BioChemMed

Technium BioChemMed Vol. 2, Issue 4 pp.9-16 (2021)
ISSN: 2734-7990
www.techniumscience.com

Pre-Cancer Diagnosis via Gene Mutations Applied Ensemble Algorithms

Shadan Mohammed Jihad Abdalwahid [1], Sami Ismael [2], Shahab Wahhab Kareem [1, 3]

[1] Department of Technical Information Systems Engineering, Erbil Technical Engineering College, Erbil Polytechnic University, Erbil, Iraq. shadan.abdalwahid@epu.edu.iq

[2] Technical Institute of Bardarash, Duhok Polytechnic University, Duhok, Iraq. sami.hussein@dpu.edu.krd

[3] Information Technology Dept., College of Engineering and Computer Science, Lebanese French University.shahab.kareem@epu.edu.iq

Abstract: According to the current study, individuals with cancer who have a gene mutation have a bad prognosis. Young women with breast cancer had a poorer prognosis than older women, although it is unknown if the p53 gene mutation contributed to this. Due in part to the devastation of cancer, the appropriate technology may help cancer sufferers in regaining their lives. Researchers seek mutations in cancer-causing gene sequences to identify the precancerous stage. While genetic testing may be used to forecast some kinds of cancer, there is presently no effective technique for identifying all cancers caused by gene mutations. It is one of the most often discovered genetic anomalies in human cancer is a malfunction in the action of the protein P53. As a consequence, the Universal Mutation Database is used to identify gene mutations (UMDCell-line2010).

The issue is that, although many basic databases (for example, Excel format databases) exist that include datasets of TP53 gene mutations associated with disease (cancer), this huge database is incapable of detecting cancer. Thus, the purpose the objective of this study is to create an approach for data mining that utilizes a neural network to ascertain the pre-cancerous state. To begin, bioinformatics techniques such as BLAST, CLUSTALW, and NCBI were used to determine whether or not there were any malignant mutations; second, the proposed method was carried out in two stages: to begin, bioinformatics techniques such as BLAST, CLUSTALW, and NCBI were used to determine whether or not there were any malignant mutations; and third, the proposed method was carried out in two stages: to begin, bioinformatics techniques such as To begin, bioinformatics tools such as BLAST and CLUSTAL

Vote Algorithms were utilized to classify pre-cancer by malignant mutations in the disease's early stages. The writers teach and evaluate their subjects using a variety of situations, including cross-validation and percentages. This page contains a review of the algorithms discussed before.

Keyword: - Ensemble algorithm, Gene mutation, TP53, Cancer disease, Machine learning.

Technium BioChemMed Vol. 2, Issue 4 pp.9-16 (2021)
ISSN: 2734-7990
www.techniumscience.com

I. INTRODUCTION

Tumor suppressor gene mutations are one among the most popular commonly found mutations in human malignancies [1]. In human breast cancer, mutations in the p53 gene, as well as overexpression of the p53 protein, are commonly seen. [2]. Numerous studies have connected p53 gene mutations and protein accumulation to a poor prognosis; some have even shown that p53 mutations are a separate prognostic factor in breast cancer (p53 alterations). [1] [2] [3] [4] [5] [6] Young women are more prone to develop breast cancer than older women with breast cancer to have unfavorable histological characteristics and a bad prognosis [7] [8] [9] [10] [11] [12] [13] [14] As a result of these results, we postulated that p53 gene alterations might contribute to the development of breast cancer in young women. While the bulk of research has concentrated on the association between p53 mutations and the prognosis of breast cancer, just a few studies have examined the linkage between p53 mutations and breast cancer in its early stages. Also worth noting is the wide variation in the reported frequency, kind, and location of p53 gene mutations in human breast cancer. The purpose of this study aimed to investigate whether there is a correlation between the pattern of p53 gene alterations and the development of early-onset breast cancer in females. Mutations in the TP53 tumor suppressor gene are commonly found in human malignancies, making it one of the most frequently changed genes in the body [11]. According to [12] and [13], the TP53 gene mutation frequency in PCa is quite modest, about 30%. Furthermore, unlike in other malignancies such as bladder cancer, the prevalence of TP53 mutations in prostate tumor tissue does not rise statistically significantly when tumor grading and staging increase. Mutations in the TP53 gene have been found to impact cell proliferation activation, DNA repair suppression, and apoptosis induction [14] [15] [13] [18] [19]. As a consequence, it has been asserted that TP53 mutations promote tumor development [16] [17]. Overexpression of the p53 protein was associated with tumor progression in patients with PCa [18], with p53 overexpression, histological grading, and tumor stage all being significant prognostic factors for survival in univariate analysis, but only p53 overexpression remaining a significant independent predictor of survival in multivariate analysis [19]. Previous research discovered a low prevalence of TP53 mutations in benign prostatic hyperplasia, ranging between 16.5 and 19.0 percent [19] [20], with individuals who had mutations having a greater risk of developing PCa later in life than those who did not have mutations [19] [20]. Recently, it was proposed that TP53 plays a role in the control of PSA [15]. Patients with prostate cancer are detected at a disproportionately high incidence when their blood PSA level is less than 4.0 ng/l [21]. On the other hand, the PSA level is the most sensitive tumor marker for prostate cancer [22] [16]. After radical prostate cancer surgery, the PSA level decreases to less than 0.1 ng/l. [17] [9] A reduction in PSA of 0.2 g/l after curative PCa therapy is considered to be suggestive of tumor development. Before therapy, a high level of prostate-specific antigen (PSA) was also considered a risk factor for tumor development [23]. We provide results from a follow-up study in this communication. The purpose of this study is to investigate the effect of TP53 mutation status, pretreatment PSA level, age, tumor grading, and staging on tumor progression in patients with prostate cancer.

II. LITERATURE REVIEW

The authors presented their findings in [24]. Twenty TP53 mutations were discovered in Chinese breast cancer families using high-throughput next-generation sequencing, which was previously thought to be due to germline mutations. Early-onset, hormone receptor-positive breast cancer was found in the vast majority of TP53 carriers, as well as a strong family history of cancer in their families. 11 patients had a germline mutation, with 6 of them being de novo germline mutations, which was the highest number found. A further case was suspected to have been induced by chemotherapy or radiation because the patient had no significant family history of cancer and aberrant clonal expansion can commonly be associated with TP53 mutations, which are common in cancer. In addition, we have identified one case of mosaic LFS (Low Fatigue Syndrome). There have been two novel mutations identified in patients with early-onset ALS (c.524 547dup and c.529 546del).
TP53 is found to be frequently altered in patients with esophageal squamous cell carcinoma, according to [14]. (ESCC). However, the genetic landscape of the TP53 mutation and its consequences for patients

are still up in the air. Next-generation sequencing (NGS) was used to identify somatic mutations in the TP53 gene in 161 patients with resectable ESCC, and immunohistochemistry was used to confirm the mutations (IHC). Patients were divided into seven groups based on their TP53 mutations, which were further subdivided into “disruptive” and “non-disruptive” types based on the extent to which the mutation affected the encoded protein. The association between the TP53 mutation and clinicopathological characteristics as well as disease outcome was examined.

As reported in [25], there is now compelling evidence that mutations not only render the p53 tumor-suppressive functions ineffective but that they can also endow mutant proteins with novel activities in some cases. Different missense mutations in the tumor suppressor gene p53 may confer distinct activities, providing insight into the mutagenic events that contribute to tumor progression. We present a comprehensive overview of the mechanisms by which mutant p53 exerts its cell-damaging effects, with a particular emphasis on the rapidly expanding mutant p53 transcriptome, and discuss the biological and clinical consequences of mutant p53 gain of function.

As previously reported in [27], there is a statistically significant association between a lack of response to 5-fluorouracil or mitomycin and mutations affecting the L2/L3 domains of the p53 protein. As a result of our previous discovery that such mutations predict resistance to weekly doxorubicin, our findings suggest that mutations affecting this specific domain of the p53 protein may cause resistance to a variety of cytotoxic agents used in breast cancer treatment.

In [27], the authors discuss the significance of p53 mutations in some of the tumors listed above, intending to elucidate how p53 contributes to the progression of cancer. TP53 status as a prognostic marker and its role as a predictor of response to therapy is also discussed in this article. p53 function abnormalities are linked to the development of other non-neoplastic diseases, according to the evidence presented in this paper.

[27] investigated this issue by analyzing data from 1537 patients who had received intensive treatment as part of the German-Austrian AML study group's protocols. It was determined that TP53 mutations were classified according to their impact on protein structure, as well as according to the evolutionary action (EAp53) and relative fitness scores (RFS). A total of 108 TP53 mutations were found in 98/1537 (6.4 percent) of the patients. While there was no difference in overall survival between patients with low-risk and high-risk AML-specific RFS, there was a difference in overall survival between patients with low-risk and high-risk AML-specific RFS.

III. METHODOLOGY

Two distinct methods have been utilized to diagnose the precancerous stage. The first technique is used to assess whether or not an individual's DNA has cancer-causing mutations. It is utilized in the second approach, which detects mutations, to establish if a patient's gene mutation is linked with a particular illness (cancer), such as lung cancer, head, and neck cancer, breast cancer, or other kinds of cancer.

A- Bioinformatics Tools:

1) BLAST: The basic local alignment search tool (BLAST) is a helpful tool for aligning sequences. This software may be used to identify the classification and homolog of a query sequence by extracting comparable portions of an input protein (or DNA) sequence from protein (or DNA) databases.

2) CLUSTALW: The capacity to conduct multiple sequence alignment is a prerequisite for other bioinformatics tasks like phylogenetic analysis and structure prediction (MSA). Even though there are many ways for applying MSA, the CLUSTALW test is the first tool for determining whether or not a person has a harmful mutation. According to the hypothesis, "two proteins with radically differing amino acid sequences may nevertheless be physiologically identical (Homology)." CLUSTALW is used to detect whether a gene is mutated or not. The TP53 gene mutation significantly raises the risk of developing cancer in the body. CLUSTALW assigns a kind of sequence to the normal sequence of each gene (without mutation), while the individual's gene assigns a different type of sequence. After then, the alignment between them is evaluated to see whether there is a match.

B- Machine Learning approach (Ensemble algorithms):

While bioinformatics techniques are used to identify dangerous mutations in a person's genes (the first method), these tools are inadequate for forecasting the likelihood of getting malignancies linked with

the malignant mutation. As a result, the suggested method needs the second approach, which is based on ensemble algorithms, since the first stage's mutations must be categorized using machine learning techniques for the technique to be successful. Vote Class is the technique for merging classifiers. There are many combinations of probability estimates for categorization. If a base classifier is unable to accept instance weights, or if the instance weights are not uniform, the data will be resampled with replacement using the weights before being given to the base classifier. The algorithms' parameters are as follows:-
Seed: The seed for the random number generator to utilize.
pre-built classifiers: The serialized classifiers that should be included. When this classifier runs, it is possible to incorporate several serialized classifiers alongside those that are created from scratch. It is not necessary to incorporate pre-built classifiers in cross-validation since their models are static and do not change from the fold to fold.
Classifiers: The classifiers to be used as a starting point.
numDecimalPlaces: The number of decimal places to utilize in the model's output.
batchSize: If the batch prediction is used, this is the desired number of instances to handle. While more or fewer instances may be supplied, this allows implementations to define their desired batch size.
combination
The combination rule was utilized.
doNotPrintModels: In the output debug, omit to print the particular trees: When true is specified, the classifier may report extra information to the console.
If set to true, the capabilities of the classifier are not verified before the classifier is constructed (Use with caution to reduce runtime).

IV. EXPERIMENTAL RESULT

The authors provide the results of the stated algorithms in this part using the following criteria:-
1- Cross-Validation (CV): the authors used a variety of CV scenarios (5, 10, and 15) to determine the training and testing conditions. The outcome is summarized in Table 1.
2- Percentage Splitting: The authors also used a percentage split (66 percent, 75 percent, and 85 percent) to determine the training and testing scenarios. The outcome is given in Table 2.

Based on the result in Tables 1 and 2 it shows that the Percentage splitting is better when 85% is. Then the 75% of the splitting percentage is second best.

Table 1 Scenario of different cross-validation of the Vote algorithms

	Cross Validation =5	Cross Validation =10	Cross Validation =15
Correlation coefficient	-0.0669	-0.0918	-0.1083
Mean absolute error	16.8648	16.8653	16.8617
Root mean squared error	20.5966	20.5911	20.5886
Relative absolute error	100%	100%	100%
Root relative squared error	100%	100%	100%
Total Number of Instances	1438	1438	1438

Technium BioChemMed Vol. 2, Issue 4 pp.9-16 (2021)
ISSN: 2734-7990
www.techniumscience.com

Table 2 Scenario of different Percentage Splitting of the Vote algorithms

	Percentage 66%	Percentage 75%	Percentage 85%
Correlation coefficient	0	0	0
Mean absolute error	16.8103	17.1875	17.1913
Root mean squared error	20.8342	21.1173	21.2349
Relative absolute error	100%	100%	100%
Root relative squared error	100%	100%	100%
Total Number of Instances	489	359	216

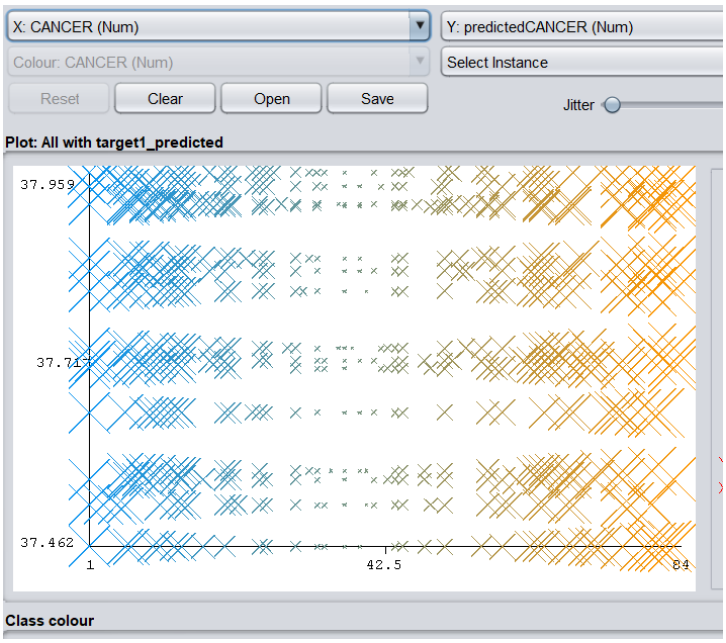

Figure 1 Regression of Vote structure of the CV=5

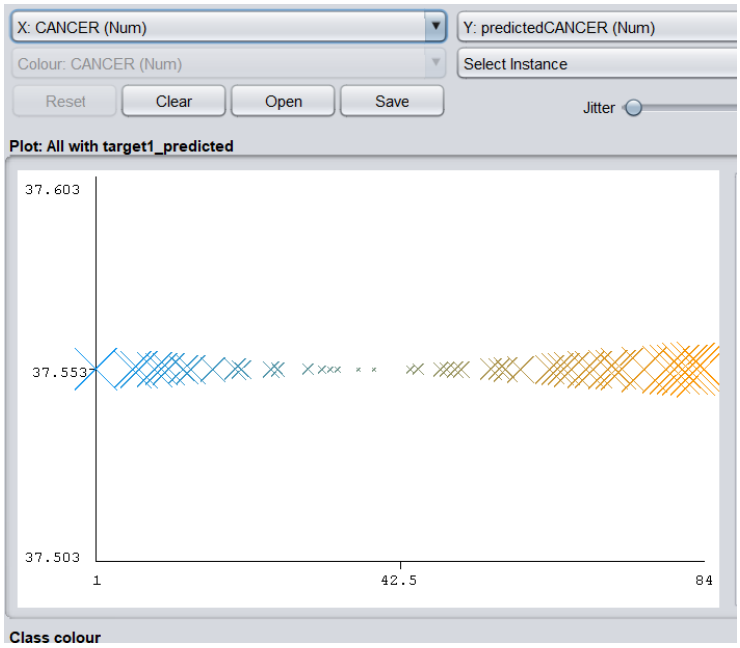

Figure 2 Regression of Vote structure of the Percentage Splitting =66%

V.CONCLUSION

If p53 is suppressed in normal tissues, it may be feasible to mitigate the apoptotic adverse effects of genotoxic cancer treatment. Alternatively, mt p53 may serve as a selective therapeutic target for cancer patients as a tumor-specific molecule. It is important to identify individuals at risk in the aging population, and it may be able to provide lifestyle recommendations based on the efficiency of p53 function.

A novel model of bioinformatics tools and a Machine Learning algorithm is utilized in this suggested approach to detect cancer in its early stages by predicting changed P53 gene expression levels in the blood. This method protects individuals against toxins and radiation while also allowing them to retain their self-control as they age. Additionally, an early diagnosis may aid in the creation of a patient-specific treatment plan. The methods described here are a unique and critical way for building a database for any local place, regardless of its size or breadth, anywhere in the globe. Due to the hereditary nature of cancer, this database would include information on each family's history, as well as information regarding genetic illnesses that have happened in the family. Then, a comprehensive database system capable of predicting genetic illness in its early stages will be created.

VI.REFERENCES

[1] Baugh, E., Ke, H., Levine, A., „Why are there hotspot mutations in the TP53 gene in human cancers?. Cell Death Differ 25, 154–160 (2018).," *Cell Death Differ 25, 154–160 (2018).*, pp. 154-160, 25 2018.

[2] Ecke TH, Schlechte HH, Hubsch A, Lenk SV, Schiemenz K, Rudolph BD and Miller K, „ TP53 Mutation in prostate needle biopsies – comparison with patients follow-up," *Anticancer Res,* pp. 4143-4148, 6 27 2007.

[3] Ava Kwong, Vivian Yvonne Shin, Cecilia Y. S. Ho, Chun Hang Au, Thomas P. Slavin, Jeffrey N. Weitzel, Tsun-Leung Chan and Edmond S. K. Ma, „Mutation screening of germline TP53 mutations in high-risk

Chinese breast cancer patients," *Kwong et al. BMC Cancer,* pp. https://doi.org/10.1186/s12885-020-07476-y, 20 2020.

[4] Amin Salih Mohammed, Shahab Wahhab Kareem, Ahmed khazal al azzawi and M. Sivaram, „Time Series Prediction Using SRE- NAR and SRE- ADALINE," *Jour of Adv Research in Dynamical & Control Systems,* 12 10 2018.

[5] Sardar M. R. K Al-Jumur, Shahab Wahhab Kareem, Raghad z.yousif, „Predicting temperature of Erbil city applying deep learning and neural network," *Indonesian Journal of Electrical Engineering and Computer Science,* pp. 944-952, 2 22 2021.

[6] Bonnier P, Romain S, Charpain C., „ Age as a prognostic factor in breast cancer: Relationship to pathologic and biologic features.," *Int J Cancer,* pp. 38-44, 1 62 1995.

[7] Evan H Baugh, Hua Ke, Arnold J Levine, Richard A Bonneau, and Chang S Chan, „Why are there hotspot mutations in the TP53 gene in human cancer," *Cell Death and Differentiation,* pp. 154-160, 25 2018.

[8] Hong Lai, Lin Lin, Mehrdad Nadji, Shanghai Lai, Edward Trapido & Lou Meng, „Mutations in the p53 Tumor Suppressor Gene and Early Onset Breast Cancer," *Cancer Biology & Therapy,* pp. 31-36, 1 1 2002.

[9] Sami H. Ismael, Shahab Wahhab Kareem, Firas H. Almukhtar, „Medical Image Classification Using Different Machine Learning Algorithms," *AL-Rafidain Journal of Computer Sciences and Mathematics,* pp. 135-147, 1 14 2020.

[10] Shahab Wahhab Kareem, Mehmet Cudi Okur, „Structure Learning of Bayesian Networks Using Elephant Swarm Water Search Algorithm," *International Journal of Swarm Intelligence Research,* pp. 19-30, 2 11 2020.

[11] Roojwan S Ismael, Rami S Youail, Shahab Wahhab Kareem, „Image encryption by using RC4 algorithm," *European Academic Research,* pp. 5833-5839, 2 4 2014.

[12] C. FA, „ Time to achieve a prostate-specific antigen nadir of 0.2 ng/ml after simultaneous irradiation for prostate cancer," *JUrol 168,* pp. 2434-2438, 8 16 2002.

[13] S. W. Kareem, NOVEL SWARM INTELLIGENCE ALGORITHMS FOR STRUCTURE LEARNING OF BAYESIAN NETWORKS AND A COMPARATIVE EVALUATION, GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES, Yasar University, 2020.

[14] Hainaut P, Hernandez T, Robinson A, Rodriguez-Tome P, Flores T, Hollstein M, Harris CC, and Montesano R, „ IARC Database of p53 gene mutations in human tumors and cell lines updated compilation, revised formats, and new visualization tools.," *Nucl Acids Res 26,* pp. 205-213, 26 1998.

[15] Hartmann A, Blazyk H, Kovach J, „The molecular epidemiology of p53 gene mutations in breast cancer.," *Trends Genet,* pp. 27-33, 13 1997.

[16] Huang, M., Jin, J., Zhang, F., Wu, Y., Xu, C., Ying, L., Su, D., „Non-disruptive mutation in TP53 DNA-binding domain is a beneficial factor of esophageal squamous cell carcinoma," *Annals of Translational Medicine,* 6 8 2020.

[17] Gurova KV Roklin OW, Krivokrysenko VI, Chumakov PM, Cohen MB, Feinstein E, and Gudkov AV, „Expression of prostate-specific antigen (PSA) is negatively regulated by p53," *Oncogene,* pp. 153-157, 21 2002.

[18] Shahab Kareem, Mehmet C Okur, „Bayesian Network Structure Learning Using Hybrid Bee Optimization and Greedy Search," in *Çukurova University,* Adana, Turkey, 2018.

[19] Shahab Wahhab Kareem, Raghad Zuhair Yousif, Shadan Mohammed Jihad Abdalwahid, „An approach for enhancing data confidentiality in Hadoop," *Indonesian Journal of Electrical Engineering and Computer Science,* p. pp. 1547~1555, 3 20 2020.

[20] JA Royds and B Iacopetta, „p53 and disease: when the guardian angel fails," *Cell Death and Differentiation,* pp. 1017-1026, 13 2006.

[21] M. JW, „Prostate-specific antigen only progression of prostate cancer," *J Urol,* pp. 1632-1642, 163 2000.

[22] Kuczyk MA, Serth J, Bokemeyer C, Machtens S, Minssen A, Bathke W, Hartmann J, and Jonas U, „ The prognostic value of p53 for long-term and recurrence-free survival following radical prostatectomy," *Eur J Cancer,* pp. 679-686, 34 1998.

[23] Walker RA, Lees E, Webb MB, Dearing SJ., „ Breast carcinomas occurring in young women (<35) are different," *Br J Cancer,* pp. 796-800, 11 74 1996.

[24] Hong Lai, Lin Lin, Mehrdad Nadji, Shenghan Lai, Edward Trapido & Lou Meng, „Mutations in the p53 Tumor Suppressor Gene and Early Onset Breast Cancer," *Cancer Biology & Therapy,* pp. 31-36, 2002.

[25] Thompson IM, Pauler DK, Goodman PJ, Tangem CM, Lucia MS, Parnes HL, Minasian LM, Ford LG, Lippman SM, Crawford ED, Crowley JJ, and Coltman CA Jr, „Prevalence of prostate cancer among men with a prostate-specific antigen level ≤4.0 ng per milliliter," *N Engl J Med,* pp. 2239-2246, 350 2004.

[26] Petr Dobes1*, Jan Podhorec, Oldrich Coufal, Andrea Jureckova, Katarina Petrakova, Borivoj Vojtesek, and Roman Hrstka, „Influence of mutation type on prognostic and predictive values of TP53 status in primary breast cancer patients," *ONCOLOGY REPORTS,* pp. 1695-1702, 32 2014.

[27] Oberpenning F, Hamm M, Schmid HP, Hertle L, and Semjonow A, „Radical prostatectomy: Survival outcome and correlation to prostate-specific antigen levels," *Anticancer Res,* pp. 4969-4972, 20 2000.

[28] Kwong, A., Shin, V.Y., Ho, C.Y.S., „Mutation screening of germline TP53 mutations in high-risk Chinese breast cancer patients," *BMC Cancer,* 1053 20 2020.

[29] William A. Freed-Pastor and Carol Prives, „Mutant p53: one name, many proteins," *GENES & DEVELOPMENT,* pp. 1268-1286, 26 2012.

[30] Stephanie Geisler, Anne-Lise Børresen-Dale, Hilde Johnsen, Turid Aas, Ju¨rgen Geisler, Lars Andreas Akslen, Gun Anker, and Per Eystein Lønning, „TP53 Gene Mutations Predict the Response to Neoadjuvant Treatment with 5 Fluorouracil and Mitomycin in Locally advanced Breast cancer," *Clinical Cancer Research,* p. 582–588, 5 9 2003.

[31] JA Royds and B Iacopetta, „p53 and disease: when the guardian angel fails," *Cell Death and Differentiation,* pp. 1017-1026, 13 2006.

[32] Dutta S, Pregartner G, Rücker FG, Heitzer E, Zebisch A, Bullinger L, Berghold A, Döhner K, Sill H. , „ Functional Classification of TP53 Mutations in Acute Myeloid Luekemia," *cancers,* p. 637, 3 12 2020.