Application of Bayes and knn classifiers in tumor detection from brain MRI images

Marta Mirkov and Ana Gavrovska, *Member, IEEE*

***Abstract*— Automatic detection of regions of interest is of great importance in computer-aided diagnosis systems. Magnetic Resonance Imaging (MRI) of head due to good soft-tissue contrast is widely used for brain tumor detection showing potential anomalies that indicate the need for further treatment. Current algorithms for processing and classification of medical images often involve complex designs of deep learning that require significant hardware resources and considerable execution time in order to assist doctors in detecting diseases. This may lead to labeling more complex cases in brain tumor detection. In this paper, statistical features are considered with application of Bayes and kNN classifiers showing comparable results having in mind publicly available brain tumor detection database.**

***Index Terms*— Magnetic Resonance Imaging, brain tumor detection, segmentation, feature extraction, machine learning.**

I. Introduction

Today, depending on the need, various medical imaging modalities are used: X-ray, fluoroscopes, mammography, computer tomography (CT) devices as well as devices based on nuclear medicine techniques – Positron Emission Tomography (PET) and Single Photon Emission Tomography (SPECT). However, equipment that does not require ionizing radiation can be used for computer-aided diagnosis systems. Magnetic Resonance Imaging (MRI) scanners employ strong magnetic fields and magnetic field gradients to generate images of the organs or the whole body, and are found very useful in diagnostics. For example, according to the World Health Organization (WHO), cancer is the second leading cause of death [1]. Cancer detection from biopsy procedures is a painful process for patients, and therefore appropriate medical imaging modalities can facilitate this procedure [2].

Images obtained from MRI scanners show satisfying soft-tissue contrast, which is suitable for brain imaging. They provide a good visualization of the posterior cranial fossa, which contains the brain stem and cerebellum. The contrast between gray and white matter makes MRI the best choice for diagnosing many central nervous system conditions, including demyelinating diseases, dementia, cerebrovascular disease, infectious disease, Alzheimer's disease, and epilepsy [3]. Since many images are taken in milliseconds, it shows how the brain reacts to different stimuli, thus enabling doctors to study the brain's functional and structural abnormalities.

In this paper, statistical feature extraction and application of two classifiers will be observed, where the aim is to create a simple algorithm for brain tumor detection [4]-[9]. One of the motivations is usefulness of texture related features in MRI images [10]. Also, one of the famous examples of hypothesis testing is the Bayes test [5], which will be implemented here, and compared to knn (k-nearest neighbors) approach [7] based on revising hand-crafted statistical features.

The paper is organized as follows. After the introduction, in Section II related work is presented. It considers traditional statistical feature extraction and machine learning usage in brain tumor detection. Steps in the experimental analysis performed in this paper are explained in Section III, where further details for feature extraction are given in Section IV. Section V is dedicated to classifiers design and performance evaluation. Obtained results and conclusion are given in Section VI and Section VII, respectively.

II. Related work

The brain tumor is an abnormal growth of cancer cells in the brain which disrupts the work of functional cells. Early detection and rapid diagnosis of tumors can help save the patient's life. Mathematical and software tools can be very successful in detecting brain abnormalities. Thus, related work is oriented towards statistical feature extraction and machine learning methods used in brain tumor detection.

Statistical features are found useful in machine learning and medical image segmentation and classification [10]-[12]. Particularly, in brain tumor detection texture is one of the most valuable features in designating the image appearance [10], [13]. It can be described statistically for the purpose of distinguishing image characteristics by the spatial allotment of gray levels. The most popular mathematical representation of image texture is co-occurrence matrix. For example, in the Gray Level Co-occurrence Matrix (GLCM), the spatial relationship of pixels is considered to examine the texture by using statistical methods. Four features can be extracted and found useful as in [10], [13]-[14]: energy, correlation, contrast, homogeneity. Namely, these features are used for extracting features and forwarding it to neuro-fuzzy models and, generally, machine and deep learning methods [15]-[18]. In [13] support vector machines are applied as classifier, where in [18] authors implemented deep convolutional neural network, and one of the publicly available datasets for brain tumor detection [19]. Using hand-crafted features are still valuable for obtaining satisfying results in brain tumor detection and more parameters may give better results. On the other hand, smaller dimension of the feature vector is important for algorithm execution, especially in the cases

Marta Mirkov is with the University of Belgrade - School of Electrical Engineering, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mail: marta.mirkov@gmail.com).

Ana Gavrovska is with the University of Belgrade - School of Electrical Engineering, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mails: anaga777@gmail.com ; anaga777@etf.rs).

where there is no need for higher complexity according to tested dataset. In this paper, Bayes and knn classifier are analyzed for classification model implementation using [19].

III. Experimental analysis

The cancer tissue is expected to stand out from the normal part of image, but the question of choice of MRI features still remains. The experimental analysis performed here consists of:

- tumor segmentation based on labeled regions in images,
- hand-crafted feature extraction,
- classification and evaluation of results on available dataset.

The proposed work is tested with Brain MRI Images for Brain Tumor Detection dataset [19], containing 98 brain MRI images with healthy tissue and 155 brain MRI images with cancer tissue. Among features for segmentation task solidity of labeled regions is applied. Having in mind traditional hand-crafted features for classification: energy, correlation, contrast and homogeneity, two classifiers are tested in initial phase. Based on preliminary segmentation analysis, two additional statistical features are added for the binary classification. Both tested classifiers, Bayes and knn, are evaluated on the test set, which is made from 30 % of the whole dataset selected in a random manner, where the rest was used for the training set. For performance evaluation true positive rate (TPR), true negative rate (TNR) and balanced accuracy (BACC) are calculated as in (1)-(3), respectively:

1) *True Positive Rate- TPR*

$$Sensitivity = \frac{TP}{P}, \tag{1}$$

where TP represents true positive, the number of samples that were positive and detected as positive, and P represents the whole set of positive samples;

2) *True Negative Rate- TNR*

$$Specificity = \frac{TN}{N}, \tag{2}$$

where TN represents true negative, the number of samples that were negative and detected as negative, and N represents the whole set of negative samples;

3) *Balanced Accuracy- BACC*

$$BACC = \frac{TPR + TNR}{2}. \tag{3}$$

IV. Proposed tumor segmentation

Proposed brain tumor segmentation is consisted of several steps. Firstly, MRI image is preprocessed using: high-pass filtering and image intensity adjustment, and then, after image binarization, connected regions are labeled. Finally, solidity is implemented for tumor segmentation.

A. MRI image preprocessing and labeling

High-frequency filtering highlights sudden changes in the image by passing high-frequency components [4]. A fifth order Gaussian HP filter with cutoff frequency of 55 Hz was used. Although MRI images provide good tissue contrast, adjusting image intensity is important for tumor segmentation. By saturating the bottom 2% and the top 1% of all pixel values the grayscale range of image is shortened and the contrast is enhanced, and therefore it highlights tumor areas.

Because the image contrast is adjusted, the histogram of the image has a bimodal distribution with a deep and sharp valley between the two peaks, which enabled using Otsu's method for automatic thresholding [4]. Finally, binary image has different white regions where some of them might represent tumor tissue. Pixels are connected and are part of one region if their edges touch. Two adjacent pixels are part of the same object if they are both of the same intensity and are connected along all directions.

B. Solidity characterization

For each labeled region, solidity can be calculated. Solidity is a measurement of the overall concavity of a particle. It is defined as the image area divided by the convex hull area. As the object form digresses from a closed circle, the convex hull area increases, and the calculated solidity decreases. Images with high solidity are more likely to contain tumor regions [13, 23]. A tumor can successfully be detected by comparing the calculated solidity of the labeled image with a higher value (closer to 1). In this case, the tumor is detected if solidity is higher than 0.6. When this area represents a tumor, it is the region or candidate with more white pixels, and when there is no tumor tissue on the image, some small regions can still be extracted. It may be assumed that differences between healthy and tumor tissue can be provided using features that describe the area and shape of this extracted region.

V. Statistical hand-crafted features

GLCM is created by calculating how often a pixel with grayscale intensity value *i* occurs horizontally adjacent to a pixel with the value *j*. Offset isn't used for defining pixel spatial relationships. The number of gray levels in the image determines the size of the GLCM. Scaling to 8 gray levels is used to reduce the number of intensity values in an image, so the size of GLCM is 8x8 pixels. The traditional statistical hand-crafted features extracted from GLCM are: energy, correlation value, contrast and homogeneity.

Energy estimates the sum of squared elements from GLCM and represents feature 1:

$$Energy = \sum_{i,j=0}^{N-1} p^2(i,j), \tag{4}$$

where N represents the number of pixels in image, i and j the location of pixel, and $p\ (i,j)$ the intensity of the pixel at the location (i,j).

The mentioned pixel pairs are then estimated using joint probabilities. It gives linear dependency of the gray levels of neighboring pixels. In general, it ranges from [-1,1]:

$$Correlation = \sum_{i,j=0}^{N-1} \frac{(i-\mu x)(j-\mu y)p(i,j)}{\sigma_x \sigma_y}, \tag{5}$$

where μ_x, μ_y, are the means and σ_x and σ_y are the standard deviations of P_x and P_y, respectively. Note that $P_x(i)$ is the ith entry in the marginal-probability matrix obtained by summing the rows of $P(i,j)$ and $P_y(i)$ is the ith entry in the marginal-probability matrix obtained by summing the rows of $P(i,j)$.

Contrast value (feature 3) estimates the local variations, i.e. sum of square variance, as in (6):

$$Contrast = \sum_{i,j=0}^{N-1} |i-j|^2\, p(i,j). \tag{6}$$

The fourth feature, homogeneity, estimates the closeness of distributed pixels.

$$Homogenity = \sum_{i,j=0}^{N-1} \frac{p(i,j)}{1+|i-j|} . \tag{7}$$

These four features are applied in brain tumor detection, but having in mind the segmentation task, in combination with two more features like the number of white pixels (feature 5) and skewness (feature 6), one may assume further improvements can be obtained. The number of white pixels in the segmented images are expected to increase the accuracy results. Also, some additional statistical parameters can be great indicators of tissue condition. Skewness represents a measure of the asymmetry of the probability distribution [20]. It can be expected that umor tissue has much higher skewness than healthy tissue, thus it is also a reliable feature for classification.

VI. Bayes and knn Classifiers and Evaluation of the Results

A. *Bayes classifier*

For implementing Bayes classifier, it is necessary to define the posterior probabilities $q_i(X)$ which represent the conditional probability that the sample X comes from the class ω_1 if its exact realization is known. Using the Bayes theorem, these probabilities can be calculated if priori probabilities of occurrence of class p_i and posterior density probability functions of measured vectors $q_i(X)$ are known [5], [11]. In this case, the first class represents images with tumor tissue, and the second represents healthy tissue. A simple decision rule can be made based on conditional probabilities:

$$q_1(X) > q_2(X) \Rightarrow X \in \omega_1 \tag{8}$$

$$q_2(X) > q_1(X) \Rightarrow X \in \omega_2 \tag{9}$$

Although the probability density functions of the classes are not known, it can be assumed that, if there is a large number of samples, they can be taken as Gaussian (according to the central limit theorem) [6]:

$$f(x) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{n/2}} e^{-(X-M)^T\Sigma^{-1}(X-M)} \tag{10}$$

where *n* represents the dimension of the vector *X*, *M* the mathematical expectation of the vector *X*, and *Σ* the covariance matrix of the feature vector. These values for both classes are obtained from the training set. For both classes, feature vectors are formed for classification.

B. *k nearest neighbors Classifier*

The most common issue in practice is information missing needed for classification based on hypothesis testing, so one may resort to non-parametric classification. One of the most popular methods is k nearest neighbors or knn. The algorithm classifies the observation point in relation to how the neighbors are classified. In the knn algorithm, k is a parameter that indicates the number of nearest samples involved in the classification [7], [21].

In Fig. 1, a new green sample that needs to be classified can be observed. The full circle in the figure represents the case when k=3: the neighbors are one blue square and two red triangles. Since there are more triangles, the green circle sample is assigned to the same class as the triangles.

However, if four is taken for k, the green circle will be classified into the blue squares class because there are more of them in the region. In conclusion, k is an essential parameter for successful classification [21]. Also, the success of the classification depends on which methods are used for defining what the nearest neighbors are. Some of the methods that are going to be considered in this paper are Euclidean, Chebyshev, Mahalanobis distance, and cosine similarity [22].

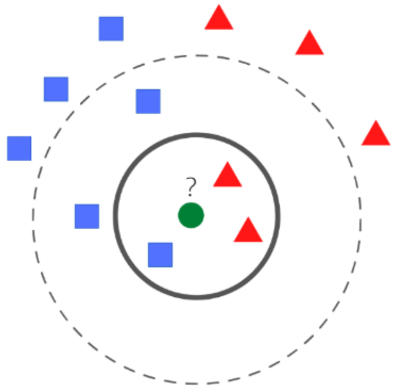

Fig. 1. Graphical representation of knn method, where green circle sample needs to be assigned to blue rectangle or red triangle class.

C. *Evaluation*

On proposed tumor segmentation steps and feature extraction, results of Bayes and knn classifiers will be shown, as well as the influence of feature vector dimension on results. The performance will be evaluated using confusion matrices and metrics described by (1)-(3).

The confusion matrix for this classification consists of two columns and two rows. Each row of the matrix represents the instances in an actual class, while each column represents the instances in a predicted class. Images with tumor tissue are labeled as 'positive', and healthy tissues are labeled as 'negative'. Four and six features are tested to observe the effects on the results using Bayes and knn classifiers. The effect of the distance type and different values of k (1 to 40) on accuracy have been also analyzed.

VII. Experimental Results

A. *Preliminary segmentation and feature inspection results*

Preliminary segmentation results are show in Fig. 2 and Fig. 3. It can be observed that tumor is correctly detected in Fig. 2 for cancer tissue example. In the image with healthy tissue in Fig.3, small regions which are not tumor are segmented as one (false) candidate. Segmented image that contains a tumor has larger white area extracted, as it can be seen in Fig. 2 and Fig. 3, so this can be also used as an effective feature that provides good separability.

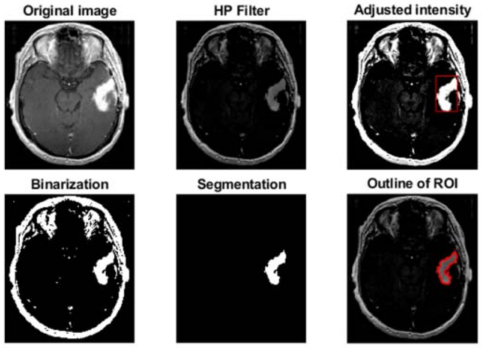

Fig.2. Results of segmentation for cancer tissue

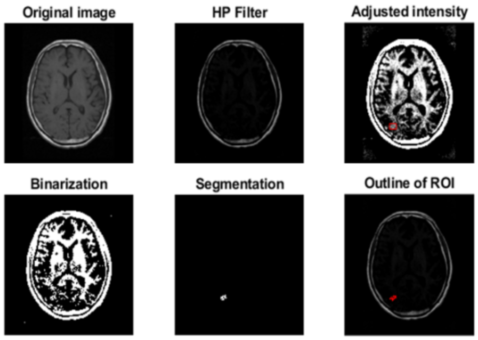

Fig. 3. Results of segmentation for normal tissue

Using the first four hand-crafted features may not be enough for high accuracy results. Adding two more features (feature 5 and feature 6) can improve results. From Fig.3 some of the separability inspection results can be seen.

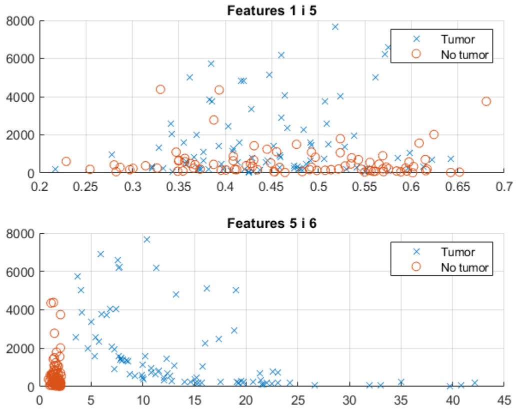

Fig. 4. Some of the separability inspection results of the selected features.

B. *Classification results using proposed method*

Classification results with only traditional or texture related features using Bayes classifier did not give expected results, as presented in Table I. Similarly is obtained for knn classifier and it is presented in Table II. For higher accuracy results, these classifiers require more information obtained through using additional features.

TABLE I
RESULTS OBTAINED WITH BAYES CLASSIFIER

Metric	TPR	TNR	BACC
Traditional four feature approach	71.4 %	82.6 %	75.8 %
The proposed method	96.6 %	100 %	98.3 %

TABLE II
RESULTS OBTAINED WITH KNN CLASSIFIER

Metric	TPR	TNR	BACC
Traditional four feature approach	89.6 %	65.5 %	77.6 %
The proposed method	100 %	96.5%	98.2 %

In the case of knn and four feature selection, the accuracy is highest for the Chebyshev distance, which is achieved for k = 39. This is illustrated in Fig. 5. These results are not suitable for practice and two additional features are applied to feature vector. The accuracy in the case with more features is highest for the cosine similarity and the best case is secured with lower number of neighbours (k = 3), which can be seen in Fig. 6, where smaller number of neighbours is a better choice.

For medical image classification, sensitivity is aimed to be high because of the need for all positives to be recognized correctly. Specificity should not be low because many false alarms are undesirable.

It is proven that only four texture related features cannot provide expected results, and adding two more features improves results. Feature vector still contains relatively small number of elements. In further experiments it is shown that the kNN algorithm with more features provides reliable results for all parameters, and compared to the Bayes classifier, it provides better sensitivity for both cases (with the lower and higher number of features), which can be seen in Table III. Slightly higher results in overall evaluation are obtained for knn for the proposed method.

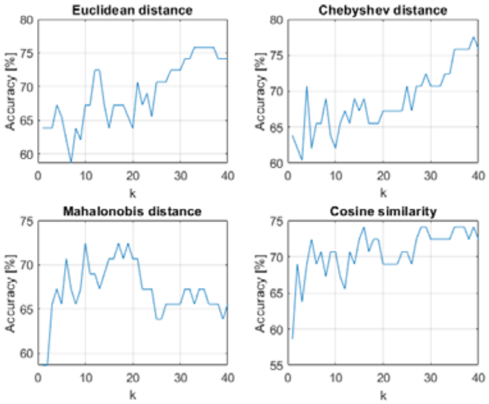

Fig. 5. Accuracy versus parameter k for four types of distances for four feature approach.

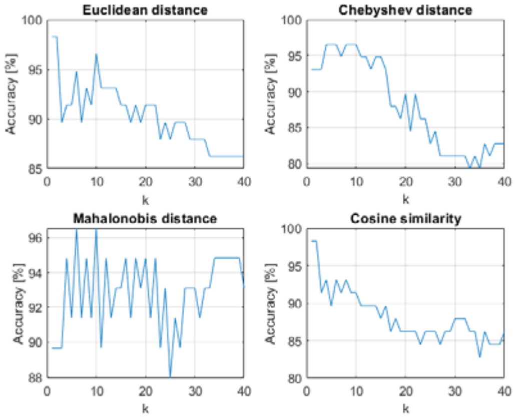

Fig. 6. Accuracy versus parameter k for four types of distances for proposed approach.

VIII. CONCLUSION

The proposed tumor segmentation in MRI head images provided an excellent base for analyzing the tumor classification methods. The proposed classification method gives surprisingly good results compared to the other methods based on machine learning tested on the same dataset. The advantage of the proposed method lies in less demanding hardware resources where traditional classification methods are used. A more diverse selection of features can further increase accuracy. Since the knn method stood out as a method with high accuracy, it is possible to test other selections and analyze the types of distances by which the classification is performed.

A possible improvement of the model would also be classifying different types of tumors and labeling more complex cases in brain tumor detection. Such data labeling for classification improvements would require help from the experts.

Acknowledgment

This work is written during the research supported and partially funded by the Ministry of Education, Science and Technological Development, Republic of Serbia. No. 2022/200103.

References

[1] H. Ritchie and M. Roser, “Causes of Death,” Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/causes-of-death', 2018.

[2] S. Webb, *The physics of medical imaging*, Taylor and Francis Group, Country: USA, 1988. https://doi.org/10.1201/9780367805838

[3] R.A. Sadek, “An improved MRI segmentation for atrophy assessment,” *International Journal of Computer Science Issues* (IJCSI) vol. 9, no. 3, pp. 569-574, 2012.

[4] M. Popović, *Digitalna obrada slike,* Akademska misao, Serbia, 2006.

[5] B. Efron, “Bayes' theorem in the 21st century,” *Science*, vol. 340, no. 6137, pp. 1177-1178, 2013.

[6] S.G. Kwak, and J.H. Kim, “Central limit theorem: the cornerstone of modern statistics,” *Korean journal of anesthesiology*, vol. 70, no. 2, pp. 144-156, 2017.

[7] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, “Knn model-based approach in classification,” OTM Confederated International Conferences “On the Move to Meaningful Internet Systems,” pp. 986-996, 2003. Springer, Berlin, Heidelberg.

[8] S. A. Medjahed, “A comparative study of feature extraction methods in images classification,” *International journal of image, graphics and signal processing* vol. 7, no. 3, pp. 16-23, 2015.

[9] P. Nair, and I. Nair, “Classification of medical image data using k nearest neighbor and finding the optimal k value,” *International journal of scientific & technology research*, vol. 9, no. 4, pp. 221-226, 2020.

[10] P.K. Bhagat, P. Choudhary, and K.M. Singh, “A comparative study for brain tumor detection in MRI images using texture features,” Chapter 13: In Advances in ubiquitous sensing applications for healthcare, *Sensors for Health Monitoring*, Academic Press, vol. 5, pp. 259-287, 2019. https://doi.org/10.1016/B978-0-12-819361-7.00013-0

[11] K. Fukunaga, *Introduction to statistical pattern recognition*, (2nd ed.), Academic Press Professional, Inc., USA. 1990.

[12] J. Jaidip, N. Patil, C. Kala, K. Pandey, A. Agarwal and A. Pradhan. "Statistical characterization of tissue images for detection and classification of cervical precancers." arXiv preprint arXiv:1112.4298 2011.

[13] K. K. Kumar, M. Devi T, and S. Maheswaran “An Efficient Method for Brain Tumor Detection Using Texture Features and SVM Classifier in MR Images*.” Asian Pacific journal of cancer prevention* :APJCP vol. 19, pp. 2789-2794, 26 Oct. 2018, doi: 10.22034/APJCP.2018.19.10.2789

[14] M. Domingo, and D. Filbert. "Classification of potential defects in automated inspection of aluminium castings using statistical pattern recognition," *8th European Conference on Non-Destructive Testing* (ECNDT 2002), pp.1-10, 2002.

[15] T.M. Hsieh, Y. M. Liu, CC Liao, F. Xiao, I-J. Chiang, J-M. Wong, "Automatic segmentation of meningioma from non-contrasted brain MRI integrating fuzzy clustering and region growing", BMC Med Inform Decis Mak 11, 54, 2011. https://doi.org/10.1186/1472-6947-11-54

[16] K. Sharma, A. Kaur, and S. Gujral, “Brain tumor detection based on machine learning algorithms,” *International Journal of Computer Applications* 103.1, pp. 7-11, 2014.

[17] R. Ranjbarzadeh, B. Kasgari, S. J. Ghoushchi, S. Anari, M. Naseri, M. Bendechache, "Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images", Sci Rep 11, 10930, 2021. https://doi.org/10.1038/s41598-021-90428-8

[18] A. Çinar, and M. Yildirim, "Detection of tumors on brain MRI images using the hybrid convolutional neural network architecture," *Medical hypotheses*, *139*, 109684, 2020.

[19] Brain MRI images for brain tumor detection dataset: https://www.kaggle.com/navoneel/brain-mri-images-for-brain-tumor-detection

[20] H. J. Baek, H. S. Kim, N. Kim, Y. J. Choi, Y. J. Kim. "Percent change of perfusion skewness and kurtosis: a potential imaging biomarker for early treatment response in patients with newly diagnosed glioblastomas." *Radiology* vol. 264, no. 3, pp. 834-843, 2012.

[21] I. K. Preeti Nair, “Classification of medical image data using k nearest neighbor and finding the optimal k value,” *International journal of scientific technology*, volume 9, 2020.

[22] R. Ehsani, and F. Drabløs, "Robust Distance Measures for k NN Classification of Cancer Data," *Cancer informatics*, *19*, 1176935120965542, 2020.

[23] M. A. Javid, S. A. Buzdar, "A novel computer aided diagnostic system for quantification of metabolites in brain cancer," *Biomedical Signal Processing and Control*, Volume 66,102401, 2021. https://doi.org/10.1016/j.bspc.2020.102401