Standard Multiple Regression Analysis Model for Cell Survival/ Death Decision of JNK Protein Using HT-29 Carcinoma Cells

Shruti Jain, D.S. Chauhan

Abstract*: Signaling by the JNK protein has been studied for more than decades with various previous reviews covering more specific aspects. For estimating the relationship among variables a statistical technique called Regression analysis (RA) is used. RA is used to determine the correlation among two or more variables. In this paper, a multiple regression analysis is used to assess the most significant contribution of JNK protein using ten different concentrations of TNF, EGF, and Insulin that control the survival/ apoptosis response of HT-29 human colon carcinoma cells. The data is analyzed using Statistica software. Data normality and the outliers were checked by visual method (histograms, box plot and Q-Q plot). Descriptive statistics (mean and standard deviation) and correlation matrix (correlation and covariance between variables) are used to get the best concentration. Standard regression analysis is used to make a model through which analysis of variance, regression coefficient & correlation coefficients were analysed and based on the p-value we come to know that 100-0-500 yields the best concentration level which helps in the analysis the cell survival/ apoptosis of JNK protein that was validated by variable importance plot.

***Index Terms*: Regression analysis, correlation, covariance, standard deviation, JNK**

I. INTRODUCTION

Data analysis can be done by different approaches. There are two types of data: categorical (numerical or binary) and continuous data. For every analysis, we have to model the system that can be done by deterministic or probabilistic modeling. Deterministic/descriptive/unsupervised (learning) modeling has no randomness, it is suitable when predicted error is negligible, hypothesize is exactly related and is analyzed through clusters of the data. Probabilistic modeling is with randomness, hypothesize has 2 components deterministic and random error. Probabilistic modeling is further divided into three different models: the predictive model, the correlation model, and the regression model. Supervised learning is also referred to as predictive modeling. While dealing with categorical target variable, the modeling algorithms are analyzed by classification techniques [1-5]. Another type of supervised learning is regression where we predict continuous outcomes [6-8]. Predictive modeling is further divided into different types: logistic, multiple, decision tree, neural network, & survival modeling (Cox regression).

Regression Analysis (RA) is a statistical method for investigating the relationship between different variables. RA can be applied to categorical or continuous data both. For continuous data, RA is divided into three types: linear regression (LR), multiple linear regression (MLR or MR) and non-linear regression (NLR) (shown in Figure 1) while for categorical data binary logistic regression is used as a non-parametric test. LR is used to model a linear relationship between dependent and independent variables but if independent variables are more than one and dependent variables are one than, MLR is used. For NLR, dependent and independent variables are not linear. NLR model is complicated than LR in terms of estimation of model diagnosis, model parameters, model selection, outlier detection, or variable selection [9-11]. Ordinary least square (OLS) method and partial least squares analysis (PLS) methods are used for calculation of LR while forward selection (FS), backward elimination (BE) and stepwise approximation (SWA) are used for MR [12-13]. PLS method is used when factors/ variables are many and highly collinear or multi-collinear. It is mostly used in the industry for soft modeling. PLS is used for LR using continuous predicted values and MR for categorical predicted values or many other combinations.

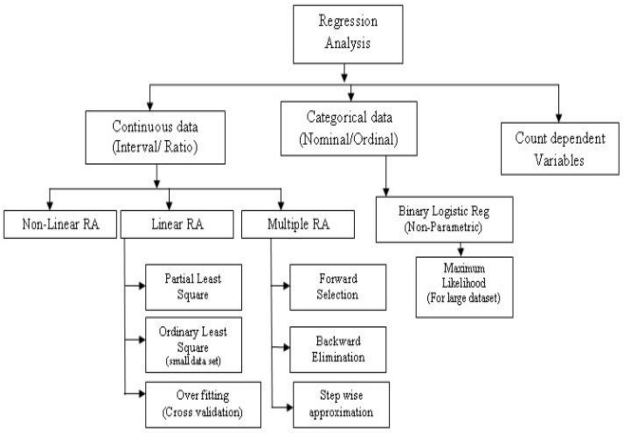
Regression Analysis
Continuous data (Interval/ Ratio)
Categorical data (Nominal/Ordinal)
Count dependent Variables
Non-Linear RA
Linear RA
Multiple RA
Binary Logistic Reg (Non-Parametric)
Partial Least Square
Ordinary Least Square (small data set)
Over fitting (Cross validation)
Forward Selection
Backward Elimination
Step wise approximation
Maximum Likelihood (For large dataset)

Figure 1: Different types of Regression Analysis and their methods

This paper presents the standard regression analysis to model a system which predicts whether JNK protein is present or absent. For experimentation, authors have considered ten different combinations of tumor necrosis factor-alpha (TNF-α), Epidermal growth factors (EGF) and insulin [14-15] mediated cell survival/ cell death response of HT-29 human colon carcinoma cells [15-16]. EGF/Insulin binds

Revised Manuscript Received on August 05, 2019.

Shruti Jain, ECE Department, Jaypee University of Information Technology, Soaln, Himachal Pradesh, India.

D.S. Chauhan, GLA Mathura, Uttar Pradesh, 281406. India

Retrieval Number H7163068819/2019©BEIESP
DOI: 10.35940/ijitee.H7163.0881019

with their receptor that activates the tyrosine kinase activity of the receptors. Grb2 and SH2 domain-contains protein which recognizes phosphorylated residues along with a guanine nucleotide exchange factor SOS and bind to the EGFR. Upon the extracellular mitogen binding to the ligand, GDP is then swapped for GTP which binds Ras to make it active. Ras activates Raf (MAP3K) that in turn activates MAP2K and than Mitogenic activated protein kinase (MAPKs). MAPK is the highly conserved family of threonine/serine protein kinase involved in main cellular processes such as differentiation, stress response, proliferation, survival, motility, and apoptosis. MAPK is a protein family that found in the maximum eukaryotic organism and is widely distributed [15-17]. The number of transcriptions factors can be activated by MAPK which control processes in the cell. MAPK are classified in three main groups: extracellular signal kinases (ERK ½, p41/42), c-jun NH2 terminal (JNK1-3) and high osmolarity glycerol pathway (p38 α, β, γ, δ / HOG). ERK is mediated by differentiation and mitogenic signals, JNK responds to ultraviolet (UV), inflammatory and stress cytokines due to this it is also known as stress activated protein kinase (SAPK) pathway. JNK activates apoptotic signaling by the up-regulation of pro-apoptotic genes via the translation of specific transcription factors or by directly modulating the mitochondrial pro and anti-apoptotic proteins through distinct phosphorylation events. Recent studies have improved our understanding the function of the JNK pathway. JNK is multifunctional kinase which is involved in many physiological processes.

Signaling by the JNK has been studied for more than decades with various previous reviews covering more specific aspects like the inhibition of JNK signaling as a therapeutic strategy in cancer or signaling in the brain. To start with, data was normalized and the outliers were removed. Descriptive statistics and correlation matrix were found to get the best concentration values. Standard multiple regression analysis models were analyzed which helps in predicting the Regression coefficients, and correlation coefficients. MLR is one of the widely used statistical techniques [18-19] in educational research. It is defined as a multivariate technique that determines the correlation between a response variable. The results were found to be same which were validated by the importance plot.

A study of data mining techniques is described in Section 2. A detailed discussion of datasets employed and the proposed methodology is explained in section 3. The standard regression analysis detection system using MLR for JNK protein is discussed in section 4. Section 5 explains the conclusion of the proposed work.

II. DATA MINING

Data mining is an approach which is used for extricating helpful in function from the large data set. Data Mining incorporates many techniques consisting of Statistical based approach, classification techniques, and clustering techniques. Further, these techniques can be applied to the dataset by different algorithms. An overview of these different techniques is shown in Figure 2.

i. Statistical based approach used principal component analysis, interpolation, and regression techniques.

ii. In clustering, there is no training set as the class labels are unknown. For example: let us assume that we have a dataset of cows belonging to different breeds and it contains the following attributes/variables: height, width, weight, and color. But we don't have the information about the breed of any of cow. So, on the basis of these four attributes, we would make clusters (number of clusters can be selected with various methods) such that each cluster would contain only those records or objects which have more similarities with each other than other clusters. Clustering algorithms can be divided into two categories which are unsupervised linear clustering and non-linear clustering. The former includes the algorithms like Gaussian clustering, Hierarchical clustering, fuzzy c-means, quality threshold, k-means etc. and later includes MST based clustering algorithms, kernel *k*-means (*k* describes the number of clusters that should be made) clustering algorithm and density-based clustering algorithm. The focus of *k*-mean is to partition a dataset in which the data in a group is more similar to each other. For partitioning, Euclidean distance can be used and the objects which are near to a certain centroid will be considered a part of that cluster.

iii. Classification Technique uses a single regression tree, decision tree, and machine learning algorithm. Classification trees / Single regression trees were divided into Classification and Regression Trees (CRT models, CART, CHI), Interactive C&RT algorithm (ICR), Interactive Exhaustive CHAID algorithm (IEC) and Chi-squared and Interactive Decision (CHAID). Averaging Trees/ Decision Trees was classified as Bagging Trees (BT) known as Averaging Trees, Random Forests (RF) known as Cleverer Averaging of Trees and Boosting Trees (BOT) known as Cleverest Averaging of Trees. Machine learning algorithms were classified as: Supervised learning, unsupervised learning and enforcement learning where they are further divided into *k*NN, Artificial Neural Networks (ANN)[20] and Support Vector Machines (SVM) [1-4]. Machine learning is considered to be a part of Artificial Intelligence. Machine Learning algorithms learn on their own experience hence do not need any human to enhance their ability. Supervised learning is useful when the class labels are known in advance and model is then trained to predict the class of a particular record. However, unsupervised learning is useful in cases where the challenge is to discover implicit relationships in a given unlabeled dataset [3-5]. In enforcement learning, only some form of feedback is available instead of proper class label or error label at each prediction step. OLS regression, Logistic Regression and Support Vector Machines are a few machine learning algorithms which are widely used in day to day applications.

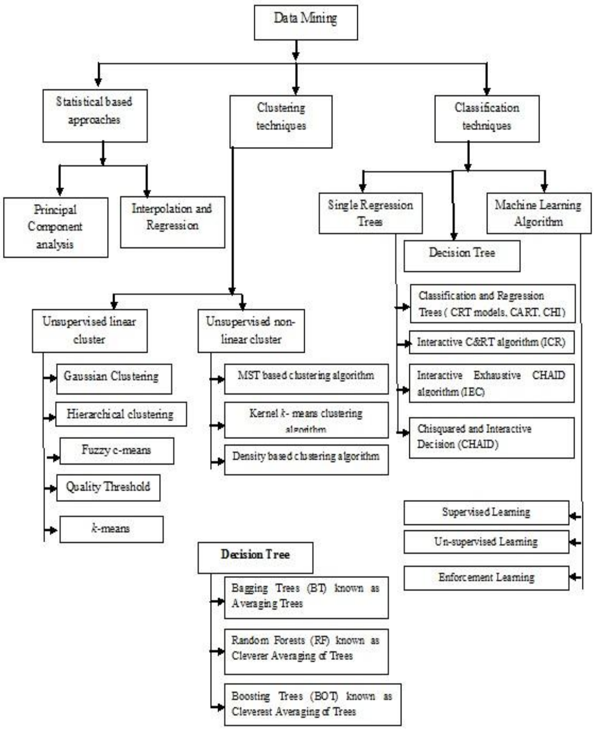
Data Mining
Statistical based approaches
Clustering techniques
Classification techniques
Principal Component analysis
Interpolation and Regression
Single Regression Trees
Machine Learning Algorithm
Decision Tree
Classification and Regression Trees (CRT models, CART, CHI)
Interactive C&RT algorithm (ICR)
Interactive Exhaustive CHAID algorithm (IEC)
Chisquared and Interactive Decision (CHAID)
Unsupervised linear cluster
Unsupervised non-linear cluster
Gaussian Clustering
Hierarchical clustering
Fuzzy c-means
Quality Threshold
k-means
MST based clustering algorithm
Kernel k- means clustering algorithm
Density based clustering algorithm
Supervised Learning
Un-supervised Learning
Enforcement Learning
Decision Tree
Bagging Trees (BT) known as Averaging Trees
Random Forests (RF) known as Cleverer Averaging of Trees
Boosting Trees (BOT) known as Cleverest Averaging of Trees

Figure 2 : A schematic representation of the classification of the data mining techniques discussed

The dataset was collected and later pre-processed. Preprocessing is a process of transforming or making data appropriate. It includes data integration, discretization, cleaning, transformation and reduction. Data is transformed into the format required for the analysis.

1. *Data Integration* : In this method different data was collected and integrated from multiple databases. This also includes removing duplicates and redundant data. Detecting and resolving data which involve conflicts. If the data is categorical than chi square test (correlation values / covariance values) was done but if it is continuous than regression analysis was performed.
2. *Data Discretization*: In this method division of range for the continuous features was done into intervals because some data mining algorithms only accept categorical attributes. Discretization of numeric/ categorical data can be done by binning method, histogram analysis or clustering analysis.
3. *Data cleaning* involves the smoothening of noisy data, filling the missing values, identify or removing the outliers. Noisy data can be solved by binning method, clustering and regression.
4. *Data Transformation*: This step involves normalization, smoothening and aggregation of data. Normalization of data can be performed by min-max approach, z-normalization, or normalization by decimal scaling.
5. *Data Reduction*: It means to remove unimportant attributes. It consists of data reduction strategies, regression and log linear model, aggregation and clustering, sampling, data compression and histograms. Data reduction strategies consist of data compression, numerosity reduction, data cube aggregation and dimensionality reduction. In dimensionality reduction there are two steps feature selection and heuristic methods. Feature selection is further divided into two techniques direct method (selection of a minimum set of feature (attribute)) and indirect method while Heuristic method involves step wise forward selection, backward elimination and or both.

III. MATERIALS AND METHODS

The principle idea in developing the proposed system is to facilitate the accurate detection of JNK providing least runtime complexity. The proposed algorithm has been implemented using Statistica software with Intel Core i5 processor, 3GHz and 8GB RAM configuration.

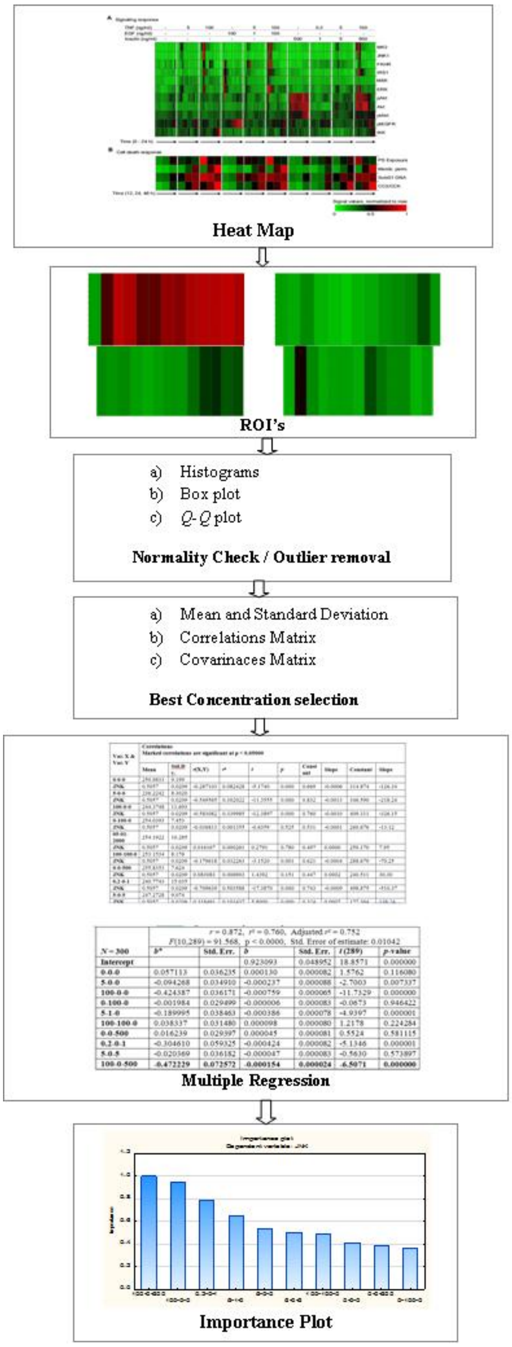

Figure 2: Schematic diagram of Proposed JNK protein Detection Method.

A. Materials

a. Data Sets: The experimental data (heat map) of cell survival/ apoptosis for different marker proteins was taken from Gaudet *et al*. (2005) of TNF, EGF or insulin treated with ten cytokine combinations. There are different marker proteins: MK2, JNK, FKHR, MEK, ERK, IRS, Akt, IKK, pAkT, ptAkT and EGFR for the HT carcinoma cells. In this paper, we are working only with JNK protein. The average signal values were normalized to the maximum value for which an excel data was prepared for 0-24 hrs. We can also say that if these proteins are absent or electronically zero (0) than it leads to cell death but if these proteins are present or electronically one (1) than there is a cell survival. The average value of four different outputs (phosphatidylserine exposure (PE), membrane permeability (MP), nuclear fragmentation (NF) and caspase substrate cleavage (CCK)) was normalized to the maximum that yields one output (cell death < 0.5, cell survival ≥ 0.5).

b. Description: There are different steps to analyze the model using Standard Multiple Regression Analysis. Initially, visual methods (histograms, box and whisker plot, Q-Q plot) are used for normality check and removing outliers. Descriptive statistics (mean and standard deviation (SD)), a correlation matrix (correlation and covariance are between variables) were analyzed to get best concentration values. Multiple regression analysis method is used to model the JNK protein. Based on the *p*-value, significant concentration value is selected. The results were validated by variable importance plot. The best concentration level should be the same as predicted by multiple linear regressions.

B. Methodology

There are different types of modeling techniques out of which, we are using Standard multiple RA for modeling our system. A schematic diagram of the proposed five stage technique is shown in Fig 3 which we have followed in this paper to get the best concentration of TNF-EGF-Insulin for the survival/death of JNK protein. Initially, a heat map from Gaudet et al was considered from where we have extracted the ROI's of JNK protein. For every concentration, there are 13 different values which shows 0-24 hr data. Using visual (histogram, Q-Q plot and box plot) approach, the normality of the data was checked and the outliers were removed. After removal of outliers, concentration values were selected using the mean, SD, correlation matrix and covariance matrix. Multiple regression analysis was performed to model the JNK protein by getting the analysis of variance, regression coefficient, and correlation coefficient and based on the p-value the best concentration was selected. A p-value ≤ 0.05 gives significant results. The results were validated by variable importance plot.

IV. RESULTS AND DISCUSSIONS

There are two types of error measurement consisting generalization error and training set error when working with regression analysis. The generalization error relates to how accurate the model will be when applied to other points while training set error relates to how close the regression is to the data being fit while. The various steps which were followed for modeling the JNK protein are normalization, removal of outliers, selection of best concentration values and calculation of regression and correlation coefficients.

A. Normalization and removal of outliers: There are different ways of normalizing the data.

a) *Histograms*: The distribution of any variable should not deviate from the normal distribution. Table 1 shows the normal distribution function values. Fig 4 (*a*) and Fig 4 (*b*) shows the histogram of normal distribution value for JNK and residuals respectively.

b) *Quantile – Quantile* (*Q-Q*) plot : The *Q-Q* plot is a graphical technique that helps in plotting the quantiles by comparing two probability distribution. If two distributions are same and linearly correlated than points will lie on $y = x$ line.

Table 1 : Normal distribution function values

Upper Boundary	Variable: JNK, Distribution: Normal *Chi-Square* = 276.04040, *df* = 6 (adjusted) , *p* = 0.00000								
	Observed	**Cumulative**	**Percent**	**Cumul. %**	**Expected**	**Cumulative**	**Percent**	**Cumul. %**	**Observed-**
≤ 0.46000	0	0	0.00000	0.0000	4.31496	4.3150	1.43832	1.4383	-4.3150
0.47000	10	10	3.33333	3.3333	8.82091	13.1359	2.94030	4.3786	1.1791
0.48000	58	68	19.33333	22.6667	19.66783	32.8037	6.55594	10.9346	38.3322
0.49000	32	100	10.66667	33.3333	35.03224	67.8359	11.67741	22.6120	-3.0322
0.50000	0	100	0.00000	33.3333	49.85081	117.6867	16.61694	39.2289	-49.8508
0.51000	13	113	4.33333	37.6667	56.67407	174.3608	18.89136	58.1203	-43.6741
0.52000	93	206	31.00000	68.6667	51.47661	225.8374	17.15887	75.2791	41.5234
0.53000	85	291	28.33333	97.0000	37.35465	263.1921	12.45155	87.7307	47.6453
0.54000	9	300	3.00000	100.0000	21.65584	284.8479	7.21861	94.9493	-12.6558
< Infinity	0	300	0.00000	100.0000	15.15208	300.0000	5.05069	100.0000	-15.1521

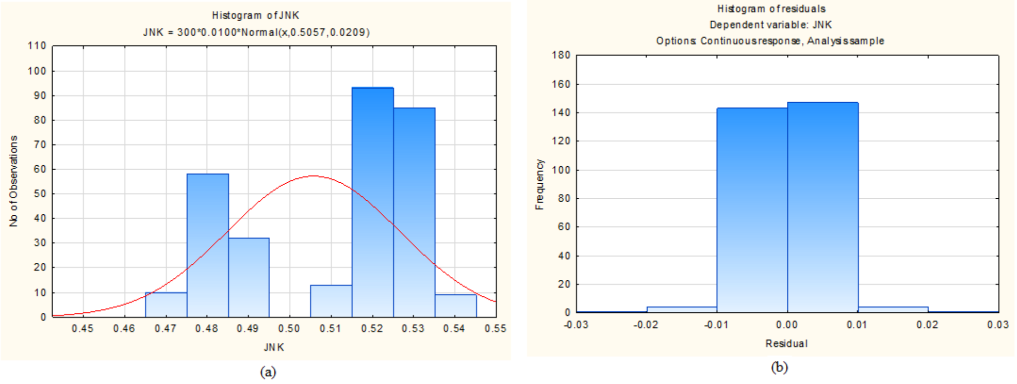

Figure 3: Histogram using normal distribution function (a) for JNK (b) Residuals

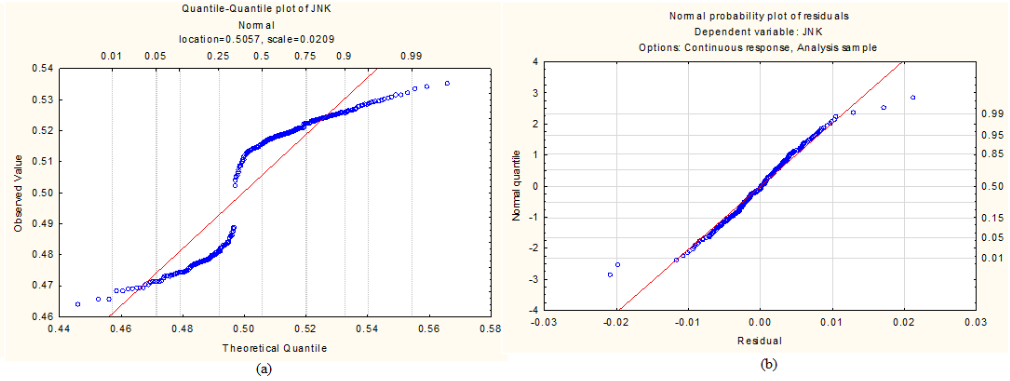

Figure 4: Q-Q plot (a) JNK (b) Residuals

The graph (Fig 5(*a*)) shows that the smallest (or largest) concentration of JNK protein is not small (or large) enough to be compatible with the normal distribution. This shows that the tails of the JNK protein distribution are too “skinny” or “thin” compared with the normal distribution. It can be concluded that the data of JNK protein are not normally distributed. Fig 5(b) shows that the *Q-Q* plot of residual of JNK protein does not match to the normal distribution. In this case, the smallest value to JNK protein is too small and the largest value is too large to be compatible with the normal distribution. This shows that the tails of the JNK protein are too “fat” or “thick” in comparison with the normal distribution. This concludes that the data is not normally distributed. To overcome this problem data is normalized so as to remove all the error occurs.

c) *Box plot*: The statistical analysis of the descriptive feature set is normalized employing the graphical method. Box plot analysis is a graphical representation of normality visualization of the feature set. Outliers are one of the major problems that every author faces while working with RA. One outlier can create a problem in estimating the values which result in parameter estimates which don’t give useful information about the data. Outliers can be generated from a simple operational mistake including small sample from a different population, and they make serious effects of statistical inference. Box plot analysis is suitable for statistical data representation indicating the 25% to 75% of the data distribution inside the rectangular box. Outliers are those feature values which do not fit in the normalized feature set range and they can be discarded for further computation.

In total we have 350 samples with ten different input combinations. After removal of outliers we get 300 samples which were used for further analysis. The Anderson darling (AD) test is used to determine if a data set follows a specified distribution. We can use the AD value to compare the fit of several distributions to determine which one is the best. However, in order to conclude one distribution is the best, its AD statistic must be substantially lower than the others. In this paper, we applied this test to the normal, weibull and log-normal distribution function. It confirms that the normal distribution give the best result after the procedure adopted to clean the data. Table 2 shows the AD statistics values before and after outlier removal using normal distribution. When the statistics are close together we use additional criteria, such as Q-Q plots, to choose between them.

Table 2: AD values before and after outlier removal using normal distribution

Normal distribution	After Outlier removal	Before Outlier removal
	22.6154	25.504

***B.* Descriptive analysis :** *Let us assume a data matrix* (X) *represented as :*

$$X = \begin{pmatrix} x_{11} & x_{12} & s_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & s_{23} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ . & . & . & \cdots & . \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{pmatrix}$$

where x_{ij} is the j^{th} variable (column) collected from the i^{th} item (rows) . The $n \times 1$ vector $\mathbf{x}_j$ gives the j^{th} variable's scores (of X for j ∈ {1, . . . , p})) for the n items and $1 \times p$ vector $x`_i$ gives the i^{th} item's scores (of X for $i \in \{1, . ., n\}$) for the p variables.

The sample mean of the j^{th} and i^{th} variable is given by Eq. (1) and Eq. (2) respectively

$$\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = n^{-1} 1_n^{'} x_j \tag{1}$$

$$\overline{x_i} = \frac{1}{p}\sum_{i=1}^{n} x_{ij} = p^{-1} x_i^{'} 1_p \tag{2}$$

where 1_n denotes as $n\times1$ vector of ones and 1_p denotes as $p\times1$ vector of ones.

a) *Mean and standard deviation* (*SD*) *calculation*: Initially mean & SD is calculated for all the ten combinations of TNF, EGF and Insulin for JNK protein. Table 3 shows the Mean & SD values.

Table 3 : Mean and SD of JNK

TNF-EGF-Insulin	Mean	SD
0-0-0	250.9833	9.19960
5-0-0	236.2242	8.30200
100-0-0	244.3748	11.69378
0-100-0	254.0393	7.45308
5-1-0	254.1922	10.28571
100-100-0	253.1534	8.17621
0-0-500	255.8353	7.62495
0.2-0-1	240.7743	15.03553
5-0-5	247.2728	9.07404
100-0-500	**204.1670**	**63.96239**

From Table 1 we can infer that 100-0-500 concentration of TNF-EGF-Insulin yields the best results.

b) *Covariance matrix* : The covariance matrix refers to the symmetric array of the numbers represented by matrix *S*.

$$S = \begin{pmatrix} s_1^2 & s_{12} & s_{13} & \cdots & s_{1p} \\ s_{21} & s_2^2 & s_{23} & \cdots & s_{2p} \\ s_{31} & s_{32} & s_3^2 & \cdots & s_{3p} \\ . & . & . & \cdots & . \\ s_{p1} & s_{p2} & s_{p3} & \cdots & s_p^2 \end{pmatrix}$$

where variance of the j^{th} variable and the covariance between the j^{th} and k^{th} variables are represented by Eq. (3) and Eq. (4) respectively

$$s_j^2 = {}^{1}\!/_{n}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)^2 \tag{3}$$

$$s_{jk} = \left({}^{1}\!/_{n}\right)\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)\left(x_{ik} - \overline{x_k}\right) \tag{4}$$

We can calculate the covariance matrix using Eq. (5).

$$S = \frac{1}{n} X_C^{'} X_C \tag{5}$$

where $$X_C = X - 1_n \,\overline{x}' = C\,X \tag{6}$$

with $\overline{x}' = \left(\overline{x}_1, \overline{x}_2 \ldots\ldots\ldots \overline{x}_p\right)$ and centering matrix (C) is expressed by Eq. (7)

$$C = I_n - n^{-1} 1_n 1_n^{'} \tag{7}$$

Covariance's between variables for different concentration of JNK is shown Table 4.

Table 4: Covariance's between variables

TNF-EGF-Insulin	Covariances									
0-0-0	84.633	12.392	-5.102	9.210	-32.968	18.479	-7.459	67.117	-27.719	350.925
5-0-0	12.392	68.923	39.666	1.218	-2.282	7.565	-0.911	55.378	-12.878	189.572
100-0-0	-5.102	39.666	136.745	-3.159	41.449	-2.820	-1.033	35.891	4.780	-19.562

0-100-0	9.210	1.218	-3.159	55.548	0.112	6.516	5.886	6.881	-4.371	44.563
5-1-0	-32.968	-2.282	41.449	0.112	105.796	-20.338	9.439	-59.332	32.009	-372.301
100-100-0	18.479	7.565	-2.820	6.516	-20.338	66.850	-2.186	39.525	-15.098	205.293
0-0-500	-7.459	-0.911	-1.033	5.886	9.439	-2.186	58.140	-12.321	7.590	-62.543
0.2-0-1	67.117	55.378	35.891	6.881	-59.332	39.525	-12.321	226.067	-66.631	809.699
5-0-5	-27.719	-12.878	4.780	-4.371	32.009	-15.098	7.590	-66.631	82.338	-347.964
100-0-500	350.925	189.572	-19.562	44.563	-372.301	205.293	-62.543	809.699	-347.964	4091.188

c) *Coorelation Matrix* : The Correlation matrix refers to the symmetric array of the numbers is expressed by matrix *R*.

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{21} & 1 & r_{23} & \cdots & r_{2p} \\ r_{31} & r_{32} & 1 & \cdots & r_{3p} \\ . & . & . & \cdots & . \\ r_{p1} & r_{p2} & r_{p3} & \cdots & 1 \end{pmatrix}$$

$$R = \frac{1}{n} X_S^{'} X_S \tag{8}$$

where $$X_S = C\,X\,D^{-1} \tag{9}$$

with centering matrix $C = I_n - n^{-1} 1_n 1_n^{'}$ and diagonal scaling matrix D = diag (s_1, $s_2 \ldots s_p$).

To quantify the relationship between feature variables, pair wise correlation test is performed and person correlation coefficients are obtained which are tabulated in Table 5.

We can calculate the correlation values by Eq. (8)

Table 5 : Pair wise correlation test for person correlation coefficients values

TNF-EGF-Insulin	Correlations									
0-0-0	1.000	0.162	-0.047	0.134	-0.348	0.246	-0.106	0.485	-0.332	0.596
5-0-0	0.162	1.000	0.409	0.020	-0.027	0.111	-0.014	0.444	-0.171	0.357
100-0-0	-0.047	0.409	1.000	-0.036	0.345	-0.029	-0.012	0.204	0.045	-0.026
0-100-0	0.134	0.020	-0.036	1.000	0.001	0.107	0.104	0.061	-0.065	0.093
5-1-0	-0.348	-0.027	0.345	0.001	1.000	-0.242	0.120	-0.384	0.343	-0.566
100-100-0	0.246	0.111	-0.029	0.107	-0.242	1.000	-0.035	0.322	-0.203	0.393
0-0-500	-0.106	-0.014	-0.012	0.104	0.120	-0.035	1.000	-0.107	0.110	-0.128
0.2-0-1	0.485	0.444	0.204	0.061	-0.384	0.322	-0.107	1.000	-0.488	0.842
5-0-5	-0.332	-0.171	0.045	-0.065	0.343	-0.203	0.110	-0.488	1.000	-0.600
100-0-500	0.596	0.357	-0.026	0.093	-0.566	0.393	-0.128	0.842	-0.600	1.000

If the value of a coefficient is close to 1, there will be a closer relation between the two data variables. If the coefficient value is the positive or negative but is very less, it implies that there might be less or no correlation between the variables. From the pair-wise Pearson correlation coefficients obtained for ten combinations are shown in Table 4, it is analyzed that 0-100-0 and 0-0-500 are least correlated to all other combinations. But further investigation is needed to be done in terms of descriptive analysis using *t*-test for optimal feature selection purpose. The best results for cell survival / death will be obtained from 100-0-500 concentration as maximum values are correlated.

C. Multiple Regression *:* The MLR considers graphical display of regression diagnosis, co linearity, variance inflation, and detection of regression outlier. The analysis of the variance shows the sum of squares (SS) and mean squares (MS) of the regression and residual error that can be are calculated by Eq. (10) and Eq. (11) and tabulated in Table 6.

$$F = \frac{MS_a}{MS_e} \qquad (10)$$

$$MS_a = \frac{SS_a}{df} \quad \text{and} \quad MS_e = \frac{SS_e}{df}$$

where MS_a is the average variability among groups, MS_e is the average variability within groups, df is the degree of freedom , SS_a is the sum of square among groups, SS_e is the error sum of square.

Table 6: Analysis of Variance using MLR

TNF-EGF-Insulin	Regression		Residual	
	SS	***F*-Ratio**	***SS***	***MS***
0-0-0	4093.4	57.51	21211.8	71.2
5-0-0	5100.4	98.01	20608.0	52.0
100-0-0	11272	113.42	29615	99
0-100-0	138.28	2.5	16470.7	55.27
5-1-0	15471	285.25	16162	54
100-100-0	1125.4	17.78	18862.9	63.3
0-0-500	0.81	0.01	17383.02	58.33
0.2-0-1	5.4	0.02	37588.7	226.8
5-0-5	8445.6	155.61	16173.5	54.3
100-0-500	**44461**	**11.24**	**1178804**	**3956**

a) *Regression Coefficients* : This is to get which of the independent variable contributes most to the prediction of JNK protein. Table 7 represents the standardized regression coefficient (*b**) and unstandardized / raw regression coefficient (*b*). The values of *b* coefficient enable us to compare the relative contribution of each independent variable in the prediction of dependent variable.

Table 7 : Regression Analysis for Dependent Variable JNK

N = 300	*r* = 0.872, *r*² = 0.760, Adjusted *r*² = 0.752 *F*(10,289) = 91.568, p < 0.0000, Std. Error of estimate: 0.01042					
	b*	**Std. Err.**	***b***	**Std. Err.**	***t* (289)**	***p*-value**
Intercept			0.923093	0.048952	18.8571	0.000000
0-0-0	0.057113	0.036235	0.000130	0.000082	1.5762	0.116080
5-0-0	-0.094268	0.034910	-0.000237	0.000088	-2.7003	0.007337
100-0-0	-0.424387	0.036171	-0.000759	0.000065	-11.7329	0.000000
0-100-0	-0.001984	0.029499	-0.000006	0.000083	-0.0673	0.946422

5-1-0	-0.189995	0.038463	-0.000386	0.000078	-4.9397	0.000001
100-100-0	0.038337	0.031480	0.000098	0.000080	1.2178	0.224284
0-0-500	0.016239	0.029397	0.000045	0.000081	0.5524	0.581115
0.2-0-1	-0.304610	0.059325	-0.000424	0.000082	-5.1346	0.000001
5-0-5	-0.020369	0.036182	-0.000047	0.000083	-0.5630	0.573897
100-0-500	-0.472229	0.072572	-0.000154	0.000024	-6.5071	0.000000

In Table 5, *r* is the multiple correlation coefficients which tell us how much concentration as whole correlation with the output value, r^2 is how much variance in JNK is accounted for by our model by concentration values as a whole. It means 76% variance in JNK protein is accounted for by the model includes all the predictors as *p* value is significant. Model as a whole is really predicting a value. Based on the *p*-value, author can say that 100-0-500 concentration values yields the best result for JNK which can be verified by importance plot.

b) *Correlation Coefficients :* Correlation coefficient become substantially inflated or deflates if extreme outliers are present in the data. This decision is somewhat subjective, a rule of thumb is that one needs to be concerned if any observation falls outside the mean ± three time standard deviation. In that case, analysis was again done with outliers and without outliers to ensure they did not affect the pattern of interconnections. The correlation is expressed by Eq. (12).

$$Cor\left(x_j, x_k\right) = \begin{cases} 1 & if\ \ j=k \\ r_{jk} & if\ \ j \neq k \end{cases} \qquad (12)$$

where r_{jk} is the Pearson correlation coefficient between variables x_j and x_k and is expressed by Eq. (13)

$$r_{jk} = \frac{s_{jk}}{s_j\, s_k} = \frac{\sum_{i=1}^{n}\left(x_{ij}-\overline{x_j}\right)\left(x_{ik}-\overline{x_k}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{ij}-\overline{x_j}\right)^2}\ \sqrt{\sum_{i=1}^{n}\left(x_{ik}-\overline{x_k}\right)^2}} \qquad (13)$$

To quantify the relationship between feature variables, correlation test is performed which are tabulated in Table 8.

Table 8: Correlation values

	Mean	**Std.Dv.**	***r*(X,Y)**	***r*²**	***t***	***p***
0-0-0	250.9833	9.199				
JNK	0.5057	0.0209	-0.287103	0.082428	-5.1740	0.000
5-0-0	236.2242	8.3020				
JNK	0.5057	0.0209	-0.549565	0.302022	-11.3555	0.000
100-0-0	244.3748	11.693				
JNK	0.5057	0.0209	-0.583082	0.339985	-12.3897	0.000
0-100-0	254.0393	7.453				
JNK	0.5057	0.0209	-0.036813	0.001355	-0.6359	0.525
5-1-0	254.1922	10.285				
JNK	0.5057	0.0209	0.016167	0.000261	0.2791	0.780
100-100-0	253.1534	8.176				
JNK	0.5057	0.0209	-0.179618	0.032263	-3.1520	0.001
0-0-500	255.8353	7.624				

JNK	0.5057	0.0209	0.083083	0.006903	1.4392	0.151
0.2-0-1	240.7743	15.035				
JNK	0.5057	0.0209	-0.709639	0.503588	-17.3870	0.000
5-0-5	247.2728	9.074				
JNK	0.5057	0.0209	0.318491	0.101437	5.8000	0.000
100-0-500	**204.1670**	**63.962**				
JNK	**0.5057**	**0.0209**	**-0.584675**	**0.341845**	**-12.4411**	**0.000**

From the Table 8 it can be interpreted that 0-0-500, 100-100-0 and 0-100-0 are not significant so can be ignored for further analysis.

t-test and Analysis of Variance (ANOVA) statistical tools were employed in this work using design of experiment application of Statistica Software. *t*-test and ANOVA are applicable for normally distributed set of continuous data. The statistical significance of *t*-test and ANOVA is given by the significance value (*p*-value) of the test which should be less than 0.05. We have calculated ANOVA using multiple regression analysis (shown in Table VI), regression analysis for JNK (shown in Table VII) and Correlation values (shown in Table VIII) which also depicts that 100-0-500 concentration will give the best results. The best concentration is used further for the experimentation which will result in cell survival / death.

D. Model Validation *:* Lastly we calculate the variable importance values from where we come to know that which concentration helps more in analysis of survival/ death for JNK protein. Fig 7 shows the variable importance plot. It was seen that 100-0-500 yields best results and 0-100-0 yields least importance which can be discarded.

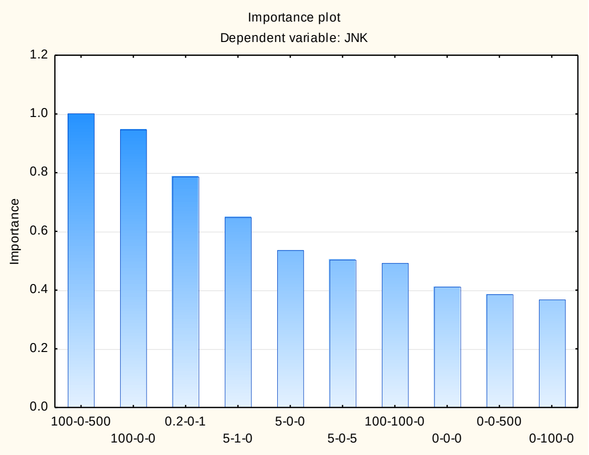

Figure 5: Variable importance plot

With the help of mean decrease in Gini, variable importance plot lists the most significant variables in descending order. The top variables contribute more to the model than the bottom ones. The prediction power is high which helps in recognizing whether value is default or not.

V. CONCLUSION

In this paper, we are proposing a model of JNK protein-mediated phosphorylation on known substrate proteins using different concentrations of TNF, EGF and Insulin (inputs) that control the survival/ apoptosis response of HT-29 human colon carcinoma cells using multiple linear regression. Exhaustive statistical investigation of the proposed approach using visual, graphical and statistical analysis is done on experimental data. Standard regression analysis was used to assess the most significant contribution assuming p-value ≤ 0.05 as significant. Based on *p*-value and importance plot, 100-0-500 concentration of TNF-EGF-Insulin was achieved for the proposed system with significantly less complexity. Results obtained reveals that this proposed system can be used for detection of survival or apoptosis for any protein. In future, the proposed algorithm will be used for other marker proteins which yield due to the combination of inputs.

REFERENCES :

1. A Dhiman, A Singh, S Dubey, S Jain, "Design of Lead II ECG Waveform and Classification Performance for Morphological features using Different Classifiers on Lead II *",Research Journal of Pharmaceutical, Biological and Chemical Sciences*,7(4), 1226- 1231: 2016.
2. S Bhusri , S Jain, J Virmani , "Classification of breast lesions using the difference of statistical features" Research Journal of Pharmaceutical , Biological and Chemical Sciences,7 (4), 1365-1372: July- Aug 2016
3. S Sharma, S Jain, S Bhusri, "Two Class Classification of Breast Lesions using Statistical and Transform Domain features", Journal of Global Pharma Technology ,9(7), pp 18-24, 2017.
4. A. V. Alvarenga, A. F. C. Infantosi, W. C. A. Pereira, and C. M. Azevedo, "Assessing the performance of morphological parameters in distinguishing breast tumors on ultrasound images," Medical engineering & physics, vol. 32, no. 1, pp. 49–56, 2010
5. SJain, "Classification of Protein Kinase B Using Discrete Wavelet Transform", International Journal of Information Technology, 10(2), 211-216, 2018.
6. S Jain. "System Modeling of AkT using Linear and Robust Regression Analysis ", Current Trends in Biotechnology and Pharmacy, 12 (2), 177-186, April 2018.
7. S Jain, "Regression modeling of different proteins using linear and multiple analysis", Network Biology. 7(4), 80-93, 2017.
8. S Jain, "Implementation of Marker Proteins Using Standardized Effect", Journal of Global Pharma Technology (JGPT), 9(5), 22-27: 2017.
9. Li Yanglong et al. Journal of Shandong Foreign Languages Teaching. 2013(2):56-51.

10. Bao Gui. Foreign Languages. 2012.28(4):63-68.
11. Wang Zongying et al. Journal of Basic English Education. 2011.13(5):7-15.
12. Ma Hongxiang. English Square·Academic research. 2013(31).
13. Qiu qiufen et al. Journal of Hunan Institute of Humanities, Sciences and Technology. 2013(2):102-105.
14. Gaudet S., Kevin J.A., John A.G., Emily P.A., Douglas L.A, and Peter S.K., "A compendium of signals and responses trigerred by prodeath and prosurvival cytokines', Manuscript M500158-MCP200, 2005.
15. S Jain. 2012, Communication of signals and responses leading to cell survival / cell death using Engineered Regulatory Networks. PhD Dissertation, Jaypee University of Information Technology, Solan, Himachal Pradesh, India.
16. S Jain, D. S. Chauhan, " Computational Analysis of MK2 protein for HT Carcinoma Cells using Pre-Processing and Characterization Techniques for Cell Death/ Survival ", Journal of Global Pharma Technology (JGPT), 10(11),13-19: 2018.
17. S Jain, "Classification of Mitogen Activated Protein Kinase using Different Wavelet Transforms (Discrete and Gabor)", Asian Journal of Microbiology, Biotechnology and Environmental Sciences. 20(2), 569-574, 2018. UGC No : 8799
18. J. T. McClave, and T. Sincich, (2006). Statistics (10th edition). Upper Saddle River, NJ: Pearson Prentice Hall.
19. A. C., Tamhane, and D. D. Dunlop, (2000). Statistics and Data Analysis: From Elementary to Intermediate (1st edition). Upper Saddle River, NJ: Pearson Prentice Hall.
20. S Jain, "Detection of Arrhythmia using Automatic Differentiation Method", International Research Journal of Natural and Applied Science, 5(7), 9-24, 2018.

AUTHORS PROFILE

Shruti Jain is Associate Professor in the Department of Electronics and Communication Engineering at Jaypee University of Information Technology, Waknaghat, H.P, India and has received her Ph.D in Biomedical Image Processing. She has a teaching experience of around 14 years and before joining JUIT, she worked as Assistant Professor in Haryana Engineering College, Jagadhari, Ambala College of Engineering, Ambala. She has specialization in Biomedical Signal Processing, Computer- Aided design of FPGA and VLSI circuits, combinatorial optimization. She has published more than 50 papers in reputed journals and 30 papers in International conferences. She is a senior member of IEEE, life member and Editor in Chief of Biomedical Engineering Society of India and member of IAENG. She has completed one externally funded project and one in pipeline. She has guided 1 PhD student and now has 5 registered students. She is a member of Editorial Board of many reputed journals. She is also a reviewer of many journals and a member of TPC of different conferences. She was awarded by Nation Builder Award in 2018-19.