RESEARCH ARTICLE

Editorial Process: Submission:03/16/2022 Acceptance:07/23/2022

A Comparative Evaluation of Cancer Classification via *TP53* Gene Mutations Using Machin Learning

Dina Yousif Mikhail[1]*, Firas H. Al-Mukhtar[2], Shahab Wahab Kareem[3]

Abstract

Objective: Cancer is one of the horrendous diseases. Classifying cancer is founded on identifying cancer-causing mutations in gene sequences. Although genetic analysis can predict certain types of cancer, there is currently no effective method for predicting cancers. Therefore, the purpose of this paper is to predict the cancer types and to find a data mining technique that uses two different machine learning algorithms for classifying cancer. Moreover, earlier detection of the mutated tumor protein P53 gene can predict treatment and gene therapy techniques. **Methods:** (UMD-2010) the Universal Mutation Database is used to diagnose mutations in genes. The challenge, however, is that the database very basic. Besides, it is an excel format database. Due to its limitations, the data base cannot be used to classify cancer. In addition, bioinformatics techniques such as pairwise alignment and BLAST are used, followed by machine learning algorithms that use neural network algorithms to classify cancer based on malignant mutations in the *TP53* gene, by selecting (12) out of (53) database fields for the *TP53* gene database in the second stage. It should be noted that the (UMDCell-line2010) database does not have one of these twelve fields (Field of gene locus). **Result:** As a Utilizing MLP and SVM for training and testing a set number of fields, the Machin learning methods were found to be an effective way to classify cancers. Where the Relative Absolute Error for MLP and SVM is 83.6005 % ,65.6605 %, the accuracy is 90 %, 93.7% respectively. **Conclusion:** Following the learning and testing stages, the mean absolute error (MAE), used to measure the errors was found in the SVM less than the (MAE) in MLP algorithm. we can conclude that using SVM is considered better than the MLP algorithm because the accuracy in SVM better than the accuracy of MLP.

Keywords: Classification- cancer- Neural Networks- bioinformatics- Machin learning

Asian Pac J Cancer Prev, **23 (7)**, 2459-2467

Introduction

Cancer is one of the major causes of death worldwide, with 9.6 million cases in 2018. In underdeveloped countries, particularly unfavourable effects are expected (Abdel-Razeq et al., 2015; M-amen et al., 2022). In the past decade, complexity has been detected in human malignancies due to an explosion in the genomic sequence and molecular data (Balmain et al., 2003). This gene is responsible 50% of all sporadic human cancers, and mutations in this gene expose its holders to the possibility of developing cancer risk throughout their life, which further clarifies its role as a tumor suppressor gene(Pitolli et al., 2019).The growth in biological data is considered as a reason for releasing solutions in many domains of computational bioinformatics. Developing a computer system is important to solve our life's problems more quickly than other systems, and, as it can be noted, biologists use computers to solve many problems. It is a combination of information technology and biology that has merged to form bioinformatics. It is a field that combines the disciplines of computer science, mathematics, and information technology that combines these disciplines. It is the process of determining genetic information and it's analysis, where data mining is a way to discover useful patterns in large databases by using intelligent techniques and algorithms (Mikhail, 2019; Neamatollahi et al., 2020). Data mining provides strong methods and techniques for different fields involving Bioinformatics(Francis Bk). Data mining in the field of cancer can assist in providing physicians with some of the information and knowledge required for accurate prediction of breast cancer recurrence and better decision-making(Pei-Tse Yang and Jia-Lien; Mosayebi A, 2020). Classification is one of the methods in data mining for classifying a specific group of items into targeted groups. Some of the most common types of classification methods are the decision tree and Bayesian networks, as well as the k-nearest neighbour and support vector machines (Neelamegam and Ramaraj, 2013; Oluwaseun and Chaubey, 2019).

*[1]Information System Engineering Department, Technical Engineering College, Erbil Polytechnic University, Erbil, Iraq. [2]Department of Information Technology & Computer science, Catholic University in Erbil, Iraq. [3]Department of Information Technology, College of Engineering and Computer Science, Lebanese French University, Erbil, Iraq. *For Correspondence: Dina .mikhail@epu.edu.iq*

In this paper, a tumor protein P53 (*TP53* gene) database and an Excel file is used to store the large amounts of data. In addition, this technique uses Machin learning algorithms to extract useful data from large data sets.

First, a normal gene protein sequence is compared to a human's gene protein sequence using the Bio Edit software to predict, diagnose, and classify cancer mutations. This means that there is a cancerous mutation in the person's gene sequence In case of a mismatch between the two protein sequences. Furthermore, bioinformatics techniques were used to complete this stage. Secondly, it is also necessary to carefully select the UMD Cell line-2010 P53 mutation database fields when training Machin Learning algorithms, such as Sequential Minimal Optimization (SMO) and Multilayer Perceptron (MLP) algorithms. The present study then provides an efficient technique for predicting and classifying cancer.

Literature Review

(Ghany and Yousif, 2016) suggested two stages to achieve this paper. The first stage bio informatics tools such as BLAST and ClustalW are used. while, in the second stage, the database of the *TP53* gene (UMD Cell line 2010 database) was used for learn and test the neural networks; which, it consisted of 12 out of 53 fields. Furthermore, a gene location field did not exist in this data base, but in this paper is added as a new field to learn and evaluate the Feed Forward Back Propagation algorithm, which can classify each mutation into a specific type of cancer. The result of training and testing data , (MSE) is (0.00000000000001) and the training rate is (1).

(Mikhail, 2019) introduced the technique for Pre-cancer Diagnosis. A neural network algorithm was used to find a data mining method that could diagnose pre-cancer. In the first step, bioinformatics tools such as BLAST in NCBI site and CLUSTALW in Bloedite application, were used to determine whether or not a person's gene sequence contained malignant mutations; in the second step, three sub-Feed Forward Back Propagation neural networks, one for each sub-dataset, were employed to classify pre-cancer via malignant mutations at precocious stages using a data mining algorithm. Moreover, it provided training rate 1 with performance (MSE) of (5.82E-12), (9.99E-12), and (9.98E-12), respectively.

(Wu and Hicks, 2021) analysed the commons genomics data of RNA-Sequence, it extracted 110 triple negative and 992 non-triple negative breast cancer tumor samples from the Cancer Genome Atlas to choose the features utilized in the validation and development of the classification models. This paper evaluated four different models of classification including Support K-nearest neighbour, Naïve Bayes, Decision tree and Vector Machines, by using some features selected at various threshold levels to learn the models for classifying the both types of breast cancer. As a results, the algorithm of SVM was able to classify breast cancer more precisely into triple negative cancer and non-triple negative breast cancer .Meanwhile, the classification errors were less than the other three algorithms.

Materials and methods

The main aim of the comparative evaluation is to classify cancer via *TP53* gene mutations using two stages, it is determined in the first stage whether or not a person's sequence contains cancer-causing mutations. Machine Learning algorithms to diagnose and determine the type of cancer was found by classifying mutations, which derived from the first technique. Below are the two approaches:

1. Bioinformatics Tools: The objectives of bioinformatics are threefold. First, bioinformatics organizes data in such a as to help researchers have access to existing information and submit new entries. The second goal is to create tools and resources to aid in data analysis. For example, after sequencing a specific protein, it is interesting to compare it to previously characterized sequences. The third goal is to use these tools to analyze data and interpret results in a biologically meaningful way(Rana, 2014; Luscombe et al., 2016).

A. BLAST: In the database, the Basic Local Alignment Search Tool (BLAST) operation is used to find homology, similarity, alignment, and annotation between the sequences of DNA or proteins. Furthermore, it helps to determine the relationships between proteins or genes (Altschul Sf; Siddesh and Editors).

B. Pairwise Sequence Alignment Algorithms is used to compare between two sequences to determine how many mutation events are required to explain the distance. Pairwise sequence alignment is the base currency of sequence comparison (Siddesh and Editors). The mutation in the person's sequence is detected by using the pairwise alignment tool. Mutations in genes increase the probability of cancer. In biology, biologists must understand that there are two types of gene sequences: the normal gene sequence (without mutations) and the person's gene sequence(Ghany and Yousif, 2016).

The main task of sequence alignment is to compare the normal DNA sequence with the person's DNA sequence in order to check whether the person's gene contains the mutation or not. If there is a match between them, then the person's DNA gene is a normal gene. Otherwise, the normal and person's sequences will be converted from DNA sequences to protein sequences and apply sequence alignment to check protein similarity between them. If the protein sequences match, then it is an indication that the person has a normal gene.. In a person's protein gene, there is a malignant mutation. This step, however, is not enough to predict cancer types. To be effective, the proposed method must include a second stage that uses machine learning algorithms to classify cancer types.

Machine learning algorithms

The changing world of data utilization, particularly in clinical healthcare, is presented by Machine Learning and Deep Learning for Health Care Analytics. It offers a wealth of real-world case studies in biomedical engineering, computer science, healthcare research, and clinical applications (Natarajan, 2017). Machine learning algorithms are required to classify cancer-related mutations that are obtained from the first stage. The learning step is done by using the following neural

network algorithms:

A. Multilayer Perceptron's (MLPS): it stands for a well-known form of neural network techniques. The neural networks of MLP have three main layers, input layer, the hidden layer, and the output layer(Morooj K. Luaibi a, 2019).The input and output layers of feedforward networks are separated by one or more hidden layers where a hyperplane in the input pattern space is represented by the output units. The MLP architecture is shown in (Fig.1). M represents many layers, each one having M nodes. "The weights from the (m-1)th layer to the mth layer are indicated by while the bias, output, and activation function of the ith neurons in the mth layer are, respectively, designated as." MLP can be used to approximate functions as well as classify linearly inseparable patterns. A backpropagation network (BP network) is an MLP is learned using the backpropagation algorithm. Figure (1) shows the relationships below. For ease of presentation, the bias vector is preceded by a plus sign. For m = 2,..., M, and the pth example:(Vishwanathan and Murty, 2002; Natarajan, 2017) .

$$\hat{y}_P = o_p^{(m)}, o_p^{(1)} = x_P, \tag{1}$$

$$\underset{\text{net}_P}{(M)} = \left[w^{(m-1)}\right]^T O_p^{(m-1)} + \theta^{(m)}, \ldots \tag{2}$$

$$0_p^{(m)} = \varphi^{(m)}(net_p^{(m)}) \tag{3}$$

$$\left(\underset{\text{net}_P}{(m)} = \underset{\text{net}_{P,1,\ldots\ldots,}}{(m)} \underset{\text{net}_{P,jm}}{(m)}\right)^T, w^{(m-1)} \text{ is} J_{(m-1)} by\, J_m\ matrix,$$

$$O_p^{(m-1)} = \left(O_{p,1\ldots\ldots,}^{(m-1)} O_{p,jm-1}^{(m-1)}\right)^T, \theta^{(m)} = \left(\theta^{(m)}{}_{1,\ldots\ldots,} \theta^{(m)}{}_{Jm}\right)^T$$

T is the bias vector, and $\varphi^{(m)}(.)$ applies $\varphi^{(m)}{}_i(.)$ to the ith component of the vector within

All $\varphi^{(m)}{}_i(.)$ are frequently chosen to have the same sigmoidal function; one can also select all $\varphi^{(m)}$ as the same sigmoidal function in the first M − 1 layer, and all $\varphi^{(m)}$ in the Mth layer as another yet continuous differentiable function.

BP learning is a supervised learning rule that is employed in the training of feedforward networks, as an example, MLPs and RNNs as well. The algorithm of BP propagates through the network to reverse the difference between the target signal and the network output. After providing data to input neurons, the network's output is compared to a given pattern. Each output unit's error is finally calculated. When the propagation of this error signal is backward, a system of closed-loop controls can be set up. Gradient descent algorithms can be used to adjust weights.

The MSE is defined as optimality's goal function. The MSE is between the actual output and the target output for all the learning pattern pairs.

$$E = \frac{1}{N}\sum_{P\in S} E_P = \frac{1}{2N}\sum_{P\in S}\left\|Y_P^{\wedge} - Y_P\right\|^2 \tag{4}$$

where N is the number of the pattern set, and

$$E_P = \frac{1}{2}\left\|Y_P^{\wedge} - Y_P\right\|^2 = \frac{1}{2} e_P^T e_p \tag{5}$$

$$e_p = Y_P^{\wedge} - Y_P, \tag{6}$$

All the parameters of the network $w^{(m-1)}$ and $\theta^{(m)}$, m = 2,..., M, are collected and represented by a matrix $W=[W_{ij}]$. The function of error E or Ep can be decreased by using the gradient-descent method. When it comes to decreasing Ep, we have

$$\Delta_P W = \lambda \frac{\partial E_P}{W} \tag{7}$$

λ represents the learning rate and it is a sufficiently small positive number

$$\delta_{P,U}^{(M)} = -e_{p,v}\emptyset_v^{'M)} \quad (net_{p,v}^{(M)}), \ m = M - 1 \tag{8}$$

W can be adjusted by

$$\frac{\partial E_p}{\partial w_{uv}^{(m)}} = -\delta_{p,v}^{(m+1)} o_{p,u}^{(m)} \tag{9}$$

Support vector machines (SVMs) using a Sequential minimal optimization (SMO) classifier

SVM is a useful classification tool and it has value in the fields of pattern classification and machine learning. It can solve the problems with complex classification. Classification is done in the input by realizing a linear or non-linear separation surface (Vishwanathan and Murty, 2002; Devi Arockia Vanitha et al., 2014; Morooj K. Luaibi a, 2019; Natarajan, 2017). SVM has many applications in real-word like bio- sequence analysis, hand-written character recognition, image classification, text categorization, etc. This is an SVM learning algorithm that is simple to construct, faster and has excellent scaling features called Sequential Minimal Optimization (SMO). The SVM theory depends on the idea of structural risk minimization (SRM)(Devi Arockia Vanitha et al., 2014; Evgeniou and Pontil, 2014; Swamy, 2014; Morooj K. Luaibi a, 2019). The SVM architecture is shown in Figure 2 .

There are numerous methods for reducing multiple binary classification tasks from a multi-class problem. One of them is SMO (sequence minimal optimization approach) to support vector classifier training. SMO has two components: an analytic process to resolve the two α_i and a method for determining which multipliers should be optimized. The feature of SMO that can solve two α_i can be carried out analytically. So, the use of numerical QP optimization is not recommended. Furthermore, SMO does not require any additional matrix storage. Because SMO does not use matrix algorithms, it is less liable to numerical precision issues (Boujelbene et al., 2008).

Results

Implementing Machine Learning algorithms for cancer classification based on mutations in geneTp53 involves the following steps:

1) The Catalogue Of Somatic Mutations In Cancer (COSMIC): the site is used to obtain the normal *TP53* gene sequence. Normal genes, gene information, and datasets are all available on the (COSMIC) website. The normal gene can be found via asking the server for the gene's name, then choosing the normal gene sequence.

2) BioEdit's tools are employed to obtain personal information about the *TP53* gene, similar to the BLAST tool at the National Centre for Biotechnology Information (NCBI).

The first approach is done to diagnose mutations by applying BioEdit tools like pairwise alignment to show the match between the normal gene and the sequences of a person's genes. As illustrated in Figure 2, an alignment of the normal *TP53* sequence with the person's gene is performed to determine if the person's gene has mutations or not. However, because the results cannot predict

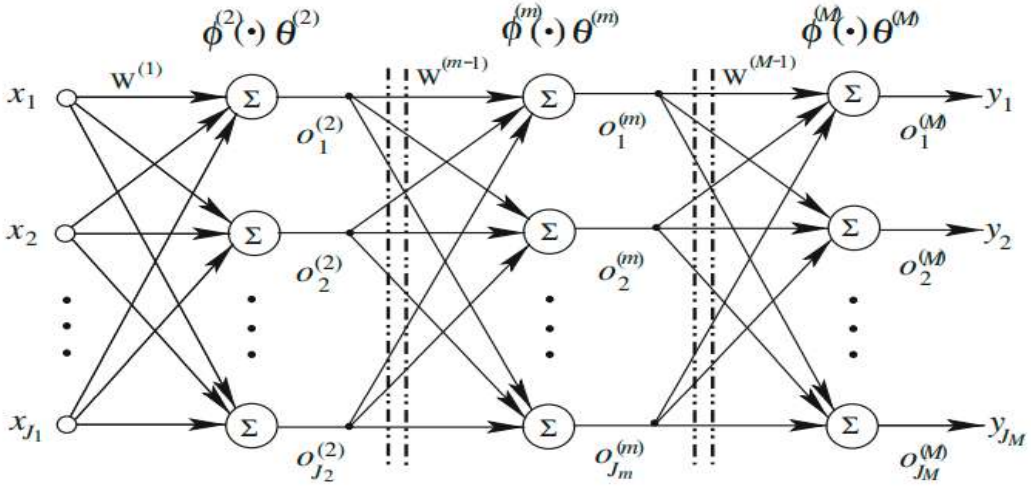

Figure 1. Shows the Architecture of MLP

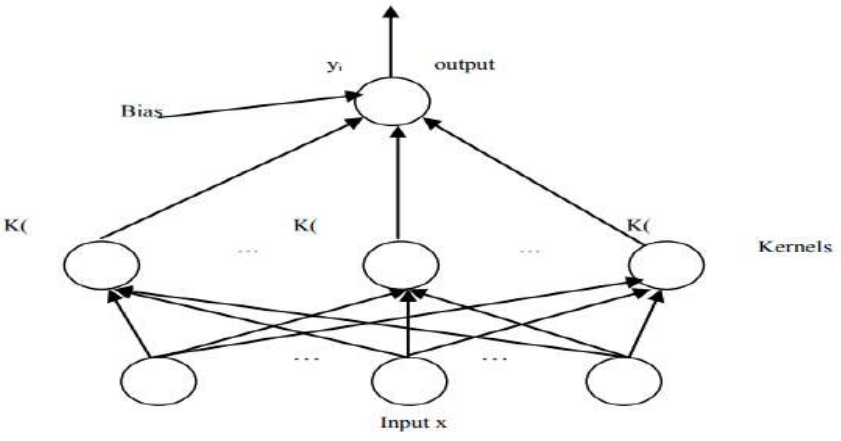

Figure 2. Shows SVM Structure

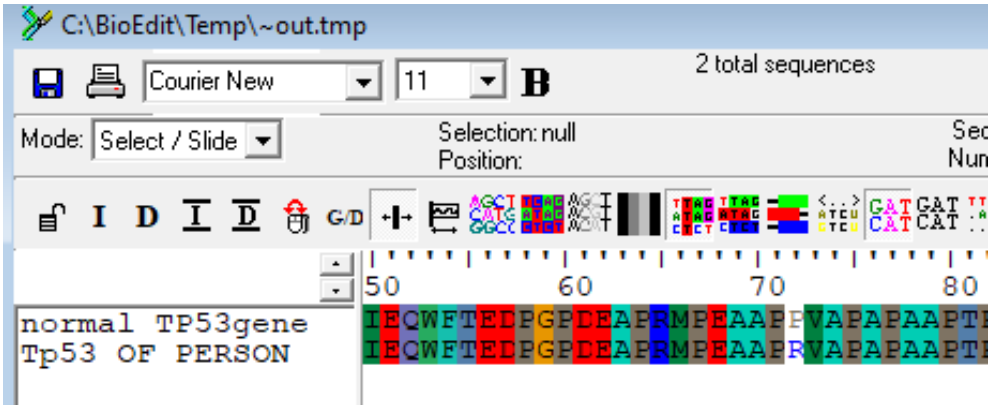

Figure 3. Shows Alignment of Normal TP53

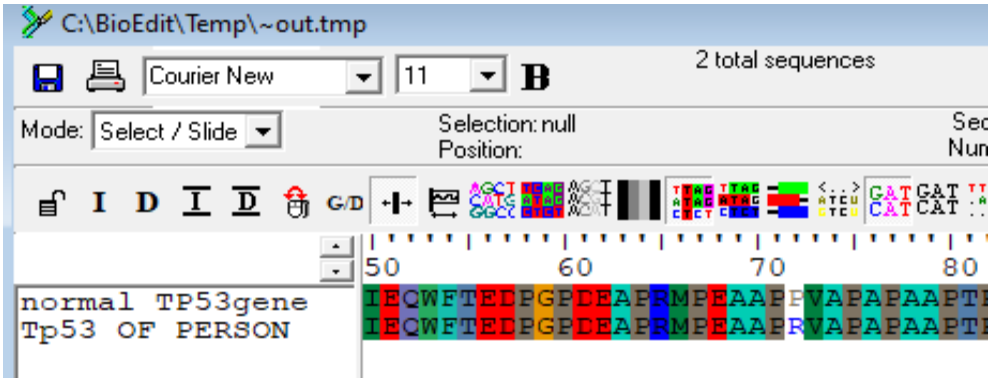

Figure 4. Shows Protein Sequence Contains a Cancerous Mutation

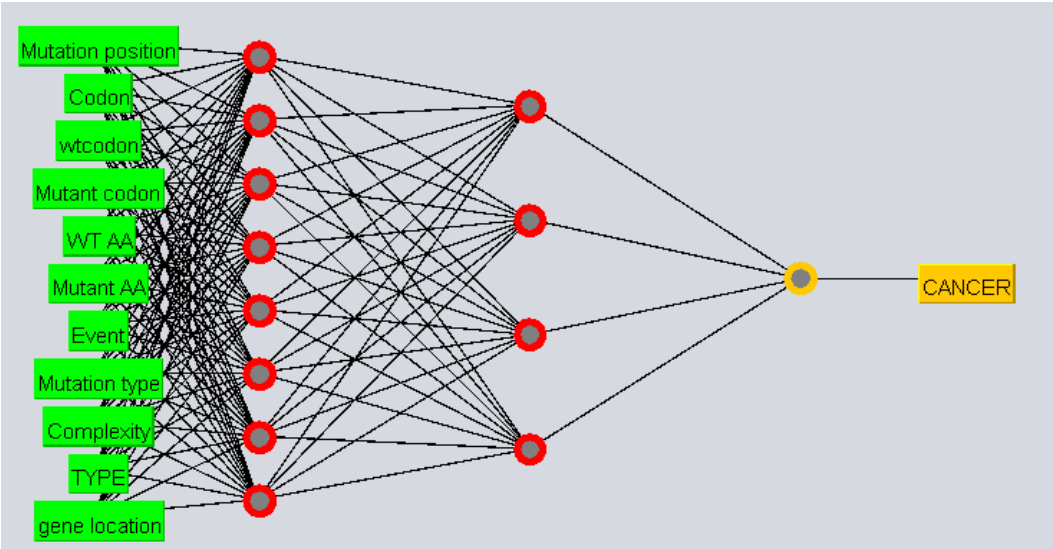

Figure 5. Shows Mlp Structure

whether a mutation impacts protein function or not, this stage is insufficient. As a result, the *TP53* gene sequence in both normal and abnormal people is translated into the tumor protein P53. The pairwise alignment function in the BioEdit package is then used to determine whether or not the person's protein sequence contains a cancerous

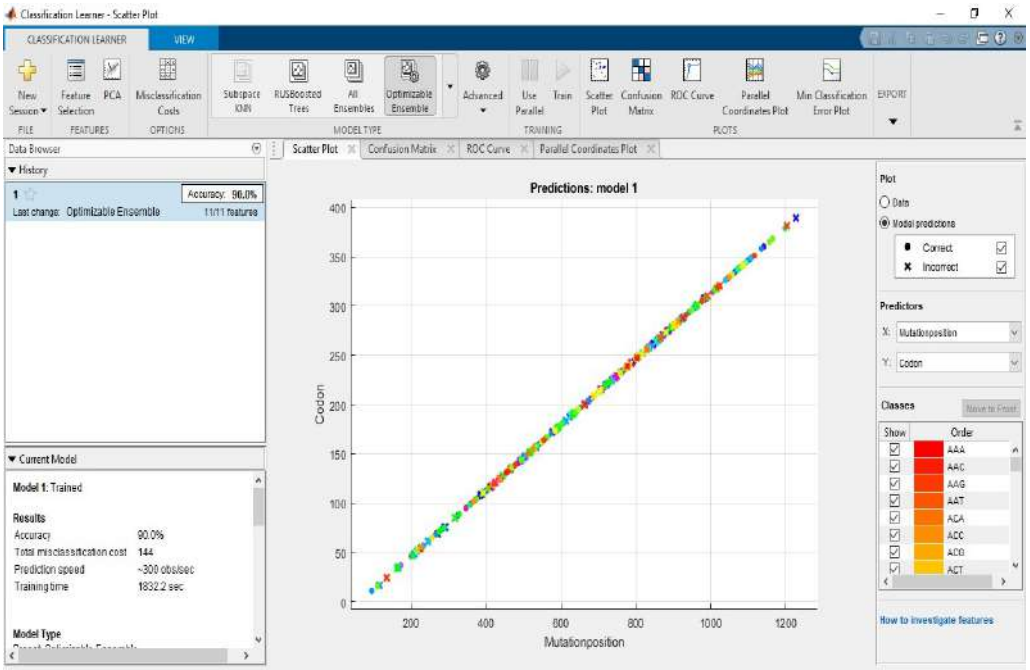

Figure 6. Shows Multi Layer Perceptrons NN Accuracy

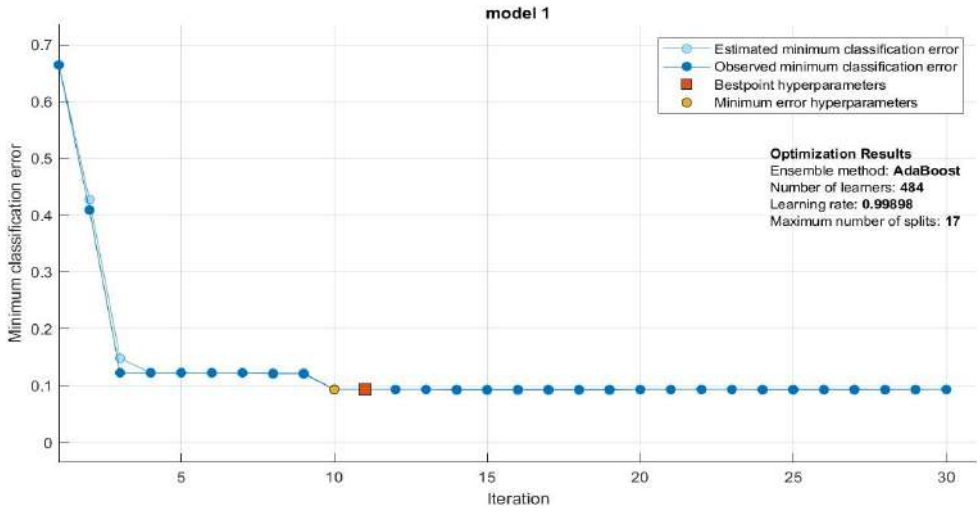

Figure 7. Shows the Minimum Classification Error

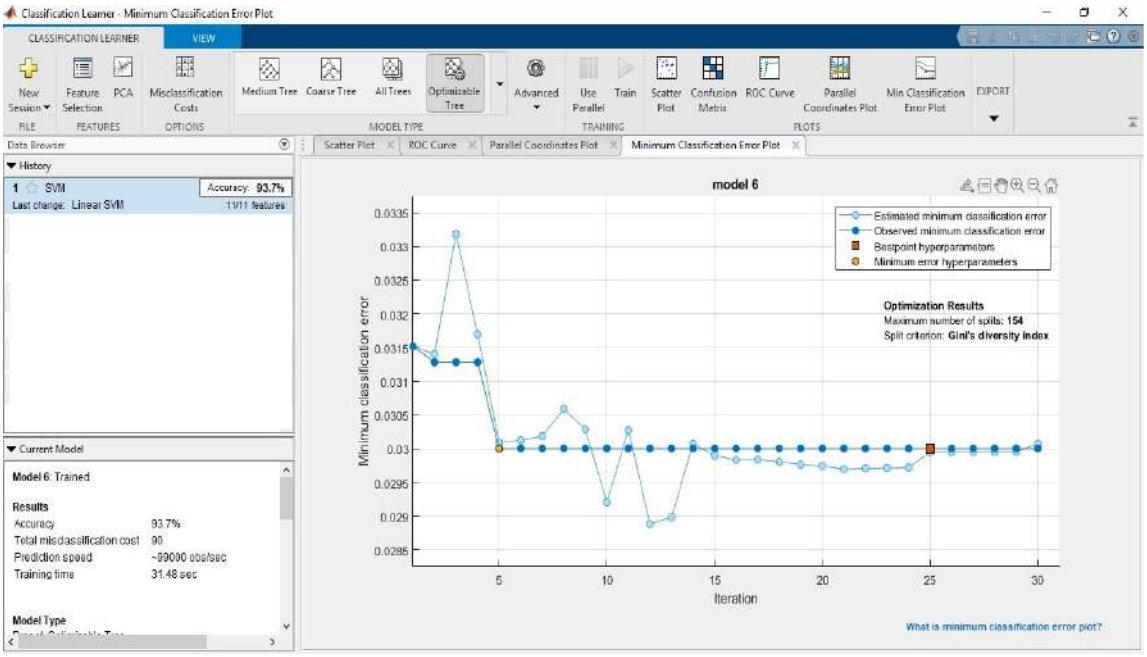

Figure 8. SVMNN Accuracy and Minimum Classification Error

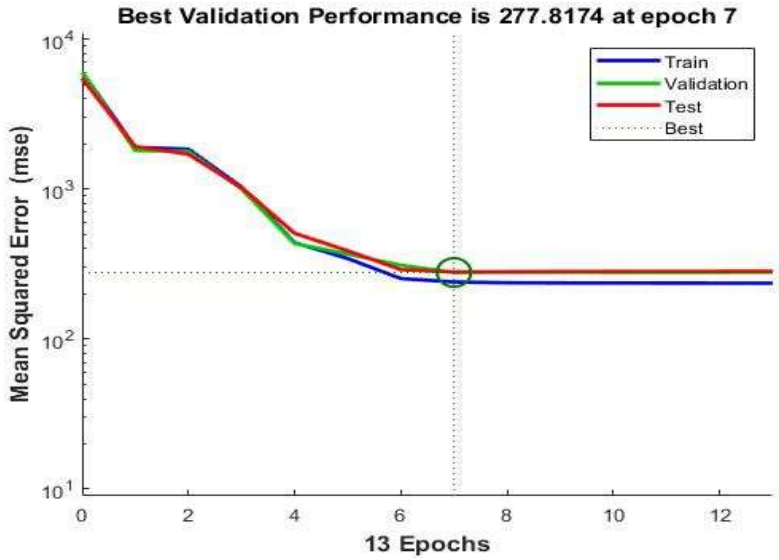

Figure 9. Shows SVMNN Mean Squared Error

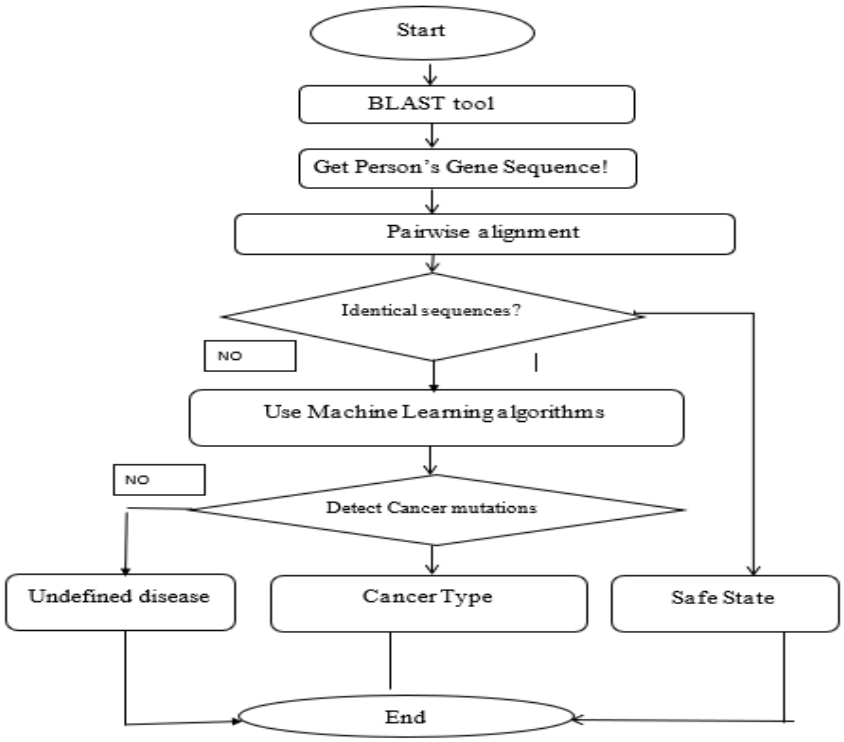

Figure 10. Shows Main Steps Flowchart of Suggested Method

Table 1. Shows Comparison of MLP and SVM Algorithm

Column1	MlP (two hidden layers)	SVM
	The percentage split 75%	The percentage split 75%
Correlation coefficient	0.5669	0.6342
Mean absolute error	13.2958	11.6237
Root mean squared error	17.9234	17.1052
Relative absolute error	77.36%	67.63%
Root relative squared error	84.88%	81.00%
Total Number of Instances	359	359
	Percentage split 85%	Percentage split 85%
Correlation coefficient	0.6133	0.6402
Mean absolute error	14.372	11.2879
Root mean squared error	18.9334	17.0521
Relative absolute error	83.60%	65.66%
Root relative squared error	89.16%	80.30%
Total Number of Instances	216	216
Cross-validation (10 fold)		
Correlation coefficient	0.5995	0.6419
Mean absolute error	12.7017	11.2861
Root mean squared error	17.1316	16.4005
Relative absolute error	75.31%	66.92%
Root relative squared error	83.20%	79.65%
Total Number of Instances	1438	1438
Cross-validation (15fold)		
Correlation coefficient	0.61	0.6415
Mean absolute error	12.4953	11.2775
Root mean squared error	16.8053	16.4056
Relative absolute error	74.10%	66.88%
Root relative squared error	81.62%	79.68%
Total Number of Instances	1438	1438

mutation as shown in Figure 3.

Figure 3 and4 demonstrate that a malignant mutation (CCC CGC), present in the codon 172, transforms from (P) amino acid (in Normal's protein sequence) to (R) amino acid (in Person's protein sequence).

4) step (3) does not classify the type of cancer because it is used to diagnose the malignant mutation in the person's sequence. This has something to do with the *TP53* gene database. So, the (UMD_Cell_line_2010) database is commonly employed to learn (MLPS) and (SVMs) neural networks, which is comprised of 53 fields and 1448 entries. A comprehensive and up-to-date database can be found at the following URL: http://p53free.fr/Database/p53MUTMAT.html . The UMD Cell line 2010 database is commonly employed to select 11 of the 12 fields for learning and testing NNs. To create precise and efficient outcomes in cancer classification, the (gene location field) field was added to the (11) fields specified.

5) By using the structure of Multilayer Perceptron's NN as shown in Figure (5) and the Support vector machines (SVMs) structure employing the Sequential minimal optimization (SMO) classifier, the malicious mutations for cancer are classified successfully to reach an optimal classifier for classification of cancer. Furthermore Regression state and the accuracy of Multilayer Perceptron's NN and the minimum classification error are shown in Figures(6) and 7) respectively, while the SVMNN accuracy and Mean Squared Error are shown in Figures(8) and (9) respectively.

The main steps of the suggested method for classifying cancer types are shown in Figure 10.

Discussion

Both structures have completed the training and learning stages. As a result of MLP and SVM being utilized for training and testing a set number of fields, which is twelve of the fifty-three fields in each *TP53* database record, e- learning methods are an effective way to classify cancers based on mutations. P53 database data was presented as columns and records in an Excel spreadsheet. To further improve accuracy, this paper populates the UMD *TP53* database with a new field called Gene Location. Table 1 also shows the results of learning and testing the proposed cancer classification method. As mentioned in the literature review, there are many proposed methods . But, some of these methods can classify two types of cancers or use a single Machine Learning algorithm for classification. While, In the present paper, however, two machine learning algorithms were learned and evaluated to classify 32 types of cancers.

Author Contribution Statement

All authors contributed equally in this study.

Acknowledgements

Approval

It was not approved by any scientific Body.

Ethical Declaration

there is no ethical committee to approve the research

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Study Registration

The study was not registered in any registering dataset .

Conflict of Interest

The authors declare that they have no conflicts of interest. The stakeholders had no role in the design, collection, analysis, or interpretation of dated in the writing of the manuscript or in the decision to publish the results.

References

Abdel-Razeq H, Attiga F, Mansour A (2015). Cancer care in Jordan. *Hematol Oncol Stem Cell Ther*, **8**, 64-70.

Altschul Sf GWMWMEWLDJ Basic local alignment search tool. *J Mol Biol*, **215**.

Balmain A, Gray J, Ponder B (2003). The genetics and genomics of cancer. *Nat Genet*, **33**, 238-44.

Boujelbene SZ, Mezghani DBA, Ellouze N (2008). Vowel phoneme classification using SMO algorithm for training support Vector Machines. 2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications, ICTTA.

Devi Arockia Vanitha C, Devaraj D, Venkatesulu M (2014). Gene expression data classification using Support Vector Machine and mutual information-based gene selection. *Procedia Comput Sci*, **47**, 13-21.

Evgeniou T, Pontil M (2014). Support Vector Machines : Theory and Applications WORKSHOP ON SUPPORT VECTOR MACHINES : THEORY AND APPLICATIONS.

France database of *TP53* gene [Online]. Available: http://p53.fr/tp53-database [Accessed 23/10/2021.

Francis Bk BSS. Predicting Academic Performance of Students Using a Hybrid Data Mining Approach. *J Med Syst*.

Ghany A, Yousif D (2016). Effective Data Mining Technique for Classification Cancers via Mutations in Gene using Neural Network. *Int J Adv Comput Sci Appl*, **7**, 69-76.

K.-L. Du MNsS (2014). Multilayer Perceptrons : Architecture and Error Backpropagation.

Luscombe NM, Greenbaum D, Gerstein M (2016). What is bioinformatics ? An introduction and overview What is bioinformatics ? An introduction and overview.

M-amen K, Abdullah OS, Amin AMS, et al (2022). Cancer Incidence in the Kurdistan Region of Iraq: Results of a Seven-Year Cancer Registration in Erbil and Duhok *Governorates*, **23**, 601-15.

Mikhail DY (2019). Pre-cancer Diagnosis via *TP53* Gene Mutations by Using Bioinformatics Neural Network. Proceedings of the 5th International Engineering Conference, IEC 2019, pp 136-41.

Morooj K. Luaibi a FGMb (2019). FACIAL RECOGNITION BASED ON DWT – HOG – PCA FEATURES WITH MLP CLASSIFIER.

Mosayebi A MBBNAKHSH (2020). Modeling and comparing data mining algorithms for prediction of recurrence of breast cancer. *PLoS One*, **15**, 10.

Natarajan P (2017). Demystifying Big Data, Machine Learning,

and Deep Learning for Healthcare Analytics, CRC Press.
Neamatollahi P, Hadi M, Naghibzadeh M (2020). Efficient Pattern Matching Algorithms for DNA Sequences. 2020 25th International Computer Conference, Computer Society of Iran, CSICC 2020, 8.
Neelamegam S, Ramaraj E (2013). Classification algorithm in Data mining: An Overview. I*nt J P2P Network Trends Technol (IJPTT)*, **4**, 369-74.
Oluwaseun A, Chaubey MS (2019). Data Mining Classification Techniques on the. *Global Sci J*, **7**, 79-95.
Pei-Tse Yang W-SWC-CWY-NSC-HH, Jia-Lien H. Breast cancer recurrence prediction with ensemble methods and cost-sensitive learning. open mid(wars), **16**.
Pitolli C, Wang Y, Mancini M, et al (2019). Do mutations turn p53 into an oncogene?. *Int J Mol Sci*, **20**.
Rana J (2014). Introduction To Bioinformatics.
Siddesh GM, Editors SRM (Algorithms for Intelligent Systems) K. G. Srinivasa (editor), G. M. Siddesh (editor), S. R. Manisekhar (editor) - Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, .pdf.
Swamy K-LDMNS 2014. Neural Networks and Statistical Learning, Springer.
Vishwanathan SVN, Murty MN (2002). SSVM: A simple SVM algorithm. Proceedings of the International Joint Conference on Neural Networks, **3**, pp 2393-8.
Wu J, Hicks C (2021). Breast Cancer Type Classification Using Machine Learning. *J Pers Med*, **11**.