بسم الله الرحمن الرحيم

Discovering and analysing Regulatory genes involved in Allergic contact dermatitis disease due to Nickel exposure

Muhammad Kundi
Auckland University of Technology (AUT)
pqb2700@aut.ac.nz

Abstract— **Genetic regulatory network inference is based on gene expression data.This inference can be challenging if we are dealing with the analysis and classification of Spatio-and Spectro-Temporal Data (SSTD) as traditional machine learning techniques like Support Vector Machine and Multi-Layer Perceptron can ignore the spatial and temporal components. To deal with SSTD data, researchers have developed a new type of data processing architecture under the name of brain like Spatio-Temporal Data Machines (STDM).The framework used for this algorithm is known as "NeuCube".In this paper we are trying to find the regulatory genes involved in Allergic contact dermatitis disease patients exposed to Nickel. The experimental data provided has 4 time points and two classes of subjects who are exposed to nickel. This data is an example of SSTD with binary classification. Therefore, NeuCube framework is used for classification and identification of the regulatory genes. With the help of Signal Noise Ratio (SNR) , top 8 genes are selected using NeuCom framework.These 8 selected genes are used for classification experiment in the NeuCube framework.The accuracy of clustering the data in this binary classification model is 100%. CST9 ,ADAM28 and DDX59 are found to be the top 3 regulating genes making contribution of 27,25 and 18 percent respectively. Also strong interaction is found in CST9 with DDX59 ,CST29 with ADAM28 and DDX59 with TNPO1 .**

I. INTRODUCTION

A challenge in gene expression studies is the identification of discriminative genes, which may be later used as predictors (inputs) to classification models. Removing irrelevant features may lead to improved accuracy and increased interpretability of the classification model. However, this task is challenging, especially when we are dealing with classification of Spatio- and Spectro-Temporal Data(SSTD) such as gene data.

Most widely used model for temporal data is based on traditional artificial neural networks (ANN). But this cannot analyse complex spatial/spectral components as they overlook the temporal dimension . [1]

Large extent of information in Spectro-temporal data is implicit in the relationship between the spectral and temporal part of the SSTD. [2]. This information is lost if we use traditional machine learning techniques such as Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM). [3]. In order to analyse and model these dynamics in a suitable way we get inspiration from the most powerful adaptive learning device i.e. the human brain. Human brain has the ability to learn from experience, work in complicated and noisy environments and above all has the ability to adapt. [4]

This novel type of data processing architectures and algorithms is developed by The Knowledge Engineering and Discovery Research Institute (KEDRI)1) .Its given the title of Brain-like Spatio-Temporal Data Machines (STDM). STDM is actually the architecture for predictive data modelling of temporal or spatio/spectro-temporal data (SSTD)

The STDM developed by KEDRI is a brain-like computational framework which uses artificial spiking neural networks (SNN) as the computational measurements. [5]

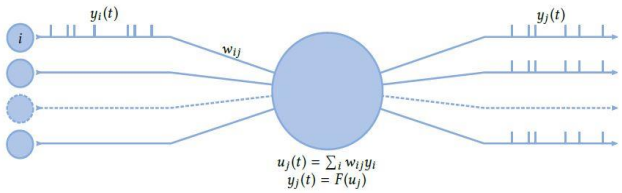

Figure 1:Simplified spiking (temporally dependent) artificial neuron.$\boldsymbol{y_{i(t)}}$ are the input spike train from neuron i; $\boldsymbol{w_{i(j)}}$ is the weight of an synapse between neuron i & j . $\boldsymbol{u_{j(t)}}$is the internal membrane voltage .F($\boldsymbol{u_j}$)is the neuron model $\boldsymbol{y_{j(t)}}$is the output spike train from postsynaptic neuron j. **[6]**

In order to model large scale efficient SSTD, a common design of evolving spatio/spectro-temporal data machines (eSTDM) is proposed as shown in Fig 3.

The main features of eSTDM design are

1. Changing multivariate date in to sequence of spikes
2. Spatio-temporal patterns of data learned through unsupervised learning in a SNN reservoir known as the "Cube".

3. Using supervised learning of regression and classification output system for regression/classification models.
4. Doing optimization using the tested accuracy of the model as a feedback to enhance performance of this system. [5, p. 3]

In 2014, a new platform NeuCube (Kasabov,2014) was developed for evolving spike neural network eSNN system.This was basically developed for spatio-temporal brain data modelling but further development was made for other types of data as well.Fig-2 shows architecture of NeuCom.

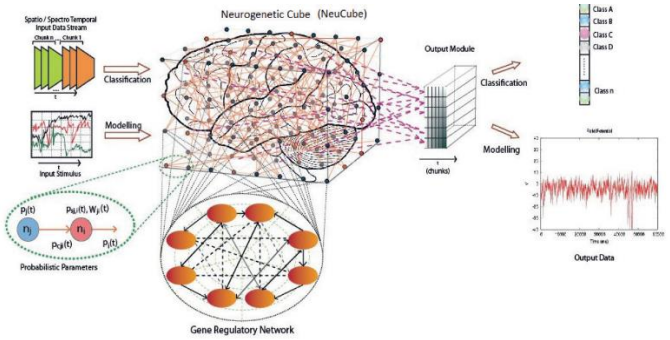

Figure 2 NeuCom Architecture. From left first section is Input Encoding, then NeuCube reservoir and third is Output and Classification section. **[12]**

In this paper we will try to discover the regulator genes and their interactions in the skin of nickel allergic patients. The data collected as a result of analysis of skin biopsies from nickel-allergic patients whose skins were exposed to nickel, for different time periods. In this experiment we have to deal Spatio- and Spectro-Temporal Data(SSTD) as this involves measuring spatial and spectral variables over time. As discussed before, using traditional machine learning models for SSTD data may result in loss of information. Therefore ,to find the effective genes and their interaction of nickel allergic patients we will make use of STDM uses artificial spiking neural networks (SNN) for computational measurements. We will achieve our goal using NeuCube which is the latest architecture for STDM methodology.

The rest of the paper is structured in the following sections. In part-II, information about the data is described. In this section pre-processing steps of data using MS Excel and Matlab along with feature selection using NeuCom are described. In part-IV ,data analysis procedure using NeuCube is discussed followed by part-V which tells about the results of the experiment. In the final part-VI, conclusion and future work is discussed.

II. Task Speciation And Description of Data

Allergic contact dermatitis (ACD) is a form of contact dermatitis induced by contact with a substance; the other type is irritant contact dermatitis (ICD).ACD is accepted as the most prevalent type of immunotoxicity found in humans, although less frequent than ICD. [7]

The aim of the study is to find the biological mechanism of the disease that shows the morphological features that is leading towards the diseased condition. In other words we are trying to find the genes involved in the inflammatory response resulting in allergic contact dermatitis (ACD).The interaction between different features is also to be discovered. For this analysis of ACD,one group of patients is exposed to Nickel and the other is the control class .The skin biopsies samples are taken at four different time points i.e no exposure, 7 hours, 48 hours and 96 hours. seven nickel-allergic patients and five nonallergic controls . Results are to be analysed to get an insight into molecular mechanism underlying pathogenesis of ACD. The experiment took place in Gentofte Hospital, University of Copenhagen, Copenhagen by The National Allergy Research Centre [8]

The full-dataset includes 34 samples with 54675 features.The data can be divided in two classes i.e “Nickel-allergic” and “Non allergic control”.Description can be summarized in the following table.

Figure 3 : Allergic contact dermatitis: time course **[9]**

Since the number of nickel-allergic patients is different from the non-allergic controls, I have to deal with missing data. I separated the control class with nickel class in separate MS Excel files.So we have a total of 8 files, 4 for control class and 4 for Nickel class. Each class has 4 time stamps i.e 0,7,48 and 96 hrs. Some of these individual class files have missing time stamps. To keep data consistent, I inserted the missing time stamp column and filled it with NaN values.

First of all, I wanted to find select the most significant features using Signal to Noise Ratio(SNR) feature in NeuCom.To prepare data for this task, I imported the MS Excel files in Matlab and saved the output as transposed Numeric Matrix and filled the NaN values using linear interpolation. Then I concatenated the arrays vertically using their mean values and saved the output as "data_for_neucom”. Created class labels where 1 represented the “Nickel-allergic” and 2 represented the “Non allergic control” and concatenated these class labels horizontally with the previously saved "data_for_neucom" .

Now I used “NueCom” framework to find the most suited genes for our classification model. First I chose to display the top 20 features in the data with less SNR.The result of corresponding SNR values is shown below

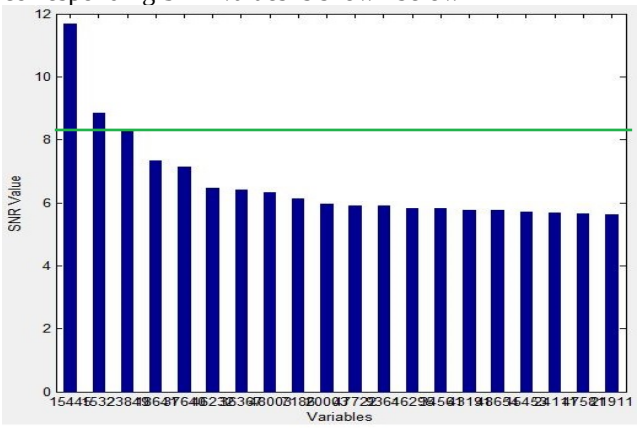

Figure 4 NeuCom output for features to be selected among 20 features (starting from the variable with the highest SNR value).

We can see that the SNR ratio values for this dataset shows that majority of the features have SNR> 8 (or 9 approximately).As a result, I decided to use the 8 top features with least SNR values for the experiment in NeuCube.Fig-4 shows the result of selected 8 genes with lowest SNR values.

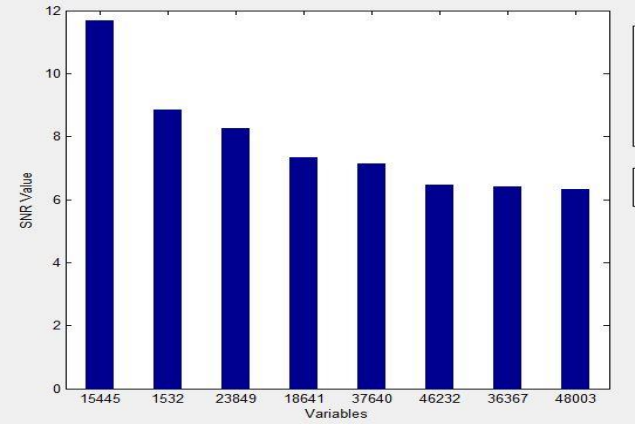

Figure 5 NeuCom output for 8 features to be selected (starting from the variable with the highest SNR)

The selected genes names are shown in the following table 1.

ADAM28	DDX59
CST9	BE467566
SPRR1A	ZBTB34
TNPO1	NCS1

Table 1 List of 8 selected genes after SNR analysis

III. EXPERIMENTAL DESIGN

After selecting the most significant features, NeuCube(v1.3)was used for encoding the spike, initializing and training the Cube and finally initializing and training the classifier. Mainly two experiments were performed with different parameters for tuning purpose. These two experiments were carried out using different spiking thresholds. Further changes in parameters is done for each experiment in order to improve the results. Experiment-1 was performed using Spiking Threshold of 0.02.and Experiment-2 was performed using the spiking threshold of 0.01.

The steps involved in the experiment can be explained clearly by briefly looking in to the block diagram of NeuCube.

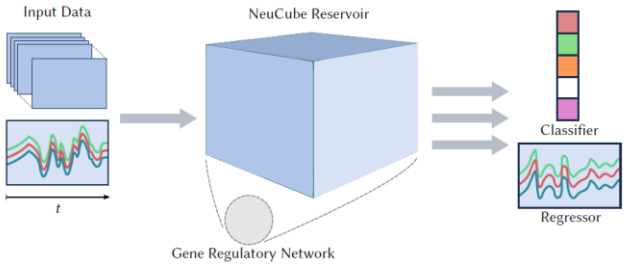

Figure 6 : Block diagram of NeuCube.In the left,we see input encoding module feeding the main NeuCube 3-D reservoir. On the right, we the out put or classification component along with the optional gene regulatory network. **[10]**

There are 3 main sections of NeuCube framework i.e..

1. Encoding section for Input data
2. 3D Spiking Neural Network section
3. Output for Regression/Classification

The input module in NeuCube, transforms the input data into an array of spikes. This spacio-temporal data is then fed into the main module i.e. the Cube. The cube reservoir captures these spatial links through its 3D representation of neurons. Finally, the classification section provides the classification of the data. There is also an optional gene regulatory component which is modulating the behaviour of the reservoir. [1]

The above three stages are involved in both the experiments carried out in NeuCube with the data of nickel exposure experiment for Allergic contact dermatitis.

After the previously described feature selection, the data is equally divided into a training set (i.e. seen set) and the testing set (i.e. unseen set) training data. This preparation is done using Matlab.4 data points and 8 features are selected. 4 samples are selected for the "Nickel-allergic" and 4 for the "Non allergic control class.The data is loaded onto NeuCube and Classification task type is selected.

A. *Experiment -I*

The data is first transformed into a train of spikes using "Thresholding Representation" (TR) method as encoding method with the spiking threshold as 0.02. In Fig 7 the spike encoding visualization is shown for feature CCST9 with sample 8.

After, encoding, the next step is to initialize the cube. In this step SSTD is mapped onto a 3D SNN cube. I used automatic Neuron Coordinate with 10 neurons on each X,Y and Z coordinate. The following result is obtained with the 8 features.

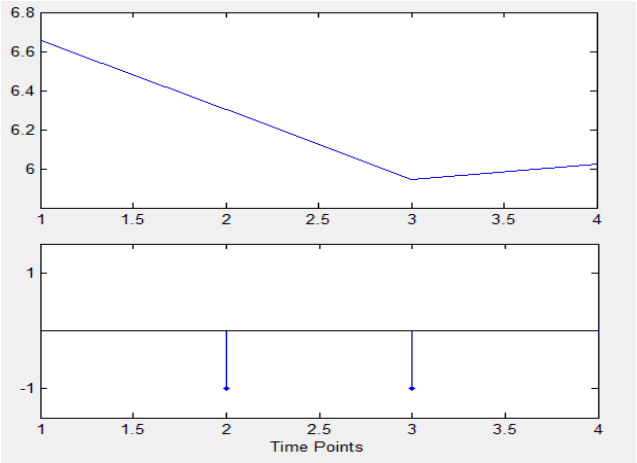

Figure 7 Spike encoding visualisation of feature CST9

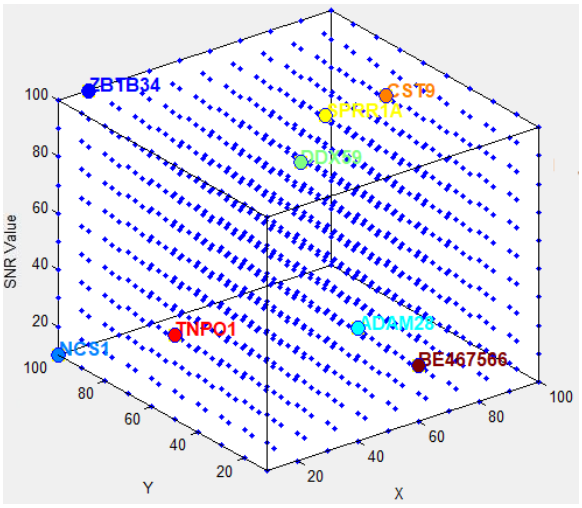

Figure 7 Cube Initialisation for experiment-I

Next, process after initialisation is training the cube.As mentioned before that the data is evenly divided in to training and testing .Therefore 50% data is used for training. 3 out 6parameters values were adjusted for training i.e. Potential Leak Rate=0.002, Training iteration=6 and LDC Probability=1.The rest of the parameters were kept at default value i.e. STDP Rate=0.01, Firing Threshold=0.5,Refractory Time=6.

After training the the Cube, its time to train the classifier.I used deSNNs classifier with adjusted parameters of Mod=0.8, Drift=0.005,K=4 and Sigma=1.

Final step is to verify our classification. After verifying I got 75% accuracy result for this experiment. The result is visualised in Fig-10

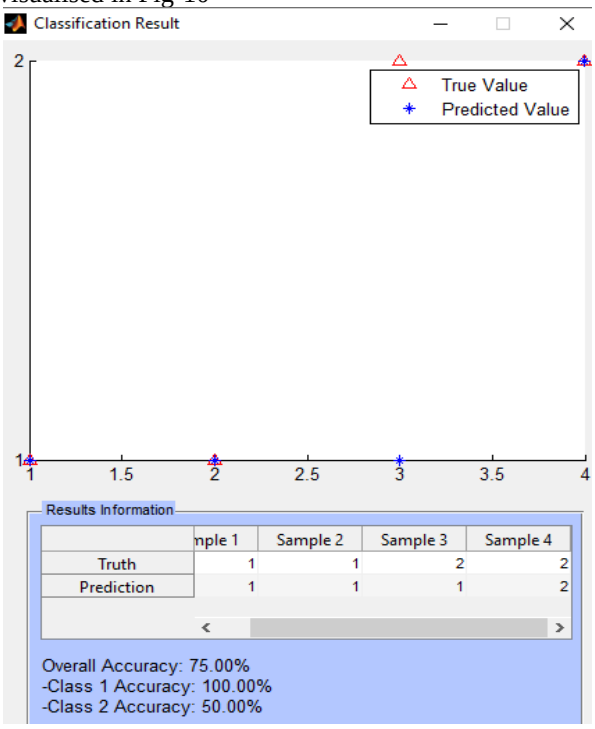

Figure 8 Visualisation of Cube after training(the visualisation show no difference than the visualisation after Cube initialization

B. *Experiment -II*

In experiment-ii,the data is first transformed into a train of spikes using "Thresholding Representation" (TR) method as encoding method but with the spiking threshold = 0.01.

Next, in the training step,I have kept the same 50% ratio between training and testing data . I changed the parameters of Potential Leak Rate=0.002, Training iteration=1 and LDC Probability=0.The rest of the parameters are not changed i.e. STDP Rate=0.01, Firing Threshold=0.5 ,Refractory Time=6.

To train the classifier in experiment-,I used deSNNs classifier with Mod=0.2, Drift=0.01,K=6 and Sigma=0.1.

When this classifier for Experiment-2 is verified 75% accuracy is achieved.

Keeping all the parameters same, I tried to tune the DeSNNs Classifier parameters.I changed the value of Drift=0.005 and K=3.After verification I got 100% accuracy for this Classification experiment.The result is shown in figure-11.

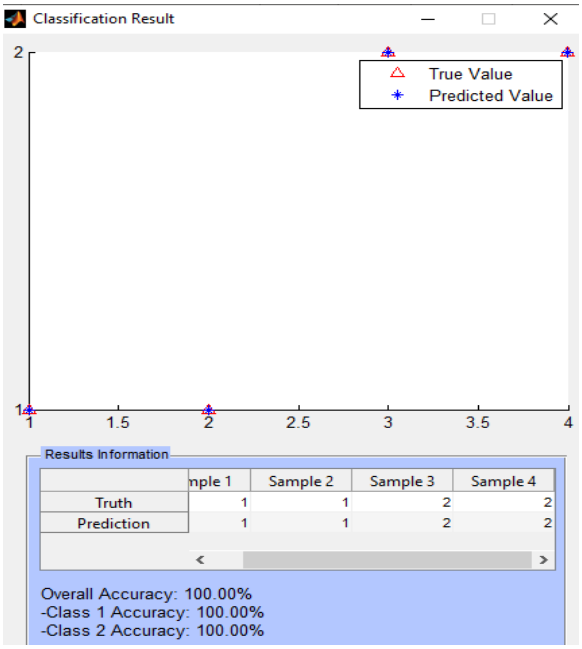

Figure 9 Classification accuracy result of Experiment-2 in NeuCube

Because of 100% accuracy, I only selected results from experiment-2 for further analysis.

IV. RESULTS AND DISCUSSION

Experiment-2 showed the following results when we verified the accuracy in NeuCube.

	Sample 1	Sample 2	Sample 3	Sample 4
Actual	1	1	2	2
Prediction	1	1	2	2

Table 2 Classification accuracy results from NeuCube experiment.

As already mentioned, Class 1 represents the "Nickel-allergic" and Class 2 represented the "Non allergic control"

The result can be more elaborated in the form of confusion matrix which is depicted in Fig-13.

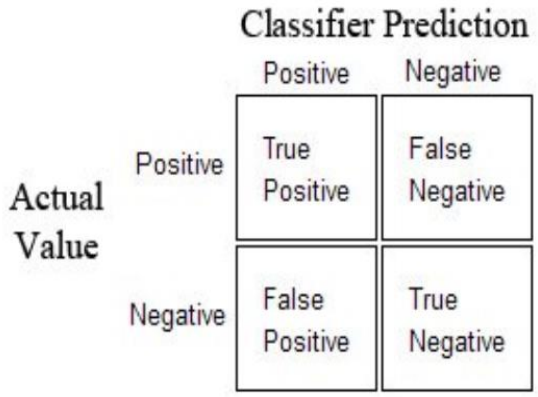

Figure 10 Confusion Matrix (Source: Researchgate)

For the results of experiment-2,the confusion matrix is shown in Table 4:

	Predicted Allergic	Predicted Non Allergic Control
Actual Allergic	2	0
Actual Non Allergic Control	0	2

Table 3 Confusion Matrix for Classification Experiment

This result shows that both Class 1 ("Nickel-allergic") and Class 2 ("Non allergic control") have 100% percent accuracy.

Further measurement of our results can be done with "precision" and "recall" calculations. These can be explained from Fig-12

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

Figure 11 Precision and Recall definition. Here tp=True positive ,fp=False positive & fn=False negative. **[11]**

This experiment has the perfect precision value of 1.0. This shows that all the information received was relevant .But precision does not tell us whether all relevant information was received. For this purpose we use the "recall". This experiment also has a perfect recall score of 1.0 which show that all relevant data was retrieved by the search.

We can say that we were able to correctly classify our data with in the two classes of "Nickel-allergic" and "Non allergic control". Based on these results we can identify the regulatory genes involved in Allergic contact dermatitis disease.

The GRN (Gene Regulatory Network) as the result of this experiment for these genes can be visualised in NeuCube.This GRN is shown in Fig 13.

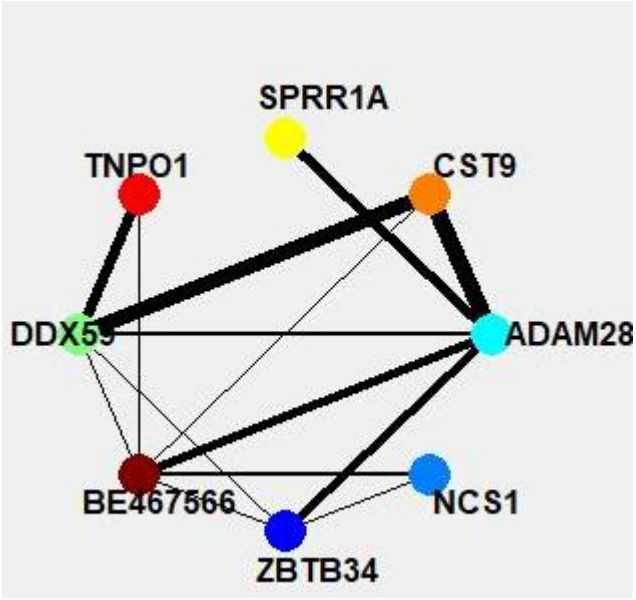

Figure 12 GRN visualisation showing total interaction between genes. Stronger lines show stronger interaction

The thickness of the connecting lines shows the strength of interaction between the genes.This result shows strong interaction between CST9 with DDX59 , CST9 with ADAM28. Second level of stronger interaction is shown between ADAM28 with BE467566 and DDX59 with TNPO1.We can further verify these interactions if we check the proportions of the regulatory genes.

NeuCom also has the ability to show the proportions of the regulatory genes as a result of the classification experiment.This is shown in Fig 14.

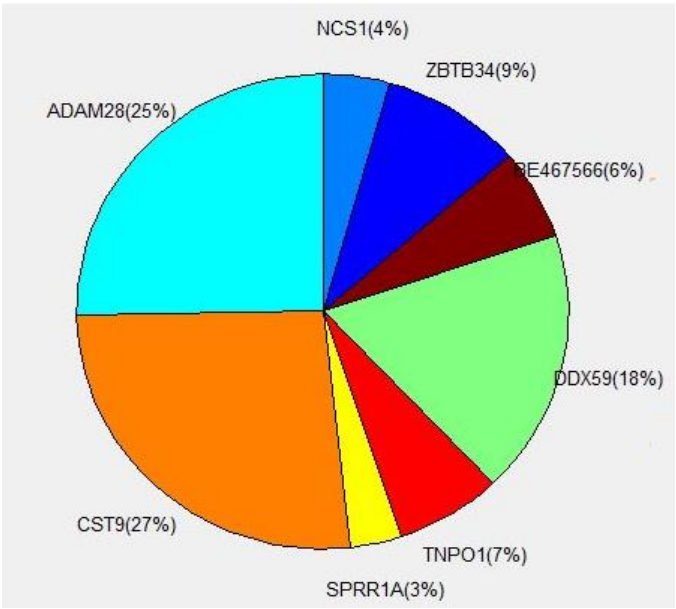

Figure 13 Neuron Proportion of the top 8 regulatory genes

This result shows that CST9 ,ADAM28,DDX59 & ZBTB34 have the most proportion among the regulatory genes as 27 , 25 18 and 9 percent respectively. This also verifies about the strong interaction among these genes in the Fig 13. The remaining 4

of the top 8 genes i.e., TNPO1 , BE467566,NCS1 and SPR1A have 6,7,4 & 3% percent proportion.
The following table describes briefly these 4 top regulatory genes involved in the experiment of analysis of skin biopsies from nickel-allergic patients.

Ranking	Name of Genes	Description
1	CST9	The superfamily of cystatin contains proteins that contain multiple sequences identical to cystatin. Some of the leaders are effective inhibitors of cysteine protease, while others have lost this inhibitory function or perhaps never acquired it. [9]
2	ADAM28	This gene encodes a disintegrin and metalloprotease family member of the ADAM family. Members of this family are membrane-anchored proteins that are structurally associated with snake venom disintegrins and are involved in a variety of biological processes. [9]
3	DDX59	Ubiquitous expression in bone marrow (RPKM 2.6), thyroid (RPKM 2.3) and 25 other tissues. [9]
4	ZBTB34	Ubiquitous expression in bone marrow (RPKM 6.2), placenta (RPKM 3.9) and 25 other tissues. [9]
5	TPNO1	This gene encodes the karyopherin receptor complex's beta subunit which interacts with nuclear localization signals to target nuclear proteins in the nucleus. [9]

Table 4 Brief introduction of top 5 regulatory genes in the experiment

V. Conclusion and Future Direction

The performance of the classification experiment using NeuCube to find regulatory genes involved in Allergic contact dermatitis was excellent. We achieved hundred percent accuracy result. Class prediction for both "Nickel-allergic" and "Non allergic control" subjects was perfect .We found the proportion of the top regulatory genes related to the experiment to study the effect of Nickel exposure Allergic contact dermatitis patients. The top 3 regulatory genes are CST29 ADAM28 and DDX59 having the most significant proportion.

Also we found very strong interaction between CST9 with DDX59 , CST9 with ADAM28. Also significant level of interaction is discovered between ADAM28 with BE467566 and DDX59 with TNPO1.Apparently the results seems but promising but there are some shortcoming in this experiments. First shortcoming is that the number of subjects is not sufficient to draw final conclusion. Also for time series analysis we only had 4 time points.

For future work, it is suggested that for better and more reliable results, this experiment can be repeated with more number of subjects. Also more time points should be considered for this experiment because more time slots can provide more information for the time series analysis.

But as long these two requirements are not fulfilled, these discovered regulatory genes can be focused to do research in providing cure for the Nickel contact allergies.

VI. References

[1] M. G. D. ,. Z. G. D. N. K. Kasabov, "Mapping, Learning, Visualization, Classification, and Understanding of fMRI Data in the NeuCube Evolving Spatiotemporal Data Machine of Spiking Neural Networks," IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 4, pp. 887-899, April 2017.

[2] N. Scott, "Evolving Spiking Neural Networks for Spatio- and Spectro- Temporal Data Analysis :Models, Implementations, Applications," p. 1, 2015.

[3] N. Scott, "Evolving Spiking Neural Networks for Spatio- and Spectro- Temporal Data Analysis :Models, Implementations, Applications," p. 12, 2015.

[4] N. Scott, "Evolving Spiking Neural Networks for Spatio- and Spectro- Temporal Data Analysis :Models, Implementations, Applications," p. 4, 2015.

[5] N. Kasabov, N. Scott, E. Tu, S. Marks, N. Sengupta, E. Capecci, M. Othman, M. Doborjeh and N. Murli, "Evolving spatio-temporal data machines based on the NeuCube neuromorphicframework: Design methodology andselected applications. Neural Networks," 2015.

[6] N. Scott, "Evolving Spiking Neural Networks for Spatio- and Spectro- Temporal Data Analysis :Models, Implementations, Applications," p. 22, 2015.

[7] K. I, B. DA, G. GF and D. RJ, "Allergic contact dermatitis," Int. Immunopharmacol, vol. 2 , no. (2–3), p. 201–11, (2002).

[8] M. Pedersen, S. L, M. T and J. J. e. al, "Gene expression time course in the human skin during elicitation of allergic contact dermatitis," J Invest Dermatol, p. 127(11), 2007.

[9] The National Center for Biotechnology Information, "NCBI," [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/tools/profileGraph.cgi?ID=GDS2935:203290_at. [Accessed 30 10 2019].

[10] N. Scott, "Evolving Spiking Neural Networks for Spatio- and Spectro- Temporal Data Analysis :Models, Implementations, Applications," p. 71, 2015.

[11] Olson and D. L, Advanced Data Mining Techniques, 1st edition ed., Springer, 2008.

[12] N. Kasabov, "NeuCube: a spiking neural network architecture for mapping," Neural Networks, pp. 52,62-76, 2014