Applying and an Assessment for Machine Learning Techniques for Classifying Cancer Via TP53 Gene Mutations

Dina Yousif Mikhail [1], Dr. Firas Husham Almukhtar [2*], Ali Hussein Yousif [1], Dr. Shahab Wahab Kareem [1]

Corresponding author: Dr. Firas Husham Almukhtar

[1*] Information System Engineering Department, Technical Engineering College, Erbil Polytechnic University, Erbil, Iraq.

[2]Information Technology Department, Catholic University in Erbil, KRG - Iraq.

F.almukhtar@cue.edu.krd

Abstract

Prognosis of mutations plays a vital role in the detection and effective prevention of cancers. Due to mutations in the TP53 gene Database, the tumor suppressor P53 is responsible for a substantial number of human malignancies. It is so hard the ability to accurately predict and diagnose cancer from elementary data (in excel file), therefor this research proposes a functional model of Machin learning and an Artificial Neural Network for classifying cancer caused by a codon mutation in the tumor protein P53. The bagging classifier and K-nearest neighbor's classifier mechanisms have been used for learning and testing the Neural Network to obtain the best accuracy of the proposed architecture. By picking (12) of the (53) TP53 gene database fields, machine learning algorithms are used to build two classification models for the bagging and K-nearest neighbor's classifier . To be clear, it is discovered that one of these 12 fields (gene location field) is missing from the UMD TP53 Mutation Database2010; as a result, it is added to the TP53 gene database for training and testing the neural network algorithms, as a way to classify cancer types. The proposed architecture's learning and testing results demonstrate that the bagging classifier algorithm outperforms K nearest neighbors in terms of accuracy and error rates

Keywords: Artificial Neural Network, Mutation, TP53 gene, Bagging Classifier, K nearest neighbors

Number: 10.14704/nq.2022.20.7.NQ33081 **Neuro Quantology 2022; 20(7):628-634**

Introduction

On account of the fact that it is caused by a variety of organic and physical actions and responses, cancer is considered to be one of the most well-known and complex infections of today(Dhannoon, 2017). A mutation in the p53 gene is found in roughly 50–60% of human malignancies (Baugh et al., 2018). Since its discovery in1979, the p53 tumor suppressor gene has been the subject of intense debate and research, despite nearly four decades of study. The p53 gene is found on chromosome 17's short arm (17p13), more precisely, it is located between base pairs 7,571,719 and 7,590,867 on chromosome 17 as shown in Fig. 1.(Mahmood et al., 2019).

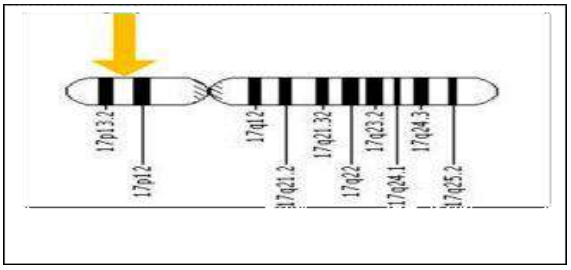

Figure 1. The TP53 gene's location on Chromosome 17

The p53 tumor suppressor regulates a wide range of cellular processes and viability, including DNA repair, cell cycle, metabolism, and apoptosis(Chen, 2016). Biologists use a programs in computer system to solve problems that are in the specialist tasks. The biologist field and

computer science field are taken together, to upholding the bioinformatics science. In addition, bioinformatics field is to plan ahead of time for biological systems and to analyze biological information, to learn more about how biological organisms work and analyze genetic information (Ghany & Yousif, 2016). Bioinformatics collects the computer sciences, statistics, information technology and mathematics. Bioinformatics takes benefits from synergies between computational and biological fields(Singh, 2015).

Computational biology can be used to predict how two proteins will interact. If the prediction is correct, then biological data from a wet lab experiment, including the proteins, should be examined using computational biology to determine how these proteins affect an organism's physiology (Mikhail, 2019).

In many fields of science, machine learning techniques are widely used, However, due to technical constraints, their application in medical literature is limited(Zhang, 2016). Machine learning (ML) is a set of algorithms that are specifically designed to predict the outcome of a situation. In addition to being easy to use, these machine learning techniques produce better results than traditional mathematical approaches to classification. In addition, the classification process relies on machine learning algorithms to achieve the best results and in the process of data analysis, machine learning techniques are used to accomplish a significant task. Depending on the input, classification techniques can predict a confident result. Using a collection of properties, the technique predicts the outcome, which is announced as either an object or a forecast characteristic(Ismael et al., 2020).

Therefore, a collection of properties is selected from UMD Cell-line-2010p53. The mutation database contains a large amount of data, which is stored in an excel sheet file, making the use of standard techniques difficult and such a massive number of data makes it unfeasible. As a result, Machin learning algorithms are utilized to make research and education procedures easier and more efficient. Finally, various machine learning methods, such as the bagging classifier and the K-nearest neighbors' classifier, are used to classify cancer types individually.

Related work:

Ayad G. Ismaeel, with Dina Y. Mikhail in (2016), try to Classify Cancers using neural networks. The proposed method was divided into two parts: first, bioinformatics tools like as CLUSTALW and BLAST were used to determine whether or not there were any malignant mutations. The second step was to use a neural network to mine data. Based on the Feed Forward Back Propagation technique, The UMD Cell line 2010 database was utilized to select 12 of the 53 TP53 gene database fields; the following data training rate (1) and Mean Square Error (MSE) were obtained as a result of the data training (0.00000000000001). This strategy was also put to the test with a cancer-causing the codon mutation 155 (ACC to CCC), which results in Head and Neck SCC Cancer(Ghany & Yousif, 2016).

Dina Y. Mikhail in (2019), proposed Pre-cancer Diagnosis method Using Bioinformatics and Neural Network. To find a data mining method that could diagnose pre-cancer, researchers used a neural network algorithm. Bioinformatics tools such as BLAST, CLUSTALW, and NCBI were used in the first step to determine whether a person's gene sequence contained malignant mutations. In the second step, Techniques of data-mining were utilized to categorize pre-cancer via mutations that cause malignancy at precocious stages utilizing one neural network for each of the three sub-datasets of forward back propagation. Furthermore, it provides training rate 1 with performance (MSE) of 5.82E-12, (9.99E-12), and (9.98E-12)(Mikhail, 2019).

Deeman Y. Mahmood, with Ayad G. Ismaeel and Abbas H. Taqi in (2019Method of Mining Proposed for Detecting a Mutant Codon 248 in TP53 Is Responsible for Cancer and Pre-Cancer. To forecast Codon mutations cause cancer and pre-cancer. (Hundreds of mutations occur in each codon, resulting in cancers), the mining technique's functional model and Artificial Neural Network was created, and this approach was applied to the mutability of hotspot codon 248 (exon7), CGG. For training and testing the proposed architecture, the Quick Propagation mechanism had been used. In this study, mutant codon 248's cancer and premalignant disease (pre-cancer) prognosis was predicted using a neural network. The classifier and Neural Network were built using Alyuda NeuroIntellegence, which is a software for neural

network simulation. It was found that the proposed architecture was 99.97 % accurate in training, 100 % accurate in validation and 99.85 % accurate in testing(Mahmood et al., 2019).

Problem Formulation (Equations and Variables)

Dataset and Pre Processing

Protein datasets excel genomes is example of primitive databases. One of these datasets is the Diseases are caused by mutations in the TP53 gene and tumor protein (cancers). In this study, bio-database (UMD_Cell_line_2010) of Cancer protein P53, with a large number of mutations (TP53), has been employed (Excel file form). As a result of the large number of fields in the P53 dataset as well as the need for data quality assurance, the following processes in data mining selection and preprocessing are required:

1. Data Selection: Due to the fact that the research concentrated on all codon mutations, and this domain's records will be the target data set. This stage is also considered a reduction of the dataset on the level of records.

2. Data Cleaning: Missing values are removed from the selected dataset, along with noise and inconsistencies in data, during this step.

3. Data Normalization: Data Normalization is the process of selecting and expressing data so that it conforms with a defined set of rules and constraints, or the rules of a particular database management system or data warehouse. Data normalization can also refer to the result produced by such processes. Many IT professionals and data analysts are concerned about irrelevant information in their source systems, which can lead to inconsistent reporting from different transactions for example(Russell, 2014).

4. Choosing a Feature Subset: The procedure for selecting features is used to identify non-profitable input fields and do not contribute significantly to the performance of the System. IT includes 53 fields (columns) and 1447 records (rows) in the UMD TP53 Mutation Database, and most of them are irrelevant to the prediction function and require an extended period of time to compute if all features are considered(De Souto et al., 2008).

According to their importance, the following features were selected to be part of the model's input set: Table I.

Table 1. Mutation fields identification.

1-Mutation position	The p53 cDNA is used as a reference for nucleotide position (1 is the A of the start ATG)
2-Codon	Placement of the codon (from 1 to 393)
3-WT codon	Normal codon base sequence
4-Mutant codon	Codons that have been mutated.
5-WT AA	Amino acid of the wild type
6-Mutant AA	Amino acid mutant.
7-Event	G>C (base change from G to C) is an example of a mutation event
8-Mutation type	One missense mutation (B), one nucleotide insertion or deletion (F), two nucleotide deletions (D) and two nucleotides insertion (I), and one tandem mutation (T).
9-Complexity	SM: Tumor with one mutational event; DMU (Double Mutation Unknown): Tumor with two p53 mutations, however their location is uncertain.; MM (Multiple Mutation): Multiple p53 mutations in the same cancer. DMD (Double Mutation Different Allele): In the same tumor, two p53 mutations occurred on two separate p53 alleles.
10-Type	Ts: Transition (a pyrimidine (C or T) is replaced with another pyrimidine, or a purine (A or G) is replaced with another purine). Tv: Transversion (a transversion mutation occurs when a pyrimidine is converted to a purine or vice versa).Fr: Frameshift mutations (deletion / insertion); Ts/Ts: two nucleotides in the same

	codon are targeted by a mutation. InF: this gene's deletion or insertion has no effect on the p53 protein's open reading frame.
11-Gene location	The location of a gene in the human body, which is obtained to test. This field does not exist in UMD p53 mutation database, it has been added in this proposed method to obtain accurate classification results and to classify specific cancer in specific cancer classification test.
12-Cancer	Cancer name

Proposed Mining Method

Data on TP53 gene mutations must be classified into various categories in order to accurately determine the type of cancer. There will be two machine learning algorithms discussed, and the mathematical model behind them will be described in the following manner:

A. Bagging classifier: It is a technique for the purpose of improving machine learning algorithms' outcomes that classify data. "Bootstrap aggregating" was the term used by Leo Breiman to describe this technique(Kotsiantis et al., 2005). In the classification state, a classification algorithm generates a classifier H: D$A\{-1,1\}$ based on the principle of a foundation set of example information D. In relation to the training set's qualifications, the bagging technique constructs a series of classifiers Hm, m=1,.., M. These classifiers are combined to create a composite classifier(Machová et al., 2006).

B. K-nearest neighbor's classifier: An unlabeled observation can be classified by assigning it to the most similar labeled example. For both the training and test datasets, characteristics of observations are gathered. For instance, When it comes to fruits, vegetables, and grains, their crunchiness and sweetness can be used to distinguish them as shown in Fig.(2)(Zhang, 2016).

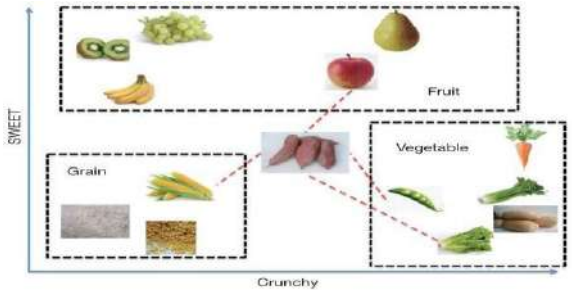

Figure 2 depicts the operation of the k-nearest neighbor's algorithm.

In this case, we're trying to figure out Sweet potatoes fall into which category? In this case, 4 nearby types of food are chosen: apple, green bean, lettuce, and corn. Sweet potato has been assigned to the vegetable class because it receives the majority of the votes. Sweet potato has been set up to the vegetable class because it receives the most votes. As can be seen, kNN's central idea is simple to grasp.

In the above example, there are two crucial ideas. To begin, we'll look at a method for determining the distance between sweet potatoes and other foods. It utilizes Euclidean distance by default, which may be determined using the equation below (1).

$$D(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$
……..(1)

Where p and q are the parameters in comparison with n features.

Distances can also be calculated by using other techniques, such as the Manhattan distance (2,3).

$\text{Sen}_i = TP_i/(TP_i + FN_i)$……..(2)

$Sp_i = TN_i/(TN_i + FP_i)$……..(3)

in which TP stands for the true positive, TN for the true negative, FP for the false positive, and FN for the false negative. The subscript I stands for "category" in this case.

In addition, k determines for the kNN algorithm how many neighbors will be selected. KNN algorithm's diagnostic performance is heavily influenced by k's choice. However, it increases the danger of overlooking minor but significant patterns due to a large k. It is important striking a balance between over-and under-adapting when choosing k(Zhang, 2016). It has been suggested by some authors that you should set k equal to the cube the square root of the training dataset's number of observations (Ghatak, 2017).

Experiment results

The Tp53 data sets are utilized in this studies are obtained from the TP53 website's UMD Cell line 2010, which is a contemporary and extensive accessible at http://p53.free.fr/Database/p53 MUT MAT.html(France Database of TP53 Gene, n.d.). The BPNN is trained and tested on (12) fields in this paper, while (11) fields are selected from a data base to be used in the training and testing. Aim for accuracy and efficiency cancer classification results, a column named (gene location field) is a new addition to the (11) fields are chosen. MATLAB program is utilized for neural networks since it has a lot of functions. The classification of cancer-causing mutations is accomplished successfully using two structures one is the bagging classifier and the second is K-nearest neighbor's classifier to produce an optimal classifier for cancer classification.

It's important to know what criteria (Correlation coefficient, Root-mean-square deviation, Mean absolute error) were used to produce the output before presenting and discussing it. Correlation coefficient is the first criterion, which measures the strength of a linear correlation between two variables. The range of values for this criterion is 0 to 1. When the absolute value of a relationship is close to1, the relationship is stronger. Null means that there is no correlation between the variables(Asuero et al., 2006).

Because MAE is a statistical fault measurement, it calculates the average distance between two real data (Z_t) points and their fitted predicted data (Z_t') points, respectively. The average of the absolute errors is used to calculate MAE, which is a good choice when there's a direct relationship between error cost and error magnitude. MAE is provided by:

$$\text{MAE} = \frac{1}{N}\sum_{t}^{N} |Z'_t - Z_t| = \frac{1}{N}\sum_{t}^{N} |e_t| \quad \text{.....(4)}$$

Assume that ($e_{1,t}$, $e_{2,t}$), t = 1,2, … m are the h-step out-of-sample forecast errors of models 1 and2, respectively, Using The loss differential between the two models using MAE as a measure of prediction loss is $dt = |e_{1t}| - |e_{2t}|$, t = 1,2, … m(Boiroju et al., 2011). The Root Mean Square Deviation (RMSD) (also known as the Root Mean Square Error (RMSE)) is used to calculate the differences between a model's predicted values and the actual model's values. These single deviations are referred to as residuals, and the RMSD works to aggregate them all into a single measure of predictive power. The RMSD of the prediction model with respect to the estimated variable Xmodel is defined as the square root of the mean squared error:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{'} (X_{obs,i} - X_{model,i})^2}{n}} \quad (5)$$

So that X_{obs} represents observed values and Xmodel represents modeled values at time/place i. The RMSD values can be utilized to tell apart between model performances during a calibration period, there will be a validation period where you can compare the performance of each model. As a result, the consequence of each error on RMSD is highly dependent on the size of the squared error; There is a disproportional impact on RMSD when there are large errors (Boiroju et al., 2011).

Relative Absolute Error (RSE) is the difference between the absolute deviation acquired from the prediction model and the one obtained by directly speculating the training sample. Prediction accuracy is inversely proportional to this factor. Maximum accuracy can be achieved by minimizing RSE:

$$RSE = \frac{\sum_{i=1}^{n} |f_i - y_i|}{\sum_{i=1}^{n} |fr_i - y_i|'} \quad \text{.....(6)}$$

To determine the root-relative squared error, use the following formula:

$$RRSE = \frac{\sum_{i=1}^{n} |f_i - y_i|^2}{\sum_{i=1}^{n} |f_i' - y_1|^2} \quad \text{.....(7)}$$

The RRSE is inversely proportional to the prediction accuracy. When the RRSE is as low as possible, the forecast accuracy can be as high as possible(Hoff et al., 2012).

Discussion results

According to Table II, the results of the bagging and k-nearest-neighbor classification algorithms are presented. Using the classification algorithms, we can determine the type of cancer.

In addition, the dataset is partitioned into two subsets: a training set and a test set.

1. Training Set = 75%, 85%
2. Testing Set = 25%, 15%

Table 2. Evaluation of cancer classification algorithms

	Bagging classifier	K nearest neighbor
Percentage split 85%		
Correlation coefficient	0.9655	0.5347
Mean absolute error	2.05	12.5537
Root mean squared error	5.6179	17.9209
Relative absolute error	11.9249 %	74.4506 %
Root relative squared error	26.4558 %	87.0429 %
Percentage split 75%		
Correlation coefficient	0.9566	0.5118
Mean absolute error	2.1575	12.8851
Root mean squared error	6.2189	19.9476
Relative absolute error	12.5527 %	74.9683 %
Root relative squared error	29.4491 %	94.4608 %

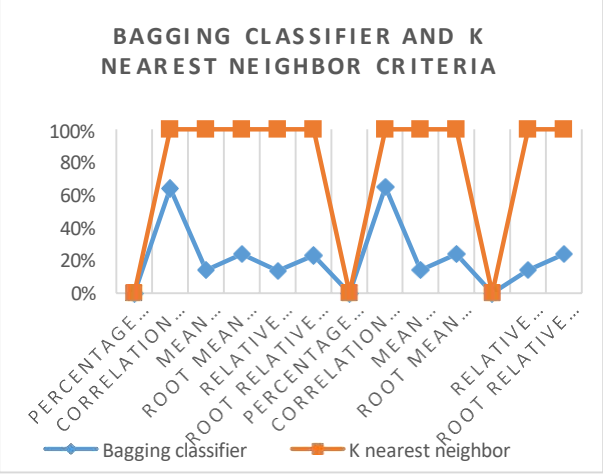

Figure 3. Bagging classifier and K nearest neighbor criteria

From the table and figure above (Table II– Fig 3), we can conclude that the bagging classifier is better than the K nearest neighbor classifier.

Conclusions

1. In this paper, two algorithms of machine learning are learned and tested separately to propose a system that automatically identifies and classifies the kinds of cancers via mutations in Tp53 gene.
2. The results of each algorithm are compared to select the best one based on various criteria. From the criteria's results, bagging classifier is better than K nearest neighbor classifier is concluded for classification the kinds of cancer.
3. The goal from this research using an extensive data set to find useful information, such as TP53 (tumor protein P53) biodata termed UMD mutations US, and categorize cancer by predicting mutated P53 genes.

References

Asuero, A. G., Sayago, A., & González, A. G. (2006). The correlation coefficient: An overview. *Critical Reviews in Analytical Chemistry*, *36*(1), 41–59. https://doi.org/10.1080/10408340500526766

Baugh, E. H., Ke, H., Levine, A. J., Bonneau, R. A., & Chan, C. S. (2018). Why are there hotspot mutations in the TP53 gene in human cancers? *Cell Death and Differentiation*, *25*(1), 154–160. https://doi.org/10.1038/cdd.2017.180

Boiroju, N. K., Yerukala, R., Venugopala Rao, M., & Krishna Reddy, M. (2011). A bootstrap test for equality of mean absolute errors. *ARPN Journal of Engineering and Applied Sciences*, *6*(5), 9–11.

Chen, J. (2016). The Cell-Cycle Arrest and Apoptotic Functions of p53 in Tumor Initiation and Progression. *Cold Spring Harbor Perspectives in Medicine*, *6*(3), a026104. https://doi.org/10.1101/cshperspect.a026104

De Souto, M. C. P., De Araujo, D. S. A., Costa, I. G., Soares, R. G. F., Ludermir, T. B., & Schliep, A. (2008). Comparative study on normalization procedures for cluster analysis of gene expression datasets. *Proceedings of the International Joint Conference on Neural Networks*, 2792–2798. https://doi.org/10.1109/IJCNN.2008.4634191

Dhannoon, B. N. (2017). *In Silico Molecular Classification of Breast and Prostate Cancers using Back Propagation Neural Network*. *September*. https://doi.org/10.7537/marscbj070317.01

France database of TP53 gene. (n.d.).

Ghany, A., & Yousif, D. (2016). Effective Data Mining Technique for Classification Cancers via Mutations in Gene using Neural Network. *International Journal of Advanced Computer Science and Applications*, *7*(7), 69–76. https://doi.org/10.14569/ijacsa.2016.070710

Ghatak, A. (2017). Machine Learning with R. In *Machine Learning with R*. https://doi.org/10.1007/978-981-10-6808-9

Hoff, T. E., Perez, R., Kleissl, J., Renne, D., & Stein, J. S. (2012). Reporting of irradiance model relative errors. *World Renewable Energy Forum, WREF 2012, Including World Renewable Energy Congress XII and Colorado Renewable Energy Society (CRES) Annual Conferen*, *2*(January 2015), 904–909.

Ismael, S. H., Kareem, S. W., & Almukhtar, F. H. (2020). Medical Image Classification Using Different Machine Learning Algorithms. *Mosul University*, *14*(1), 135–147. https://doi.org/10.33899/CSMJ.2020.164682

Kotsiantis, S. B., Tsekouras, G. E., & Pintelas, P. E. (2005). Bagging model trees for classification problems. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3746 LNCS*(January), 328–337. https://doi.org/10.1007/11573036_31

Machová, K., Barčák, F., & Bednár, P. (2006). A bagging method using decision trees in the role of base classifiers. *Acta Polytechnica Hungarica*, *3*(2), 121–132.

Mahmood, D. Y., Ismaeel, A. G., & Taqi, A. H. (2019). Mining method for cancer and pre-cancer detection caused by mutant codon 248 in TP53. *Periodicals of Engineering and Natural Sciences*, *7*(2), 522–533. https://doi.org/10.21533/pen.v7i2.546

Mikhail, D. Y. (2019). Pre-cancer Diagnosis via TP53 Gene Mutations by Using Bioinformatics Neural Network. *Proceedings of the 5th International Engineering Conference, IEC 2019*, 136–141. https://doi.org/10.1109/IEC47844.2019.8950565

Russell, G. (2014). *Normalisation. Database eLearning*. N.d. Retrived.

Singh, G. B. (2015). Fundamentals of Bioinformatics and Computational Biology. In *Methods in Molecular Biology* (Vol. 6, Issue Chapter 4). springer.

Zhang, Z. (2016). Introduction to machine learning: K-nearest neighbors. *Annals of Translational Medicine*, *4*(11). https://doi.org/10.21037/atm.2016.03.37