Machine Learning Models for Predicting Breast Cancer Risk in Women Exposed to Blue Light from Digital Screens

Seyed Ali Reza Mortazavi[1], Sedigheh Tahmasebi[2], Hossein Parsaei[3,4]*, Abdorasoul Taleie[2], Mehdi Faraz[5], Abbas Rezaianzadeh[6], Atefeh Zamani[7], Ali Zamani[3], Seyed Mohammad Javad Mortazavi[3]

[1]MD, Student research committee, Shiraz University of Medical Sciences, Shiraz, Iran
[2]MD, Breast Cancer Research Center, Shiraz University of Medical Sciences, Shiraz, Iran
[3]PhD, Department of Medical Physics and Engineering, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran
[4]PhD, Shiraz Neuroscience Research Center, Shiraz University of Medical Sciences, Shiraz, Iran
[5]MSc, Department of Medical Physics and Engineering, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran
[6]PhD, Department of Epidemiology, Shiraz University of Medical Sciences, Shiraz, Iran
[7]PhD, Department of Statistics, Shiraz University of Medical Sciences, Shiraz, Iran

ABSTRACT

Background: Nowadays, there is a growing global concern over rapidly increasing screen time (smartphones, tablets, and computers). An accumulating body of evidence indicates that prolonged exposure to short-wavelength visible light (blue component) emitted from digital screens may cause cancer. The application of machine learning (ML) methods has significantly improved the accuracy of predictions in fields such as cancer susceptibility, recurrence, and survival.

Objective: To develop an ML model for predicting the risk of breast cancer in women via several parameters related to exposure to ionizing and non-ionizing radiation.

Material and Methods: In this analytical study, three ML models Random Forest (RF), Support Vector Machine (SVM), and Multi-Layer Perceptron Neural Network (MLPNN) were used to analyze data collected from 603 cases, including 309 breast cancer cases and 294 gender and age-matched controls. Standard face-to-face interviews were performed using a standard questionnaire for data collection.

Results: The examined models RF, SVM, and MLPNN performed well for correctly classifying cases with breast cancer and the healthy ones (mean sensitivity> 97.2%, mean specificity > 96.4%, and average accuracy > 97.1%).

Conclusion: Machine learning models can be used to effectively predict the risk of breast cancer via the history of exposure to ionizing and non-ionizing radiation (including blue light and screen time issues) parameters. The performance of the developed methods is encouraging; nevertheless, further investigation is required to confirm that machine learning techniques can diagnose breast cancer with relatively high accuracies automatically.

Keywords

Artificial Intelligence; Breast Cancer; Digital Screens; Screen Time; Visible Light; Blue Light; Prognosis Prediction; Smartphone; Circadian Clocks

Introduction

Accumulating evidence shows that circadian rhythm dysregulation is associated with adverse health effects such as susceptibility to bipolar spectrum disorders (BSDs) onset, hormone secretion imbalance, sleep problems, coronary heart attacks, depression,

*Corresponding author:
Hossein Parsaei
Department of Medical Physics and Engineering, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran
E-mail: hparsaei@sums.ac.ir

Received: 26 May 2021
Accepted: 2 August 2021

carcinoma, dysplasia, metabolic disorders, and neurodegenerative diseases [1-7]. As a symbol of today's digital life, screen time (the amount of time someone spends using devices such as a smartphone, video game console, tablet, laptop, computer, or television) is rapidly increasing around the world. In children and adolescents, prolonged digital screen time is associated with specific adverse health effects including bad self-reported health status, loneliness, obesity, irritability, low mood, impaired cognitive and socioemotional development to poor school performance [8-13]. Moreover, a strong link between screen time and sleep duration, in particular in children under 6 months of age was reported [14].

In adults, there is limited evidence for an association between a sedentary lifestyle and obesity. Moreover, the reported associations are not causal [15]. In addition, considering the widespread use of digital screens in adults, substantial evidence now shows that exposure to visible light emitted from digital screens at night can be associated with some adverse health effects and impaired performance through dysregulation of the circadian rhythms. Humans are evolved predominantly under yellow light (the wavelength of ~ 570 nanometers); however, the displays of tablets, laptops, and smartphones as well as other digital screens usually emit high levels of short-wavelength blue light. The spectral profile of the visible light emitted from the displays of mobile phones, tablets, and computers adversely alters circadian physiology, alertness, and levels of cognitive performance [16]. Exposure to digital screens during bedtime has been reported to be associated with increased sleep latency, decreased sleep duration, the inefficiency of sleep, and higher rates of daytime sleep disturbances [17]. Finally, previous researches reported that prolonged exposure to short-wavelength visible light (blue region) emitted from digital screens can be linked to cancer. Therefore, there is a growing global concern over rapidly increasing screen time. Given this consideration, developing a prediction model for the risk of breast cancer is important.

Machine learning methods are successful in analyzing medical data [18-22]. Due to the ability of these techniques in modeling complex data, their applications in analyzing medical data have gained a great deal of interest these days. In this analytical study, these techniques were used to analyze cohort data collected from 603 cases, including 309 breast cancer cases and 294 gender and age-matched controls. Particularly, three machine learning models Random Forest, Support Vector Machine, and Multi-Layer Perceptron Neural Network were examined to predict breast cancer in women with various levels of exposure to ionizing and non-ionizing radiation. The effect of the parameters was estimated via the relative importance index. Details are given in the following sections.

Material and Methods

This analytical study was objected to develop a model for predicting the risk of breast cancer in women using several parameters (predictors) related to exposure to ionizing and non-ionizing radiation.

Data Collection

Data was collected from 603 cases (309 breast cancer cases and 294 gender and age-matched healthy cases). The methods used for data collection and analysis are described in detail elsewhere [23]. In brief, cases and controls were age-matched in an age-decade category. They were also matched for a family history of breast cancer. To collect data, face-to-face interviews were performed using a standard questionnaire. Ethical approval was obtained for the study from the Shiraz University of Medical Sciences (SUMS) Ethics Committee. Multivariate analysis, chi-square, and Fisher's exact tests were used for data analysis; details of the analysis are presented in [23].

Data Preprocessing

In the first step, the outliers and the samples with missing values were removed from further analysis. Graphical plots and restricted variable boundaries were applied to determine outliers. Finally, numerical features were normalized using min-max normalization to be in the range of 0 and 1.

Classification Algorithms

This research was objected to developing a model for predicting breast cancer in women with different levels of exposure to ionizing and non-ionizing radiation (including blue light and screen time issues). Three classifier algorithms were explored for this purpose: 1) Support vector machine (SVM), 2) Random Forest (RF), and 3) Multilayer perceptron (MLP).

Support Vector Machines: The Support Vector Machines (SVM) algorithm creates a maximum margin separation hyperplane separating the classes [24]. The hyperplane is determined throughout the training process. In developing an SVM classifier, several user-defined parameters such as the cost parameter (C), the type of kernel function, and its parameters should be defined. In this work, we used internal cross-validation to determine these parameters. Ultimately, the value of C was set at 2.0, radial basis function was used as the kernel function and the value of sigma for the kernel was set to 0.5.

Random Forest: Random Forest (RF) algorithms operate by fusing several randomly created decision trees created via either bagging technique or feature randomness approach [25]. In bagging, training data for an individual tree is created by randomly sampling from the training set with replacement. For classification, the algorithm determines the class label of a test sample by aggregating the output of each generated decision tree via the majority voting approach in which a class label with the most votes is chosen. A critical parameter in developing an RF algorithm is the number of individual trees. In this study, this parameter was set at 100 that was determined via internal cross-validation.

Multi-Layer Perceptron: The Multi-Layer Perceptron (MLP), is a type of artificial neural network that, in general, has three layers of neurons: input, hidden, or output. The neurons on each layer operate on the weighted sum of the outputs of the neurons in the previous layer. In training an MLP, the values of neurons' weights are adjusted so that the classification error is minimized. Here, we used the back propagation algorithm to estimate the neurons' weights [26].

In developing an MLP-based classifier, besides the neurons' weights, the number of neurons in the hidden layer should be determined. Here, the best number of neurons was found using internal cross-validation by evaluating several models with the various numbers of neurons (two to twenty). The model with the lowest error was selected.

Statistics

The developed prediction models were quantitatively evaluated using the three common indices sensitivity, specificity, and accuracy defined as:

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \quad (1)$$

$$Specificity = \frac{TN}{TN + FP} \times 100 \quad (2)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (3)$$

where the parameters *TP*, *TN*, *FP*, and *FN* are defined as follows:

TP: Number of subjects with breast cancer that were correctly classified by the model.

TN: Number of healthy subjects that were correctly classified by the model.

FP: Number of healthy individuals that were incorrectly identified subject with breast cancer by the model.

FN: Number of cases with breast cancer that were incorrectly identified as healthy cases by the model.

Statistical comparison of the values obtained for each performance index was conducted using the Friedman test at a 5% significance level.

Results

The performances of the models in terms of the three evaluation indices used are summarized in Table 1. The values presented in the Table 1 were estimated using 10-fold cross-validation, leading to an unbiased evaluation of the models and ultimately an examination of their generalization ability. Graphical comparisons of the developed prediction models in terms of the three indices sensitivity, specificity, and accuracy are provided in Figure 1.

The relative importance measure for each feature estimated using the RF model is plotted in Figure 2. Gini importance (average impurity decrease) method was used to compute feature importance.

Discussion

This paper presents three models for predicting breast cancer via the history of exposure to ionizing and non-ionizing radiation (including blue light and screen time issues) parameters.

The results provided in Table 1 and Figure 1 show that the three models performed well for correctly classifying cases with breast cancer and the healthy ones as the average sensitivity > 97.2% and the average specificity > 96.4%. These values show that the developed machine learning models and the features used to represent each case were effective.

Statistical analysis of comparing the performance of the developed methods revealed that

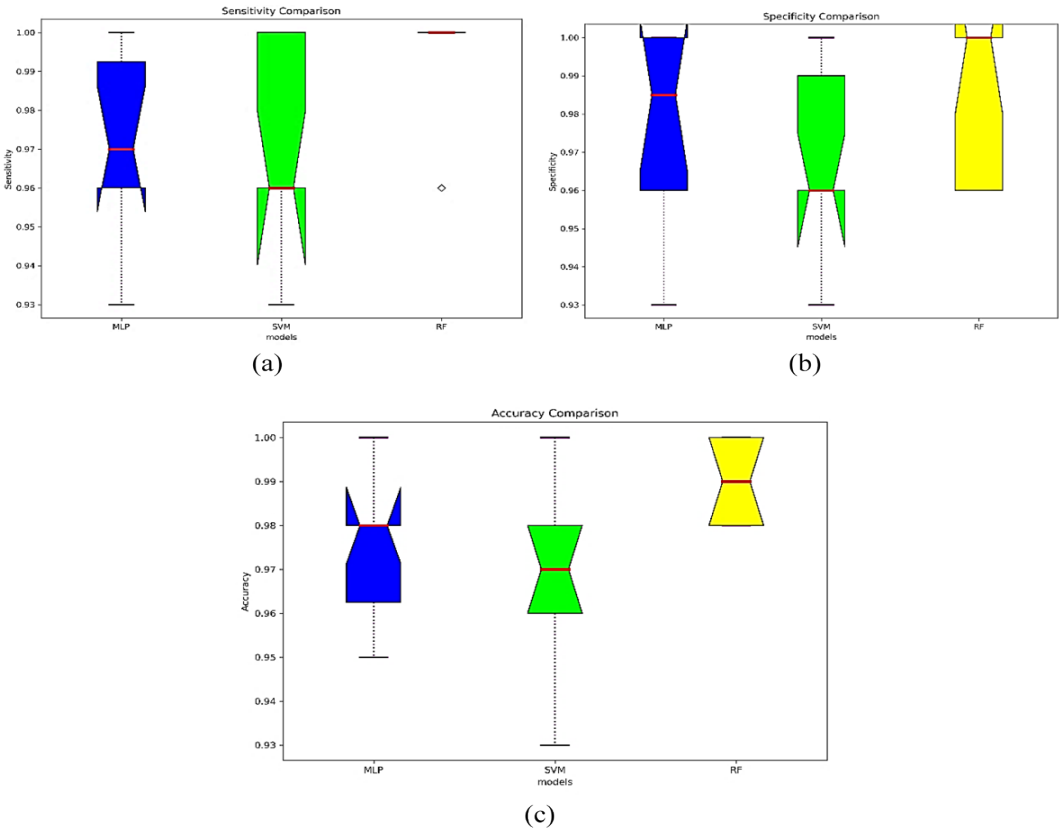

Figure 1: Graphical comparisons of the three developed prediction models in terms of sensitivity (a), specificity (b), and accuracy (c).

Figure 2: The relative variable importance for the parameters.

Table 1: Mean ± standard deviation for the performance indices of models developed for predicting breast cancer.

	Sensitivity (%)	Specificity (%)	Accuracy (%)
MLP	97.2±2.25	97.8±2.5	97.4±1.4
SVM	97.3±2.5	96.4±2.3	97.1±2.1
RF	99.6±1.3	98.4±2.1	99.0±1.1

MLP: Multilayer perceptron, SVM: Support Vector Machine, RF: Random Forest

there is no statistically significant difference between the performances of the developed models. Nevertheless, the median value for the sensitivity of the RF-based model is 100% which showed that the model categorized the majority of cancer cases correctly. The median value for the sensitivity of this model is 100% as well.

There are several reasons why the developed models performed well in this study. The predictors are effective; a key step in developing a machine learning model is selecting a set of good features that represent each example well. The features employed in this work were selected based on the literature and authors' previous study that revealed the effectiveness of these parameters on the risk of breast cancer. The data was collected carefully and precisely. Data is the backbone of developing machine learning models. As discussed in the previous section, data was collected using a standard questionnaire, cases and controls were age-matched in an age-decade category, and were matched for family history of breast cancer. Finally, the robustness of the models used could be another reason as they performed well in several medical applications [18-22].

The relative importance measure for each feature estimated using the RF model is plotted in Figure 2. Gini importance (average impurity decrease) method was used to compute feature importance. As shown, the most three important risk factors are "Exposure to BTS", "Number of Cellphones" and "Age"; this funding is according to the results of previous studies [27-29]. These overlaps with previous researches support the hypothesis that the developed model correctly identified relevant susceptibility for breast cancer due to exposer to ionizing and non-ionizing radiation. In terms of applications, the findings described here can be used to predict the risk of breast cancer based on exposure parameters (previous history of exposure to ionizing and non-ionizing radiation and particularly exposure to digital screens and blue light). Limitations of our work include limited sample size, retrospective design, and recall bias.

Conclusion

We described the development and applications of machine learning models for predicting the risk of breast cancer based on exposure

parameters (history of exposure to ionizing and non-ionizing radiation and particularly exposure to digital screens and blue light). The results showed that RF, SVM, and MLPNN are capable of discriminating subjects with breast cancer and the healthy cases with relatively high-performance values, mean sensitivity >97.2%, mean specificity > 96.4%, and average accuracy >97.1%. Further, we found that the three risk factors "Exposure to BTS", "Number of Cellphones" and "Age" are the most important.

Acknowledgment

This study is based on the findings of Dr. SAR Mortazavi's undergraduate thesis supported by Shiraz University of Medical Sciences (Grant No. 96-01-01-16623). We are extremely grateful to all the patients who participated in this study. We also gratefully acknowledge Ms. Absalan for her valuable assistance.

Authors' Contribution

SAR. Mortazavi, S. Tahmasebi, SMJ. Mortazavi, A. Taleie, A. Rezaianzadeh, A. Zamani were responsible to experimental design and data gathering. Atefeh Zamani conducted statistical analysis. M. Faraz and H. Parsaei developed Machine learning models and interpreted the results. SAR. Mortazavi, SMJ. Mortazavi, M. Faraz and H. Parsaei contributed to drafting the article. All authors provided critical feedback and approved the final manuscript.

Ethical Approval

The Shiraz University of Medical Sciences Institutional Review Board approved this study (Ethic number: IR.SUMS.MED.REC. 1398.057).

Informed consent

Obtaining informed consent from all patients participating in this clinical research was a must.

Funding

This study was funded by Shiraz University of Medical Sciences (Grant number 96-01-01-16623) awarded to SMJ. Mortazavi.

Conflict of Interest

None

References

1. Bracci M, Ciarapica V, Zabaleta ME, Tartaglione MF, Pirozzi S, et al. BRCA1 and BRCA2 Gene Expression: Diurnal Variability and Influence of Shift Work. *Cancers.* 2019;**11**(8):1146. doi: 10.3390/cancers11081146. PubMed PMID: 31405066. PubMed PMCID: PMC6721503.
2. Fonken LK, Nelson RJ. The effects of light at night on circadian clocks and metabolism. *Endocr Rev.* 2014;**35**(4):648-70. doi: 10.1210/er.2013-1051. PubMed PMID: 24673196.
3. Shanmugam V, Wafi A, Al-Taweel N, Büsselberg D. Disruption of circadian rhythm increases the risk of cancer, metabolic syndrome and cardiovascular disease. *J Local Glob Health Sci.* 2013;**2013**(1):3.
4. Sahar S, Sassone-Corsi P. Metabolism and cancer: the circadian clock connection. *Nat Rev Cancer.* 2009;**9**(12):886-96. doi: 10.1038/nrc2747. PubMed PMID: 19935677.
5. Giudice A, Crispo A, Grimaldi M, Polo A, Bimonte S, Capunzo M, et al. The Effect of Light Exposure at Night (LAN) on Carcinogenesis via Decreased Nocturnal Melatonin Synthesis. *Molecules.* 2018;**23**(6):1308. doi: 10.3390/molecules23061308. PubMed PMID: 29844288. PubMed PMCID: PMC6100442.
6. Alloy LB, Ng TH, Titone MK, Boland EM. Circadian Rhythm Dysregulation in Bipolar Spectrum Disorders. *Curr Psychiatry Rep.* 2017;**19**(4):21. doi: 10.1007/s11920-017-0772-z. PubMed PMID: 28321642. PubMed PMCID: PMC6661150.
7. Xie Y, Tang Q, Chen G, Xie M, Yu S, Zhao J, Chen L. New Insights Into the Circadian Rhythm and Its Related Diseases. *Front Physiol.* 2019;**10**:682. doi: 10.3389/fphys.2019.00682. PubMed PMID: 31293431. PubMed PMCID: PMC6603140.
8. Stiglic N, Viner RM. Effects of screentime on the health and well-being of children and adolescents: a systematic review of reviews. *BMJ Open.* 2019;**9**(1):e023191. doi: 10.1136/bmjopen-2018-023191. PubMed PMID: 30606703. PubMed PMCID: PMC6326346.
9. Klesges RC, Shelton ML, Klesges LM. Effects of television on metabolic rate: potential im-

plications for childhood obesity. *Pediatrics.* 1993;**91**(2):281-6. PubMed PMID: 8424001.

10. Iannotti RJ, Janssen I, Haug E, Kololo H, Annaheim B, Borraccino A; HBSC Physical Activity Focus Group. Interrelationships of adolescent physical activity, screen-based sedentary behaviour, and social and psychological health. *Int J Public Health.* 2009;**54**(Suppl 2):191-8. doi: 10.1007/s00038-009-5410-z. PubMed PMID: 19639256. PubMed PMCID: PMC2732761.
11. Marsh S, Ni Mhurchu C, Maddison R. The non-advertising effects of screen-based sedentary activities on acute eating behaviours in children, adolescents, and young adults. A systematic review. *Appetite.* 2013;**71**:259-73. doi: 10.1016/j.appet.2013.08.017. PubMed PMID: 24001394.
12. Wang H, Zhong J, Hu R, Fiona B, Yu M, Du H. Prevalence of high screen time and associated factors among students: a cross-sectional study in Zhejiang, China. *BMJ Open*. 2018;**8**(6):e021493. doi: 10.1136/bmjopen-2018-021493. PubMed PMID: 29921687. PubMed PMCID: PMC6009552.
13. Gentile DA, Berch ON, Choo H, Khoo A, Walsh DA. Bedroom media: One risk factor for development. *Dev Psychol*. 2017;**53**(12):2340-55. doi: 10.1037/dev0000399. PubMed PMID: 28945440.
14. Chen B, Van Dam RM, Tan CS, Chua HL, Wong PG, Bernard JY, Müller-Riemenschneider F. Screen viewing behavior and sleep duration among children aged 2 and below. *BMC Public Health.* 2019;**19**(1):59. doi: 10.1186/s12889-018-6385-6. PubMed PMID: 30642299. PubMed PMCID: PMC6332844.
15. Biddle SJH, García Bengoechea E, Pedisic Z, Bennie J, Vergeer I, Wiesner G. Screen Time, Other Sedentary Behaviours, and Obesity Risk in Adults: A Review of Reviews. *Curr Obes Rep.* 2017;**6**(2):134-47. doi: 10.1007/s13679-017-0256-9. PubMed PMID: 28421472.
16. Cajochen C, Frey S, Anders D, Späti J, Bues M, Pross A, Mager R, Wirz-Justice A, Stefani O. Evening exposure to a light-emitting diodes (LED)-backlit computer screen affects circadian physiology and cognitive performance. *J Appl Physiol (1985)*. 2011;**110**(5):1432-8. doi: 10.1152/japplphysiol.00165.2011. PubMed PMID: 21415172.
17. Krishnan B, Sanjeev RK, Latti RG. Quality of Sleep Among Bedtime Smartphone Users. *Int J Prev Med.* 2020;**11**:114. doi: 10.4103/ijpvm.IJPVM_266_19. PubMed PMID: 33088442. PubMed PMCID: PMC7554597.
18. Jo T, Nho K, Saykin AJ. Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data. *Front Aging Neurosci.* 2019;**11**:220. doi: 10.3389/fnagi.2019.00220. PubMed PMID: 31481890. PubMed PMCID: PMC6710444.
19. Sarica A, Cerasa A, Quattrone A. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. *Front Aging Neurosci.* 2017;**9**:329. doi: 10.3389/fnagi.2017.00329. PubMed PMID: 29056906. PubMed PMCID: PMC5635046.
20. Parsaei H, Faraz M, Mortazavi SMJ. A multilayer perceptron neural network–based model for predicting subjective health symptoms in people living in the vicinity of mobile phone base stations. *Ecopsychology.* 2017;**9**(2):99-105. doi: 10.1089/eco.2017.0011.
21. Aminsharifi A, Irani D, Tayebi S, Jafari Kafash T, Shabanian T, Parsaei H. Predicting the Postoperative Outcome of Percutaneous Nephrolithotomy with Machine Learning System: Software Validation and Comparative Analysis with Guy's Stone Score and the CROES Nomogram. *J Endourol*. 2020;**34**(6):692-9. doi: 10.1089/end.2019.0475. PubMed PMID: 31886708.
22. Mortazavi SMJ, Aminiazad F, Parsaei H, Mosleh-Shirazi MA. An artificial neural network-based model for predicting annual dose in healthcare workers occupationally exposed to different levels of ionizing radiation. *Radiat Prot Dosimetry*. 2020;**189**(1):98-105. doi: 10.1093/rpd/ncaa018. PubMed PMID: 32103272.
23. Mortazavi SAR. The Association of Screen Time and Female Breast Cancer - A Retrospective Case-Control Study [dissertation]. Shiraz: Shiraz University of Medical Sciences; 2021.
24. Cortes C, Vapnik V. Support-vector networks. *Machine Learning*.1995;**20**:273-97. doi: 10.1007/BF00994018.
25. Breiman L. Random Forests. *Machine Learning.* 2001;**45**:5-32. doi. 10.1023/A:1010933404324.
26. Haykin S. Neural Networks and Learning Machines. 3rd edition. New York: Pearson; 2008.
27. Mortazavi SMJ. Subjective Symptoms Related to GSM Radiation from Mobile Phone Base

Stations: a cross-sectional study. *J Biomed Phys Eng*. 2014;**4**(1):39-40. PubMed PMID: 25505767. PubMed PMCID: PMC4258853.

28. Ramirez-Vazquez R, Gonzalez-Rubio J, Arribas E, Najera A. Personal RF-EMF exposure from mobile phone base stations during temporary events. *Environ Res*. 2019;**175**:266-73. doi: 10.1016/j.envres.2019.05.033. PubMed PMID: 31146098.
29. Koppel T, Ahonen M, Carlberg M, Hedendahl LK, Hardell L. Radiofrequency radiation from nearby mobile phone base stations-a case comparison of one low and one high exposure apartment. *Oncol Lett*. 2019;**18**(5):5383-91. doi: 10.3892/ol.2019.10899. PubMed PMID: 31612047. PubMed PMCID: PMC6781513.