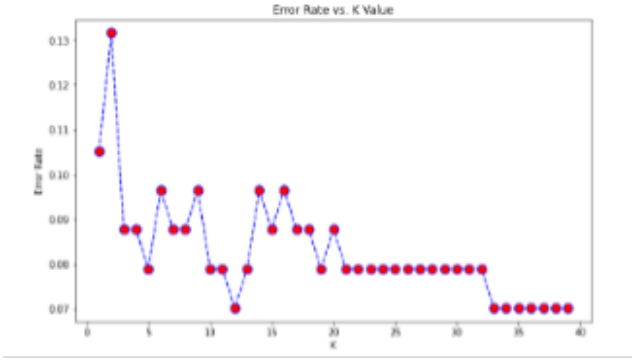

Fig. 10. Train and test.

IV. Conclusion

This paper treats the Wisconsin-Madison Breast Cancer diagnosis problem as a pattern classification problem. The KNN algorithm is used as the nonparametric classifier. The KNN algorithm assigns the class label of the new datum based on the class label that most of the K-closest training data possess. The KNN algorithm yields the best classification performance that is obtained so far on this problem. However, we believe the performance of the model can be further improved by cleaning the input data. Cleaning and evaluating data to eliminate erroneous or noise data plays a crucial role in machine learning applications specially for CNN based computer vision task [14]–[17]. In the future, the performance of the model after data cleaning should be investigated.

References

[1] Hossain, M., Syeed, M., Fatema, K. & Uddin, M. The perception of health professionals in Bangladesh toward the digitalization of the health sector. *International Journal Of Environmental Research And Public Health*. **19**, 13695 (2022)

[2] Frank, A. UCI machine learning repository. *Http://archive. Ics. Uci. Edu/ml*. (2010)

[3] Hossain, M., Hanna, M., Uraoka, N., Nakamura, T., Edelweiss, M., Brogi, E., Hameed, M., Yamaguchi, M., Ross, D. & Yagi, Y. Automatic quantification of HER2 gene amplification in invasive breast cancer from chromogenic in situ hybridization whole slide images. *Journal Of Medical Imaging*. **6**, 047501-047501 (2019)

[4] Hossain, M., Syeed, M., Fatema, K., Hossain, M. & Uddin, M. Singular nuclei segmentation for automatic HER2 quantification using CISH whole slide images. *Sensors*. **22**, 7361 (2022)

[5] Shakhawat, H., Nakamura, T., Kimura, F., Yagi, Y. & Yamaguchi, M. Automatic quality evaluation of whole slide images for the practical use of whole slide imaging scanner. *ITE Transactions On Media Technology And Applications*. **8**, 252-268 (2020)

[6] Marcano, A., Quintanilla-Dominguez, J. & Andina, D. WBCD breast cancer database classification applying artificial metaplasticity neural network. *Expert Systems With Applications*. **38**, 9573-9579 (2011)

[7] Vispute, N., Sahu, D. & Rajput, A. An empirical comparison by data mining classification techniques for diabetes data set. *International Journal Of Computer Applications*. **131**, 6-11 (2015)

[8] Kumari, M. & Singh, V. Breast cancer prediction system. *Procedia Computer Science*. **132** pp. 371-376 (2018)

[9] Yeh, W., Chang, W. & Chung, Y. A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. *Expert Systems With Applications*. **36**, 8204-8211 (2009)

[10] Kaya, Y. & Uyar, M. A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease. *Applied Soft Computing*. **13**, 3429-3438 (2013)

[11] Nahato, K., Harichandran, K., Arputharaj, K. & Others Knowledge mining from clinical datasets using rough sets and backpropagation neural network. *Computational And Mathematical Methods In Medicine*. **2015** (2015)

[12] Abbass, H. An evolutionary artificial neural networks approach for breast cancer diagnosis. *Artificial Intelligence In Medicine*. **25**, 265-281 (2002)

[13] Chen, H., Yang, B., Liu, J. & Liu, D. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Systems With Applications*. **38**, 9014-9022 (2011)

[14] Hossain, M., Nakamura, T., Kimura, F., Yagi, Y. & Yamaguchi, M. Practical image quality evaluation for whole slide imaging scanner. *Biomedical Imaging And Sensing Conference*. **10711** pp. 203-206 (2018)

[15] Shakhawat, H., Hanna, M., Ibrahim, K., Serrette, R., Ntiamoah, P., Edelweiss, M., Brogi, E., Hameed, M., Yamaguchi, M., Ross, D. & Others Automatic Grading of Invasive Breast Cancer Patients for the Decision of Therapeutic Plan. *Artificial Intelligence For Disease Diagnosis And Prognosis In Smart Healthcare*. **7** pp. 123 (2023)

[16] Shakhawat, H., Hossain, S., Kabir, A., Mahmud, S., Islam, M. & Tariq, F. Review of Artifact Detection Methods for Automated Analysis and Diagnosis in Digital Pathology. *Artificial Intelligence For Disease Diagnosis And Prognosis In Smart Healthcare*. pp. 177-202 (2023)

[17] Hossain, M., Shahriar, G., Syeed, M., Uddin, M., Hasan, M., Shivam, S. & Advani, S. Region of interest (ROI) selection using vision transformer for automatic analysis using whole slide images. *Scientific Reports*. **13**, 11314 (2023)