Pakistan Journal of Engineering and Technology (PakJET)
Multidisciplinary & Peer Reviewed
Volume: 5, Number: 1, Pages: 16- 22, Year: 2022

GVDeepNet: Unsupervised Deep Learning Techniques for Effective Genetic Variant Classification

Ghulam Muhammad[1], Umair Saeed[2], Noman Islam[3], Kamlesh Kumar[2], Fahad Hussain[2], Mansoor Ahmed Khuhro[2], Aftab Ahmed Shaikh[2], Iqra Ali[4]

[1]Department of Computer Science, Bahria University, Karachi, Pakistan

[2]Department of Computer Science, Sindh Madressatul Islam University, Karachi, Pakistan

[3]College of Computing and Information Sciences, PAF KIET, Karachi, Pakistan

[4]Department of Computer System Engineering, NED University of Engineering and Technology, Karachi, Pakistan

Corresponding author: Noman Islam (e-mail: noman.islam@kiet.edu.pk).

***Abstract*- This paper proposes a novel machine learning approach to genetic variant classification based on un-supervised learning. During the past few years many lives have been lost due to genetic diseases and the inbility to identify them. The genetic disorder is mainly because of the alteration in the common DNA nucleotide sequence. Benign and pathogenic are the common examples of these genetic variants. Deliberate changes for the gene mutation may cause unexpected results. In this paper, two unsupervised deep learning classification methods have been proposed to classify these genetic changes. For this work, self-organizing maps (SOM) and autoencoder models have been used. SOM is an unsupervised learning technique used to obtain a low dimensional representation of the data. The SOM has been implemented using MiniSOM library. Autoencoder comprises an encoder and decoder component. The information encoded by encoder is decoded using the decoder component to obtain as close representation to the input as possible. The analysis were performed on ClinVar dataset comprising 6 lac records. The dataset is publicly available at https://www.ncbi.nlm.nih.gov/clinvar/. The data was first subjected to pre-processing to handle missing and duplicate values. The result showed the good performance of autoencoder, where its accuracy is 97% (on Test Data), and SOM has an accuracy of 96% (on Test Data). It has been concluded that unsupervised deep learning models, SOM and autoencoder retain enough prediction power to classify and identify genetic variants i.e. if the underlying alternation in the gene gives positives changes or the contrary case.**

***Index Terms*-- Genetic Variant; DNA nucleotide sequence; Classification; Deep Learning; Autoencoder; Self-Organizing Map.**

I. INTRODUCTION

Genome contributes to the differences in the body, especially human's eye, and blood types. This genetic information, along with the other forms of replicating genetic information in the human body is encoded in Deoxyribonucleic acid (DNA) or Ribonucleic acid (RNA). The primary structure of chemical or hypothetical nucleic acid is based on nucleotides or modified nucleotides of genetic information [1-3].

Genetic information is widely examined to identify other critical diseases, especially cancer. Many studies on the subject of cancer genome are under observation, including the miRNA analysis [4-6]. Similarly, information related to genetic invariant is the most authenticated resource to detect diseases such as Alzheimer's disease [4]. A well-known approach in terms of changes of the gene is observing the changes in miRNA regulome that can be measured and controlled by medical treatment for cancer patients. The miRNA is exceptionally connected to multiple miRNAs and one miRNA si-lences variety of genes [5]. Another innovation in the field of genetic variant is Next-generation sequencing (NGS) technology, which has marvelous im- provements in the last decades; this NGS defines a DNA sequencing technology that has revolutionized genomic research tremendously [7-14].

Nonetheless, machine learning has been expeditiously infiltrating in medicine and related discipline. For instance 26 have used machin learning for hand gesture recognition. 27 have used machine learning for speech recognition. Likewise, it has done tremendous work in the field of genetic invariants and cardiovascular medicine. Automated risk prediction is an AI-based algorithm that is used to handle and guide clinical care; it is also helpful for the handling of complex diseases through unsupervised learning techniques, which is the primary technique used in this paper [15-17]. Alternatively, support vector machine, Decision Tree and Random Forest are a few supervised machine-learning algorithms that have potential applications for cancer detection [7]. The approach of unsupervised learning is to classify cancer Diagnosis using the Hidden Markov Model, which is a broadly used machine learning technique in which we oversee the change in the DNA in the human genome, which is a form of genetic variation [18, 19].

II. LITERATURE REVIEW

Many research works have been presented with the underlined efficient algorithms for the identification and classification of genetic variants, and this research is con- tinued with an incredible pace. Among the most prevalently used algorithms, CADD is the one that is designed to annotate coding and noncoding variants [15].

The authors in [12] proposed to identify the noncoding variants in human genetics. A deep learning framework is used to predict the noncoding version de noveo from the sequence of genetic, known as DeepSEA. A computational method for classifying the prioritization and interpretation using a deep neural network technique has been discussed [13]. A deep-learning based ab initio predictive variant model has been proposed to predict and annotate patters for the related risky diseases [11].

The human splicing code reveals new insights into the genetic determinants of disease based on computational technique counts how strongly genetic variants affect RNA splicing which is one of the main steps in gene expression causing many diseases, including the neurological disorders and cancer disease [18]. Machine learning and ANN are covering the great effects in DNA classification for cancer patients. A hybrid convolutional and recurrent deep neural network is used for quantifying the function of DNA sequences [16]. A classification technique to classify the cancer diseases using machine learning clasification algorithms artificial neural network, k-nearest neighbors, decision trees, fuzzy classifier, Navies Bayes classifier, random forest and support vector machine [7] . In another work, the DNA genome classifies and identities the cancer of prostate. The author has used metastatic castration-sensitive prostate cancer (mCSPC) to a castration-resistant (mCRPC) state that signals the lethal phenotype of prostate cancer [13]. The approach of paper is to align and find the cluster through the optimization of multiple neural networks using three viral genome and gene datasets, averaging 1300 sequences each [14].

The DNA and RNA functionalities can be quantified by implementing a hybrid convolutional and recurrent deep neural network [1]. DeepBind is the deep-learning based tool which predicts the sequence of DNA and RNA binding protein and it is capable of processing a million of such sequences [2]. Predictive analysis based on neural network is the continuously evolving field, researchers have been working to develop efficient algorithms and tools based on computational approaches to predict methylation states within a cell [16]. Methylation of the carbon-5 of cytosine (5mC) is one of the approch to find the tumor from DNA [15].

Authors in [18] have used an approach called LEAP based on supervised machine learning technique for genetic variant classification. A semi-supervised approach has been used in [17]. [11] have tested various machine learning classifiers for cancer detection based on DNA sequences. [19] have used deep learning for IDC breast cancer detection.

Based on extensive literature review, it has been found that genetic variant classification is a very important topic of research. However, the research is at infancy and a lot of studies are required to develop a benchmark solution for classification. Specifically, the authors are unable to found substantial work on using unsupervised or semi-supervised learning algorithms. So, the major contributions of this research are:

- A self organizing map based approach to genetic variant classification
- An auto-encoder based approach to genetic variant classification
- Empirical evaluation of these two approaches

III. PROPOSED METHODOLOGY

In this section, Autoencoder and SOM are discussed briefly. The dataset and steps of data pre-processing have been explained. Then there is a proposed architecture design with simulation.

A. AUTOENCODER

Autoencoder is an unsupervised deep artificial neural network. Autoencoder comprises of two main components; encoder and decoder. The encoder compresses and encodes information and then the decoder reconstructs the information back from the compressed encoded data. This data is as close to the original input as possible. Figure 2 illustrates traditional autoencoder architecture. Encoder and the decoder are the two part of autoencoder, which can be defined as ø and Ψ,

$$ø : X \rightarrow F \qquad (1)$$

$$\Psi : X \rightarrow F \qquad (2)$$

$$\theta, \Psi = \arg_{\theta,\Psi \min} "X - (\theta \mathrm{o} \Psi) X" \qquad (3)$$

B. SELF ORGANIZING MAPS

Self-organizing map (SOM) has been introduced by Teuvo Kohonen in the 1980s is sometimes called a Kohonen map. SOM is a type of artificial deep neural network (ANN). It was trained using unsupervised learning to create a low-dimensional representation of the training sample input data, called a map. SOM can be defined as the following mathematical model:

$$W_v (s + 1) = W_v (s) + \theta(u, v, s) \bullet a(s) \bullet (D(t) - W_v (s)) \qquad (4)$$

where s denotes the iteration of current, λ limit iteration, index of target is by t, target data vector of input is by D(t), node index is by v, node’s current weight vector by $W_{v,,}$ index of the matching unit known as best matching unit (BMU) by u, index of the BMU by e, angle due to the distance by θ(u, v, s), and due to the progress of the learning iteration denoted by a(s).

C. DATASETS

ClinVar [9-12] is a genetic variant annotation dataset source that is publicly available. The variants are classified by a geneticist in a genetic laboratory manually. The variants are categorized into different classes like benign, likely benign, likely pathogenic, and pathogenic, etc. The purpose is to forecast if a ClinVar variant will have conflicting class. This is a binary classification problem, where every instance in the dataset is a genetic variant. The dataset contains 40 columns as shown in Table I. CHROM, POS, REF, and ALT are some columns from the dataset. CHROM column contains Chromosome the variant is placed on. POS column contains the Position on the chromosome the variant is placed on. REF feature contains the reference Allele and the ALT column contains alternate Allele. The CLASS column contains the binary representation of the target class, where 0 describes no conflicting submissions and 1 describes conflicting submissions. The dataset contains a total of 65188 records.

D. EXPERIMENTAL ANALYSIS

16434 records belong to label 1 and 48754 records belong to label 0. This dataset is an imbalanced dataset. Some instances contain missing values. As can be seen in Fig. 2, the data of the CHROM column for class 1 and 0 is different, however, the data distribution is the same. Figure 1 shows the methodology employed for analysis. The data was first acquired which was preprocessed first. Then the model is created using SOM and auto-encoders. The experiments were performed and results are then acquired.

E. PREPROCESSING

Some columns were removed due to the sort of duplicate information reason. For instance, CLNHGVS columns contain the information that POS, REF and ALT columns already contain at the unit level. Most of the columns contain text-based information. This information was converted to enumeration via an encoding scheme as shown in Table I.

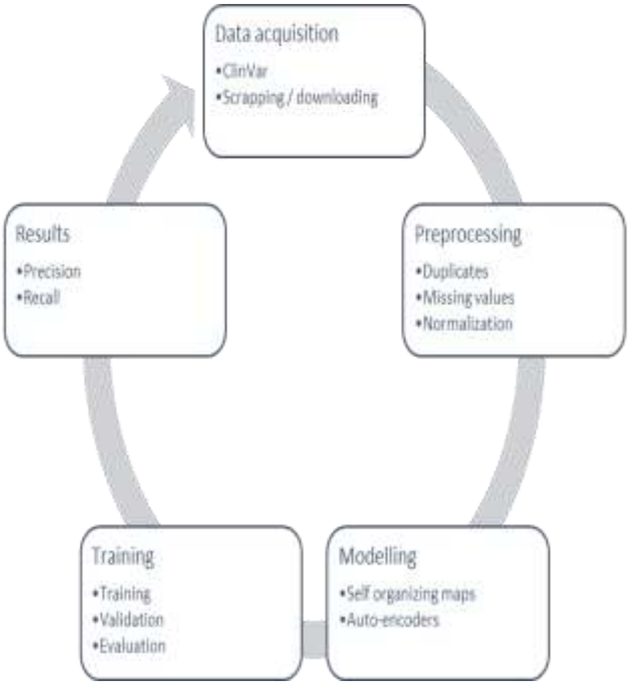


FIGURE 1: Methodology employed for analysis

Missing values in textual data type encoded as 0 (NONE). Some columns like AF ESP and AF EXAC contain decimal values. To fill missing values in such data types, we placed the value from the instances having a similar feature set. If any similar feature instance was not found, the 0.0 value was placed. The same process was followed for integer data type columns like cDNA position, CDS position, and Protein position columns.

F. MODELING AND SIMULATION

Two unsupervised deep learning algorithms were used in the proposed methodology. As mentioned, an autoencoder has two foremost components; encoders and decoders. Figure 2 shows the encoder-decoder architecture.

a. Encoder

Encoder architecture contains an input layer. The input dimension was 28, since 28 features were selected after the preprocessing step. After the input layer, dense layers were added with 50 neurons and tanh as an activation function. The dropout layer was added with a 0.2 drop out rate. Another dense layer was added with 25 neurons and relu activation function. Another dropout layer has been added with a 0.5 drop out ratio. Dense layer has been added with 12 neurons and tanh activation function. Finally, another dense layer has been added with 6 neurons and relu activation function.

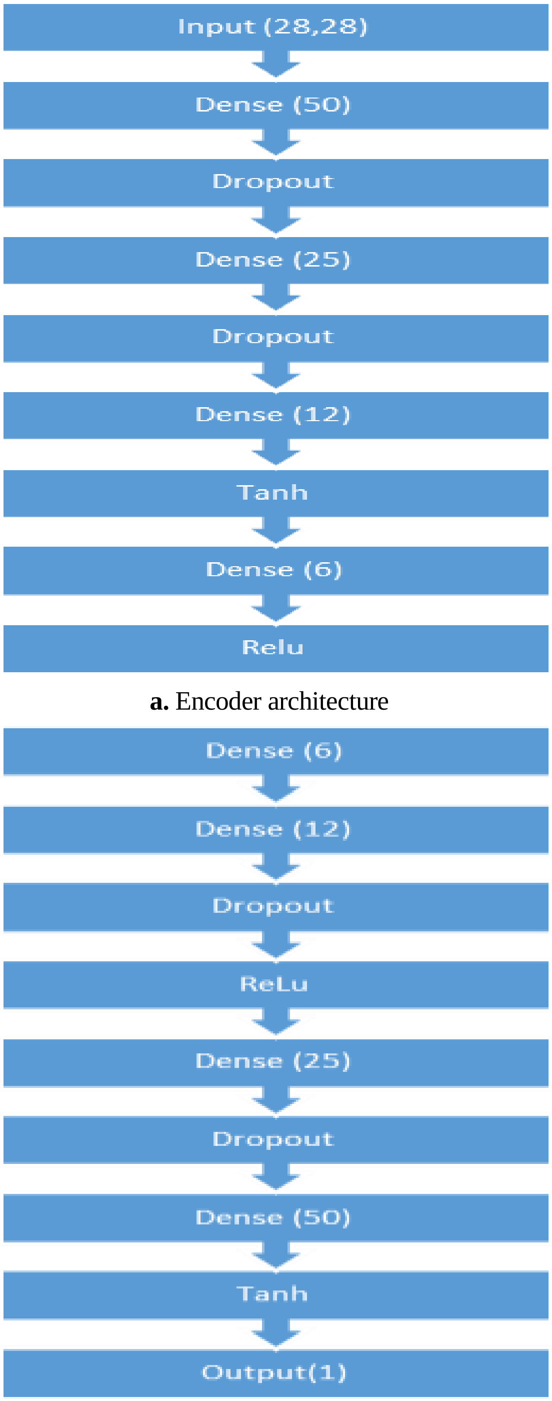
TABLE I

EXEMPLARY COLUMNS WITH THEIR RESPECTIVE DATA ENCODING

Fields	Encoding
CLNVC	Insertion - 1 Deletion - 2 Microsatellite -3 Inversion - 4 single nucleotide variant - 5 Duplication6 Insertion - 1 Deletion - 2 Microsatellite -3 Inversion- 4 single nucleotide variant - 5 Duplication 6 Indel - 7
IMPACT	Modifier - 1 High- 2 Low - 3 Moderate –4
FEATURE TYPE	MotifFeature - 1 Transcript - 2
BIOTYPE	Protein coding - 1 Misc RNA - 2
SIFT	Tolerated low confidence - 1 Tolerated - 2 Deleterious low confidence - 3 Deleterious - 4
POLYPHEN	Unknown - 1 Possibly damaging - 2 Benign - 3 Probably damaging - 4

b. Decoder

Decoder contains a dense layer with 6 neurons and relu activation function. A dense layer has been added with 12 neurons and tanh activation function. After two dense layers Dropout layer was added with a 0.5 rate. Dense layer has been added with 25 neurons and relu as an activation function. To reduce the overfitting, the dropout layer has been added with a 0.2 drop out ratio. Another Dense layer has been added with tanh activation function and 50 neurons. Finally, an output layer has been added with 1 neuron and sigmoid function due to binary classification problems. The proposed Auto Encoder architecture has been compiled with loss type *binary crossentropy* and *adadelta* optimizer. The number of epoch has been defined as 20 with batch size 500. The learning rate has been defined as 10-7. Testing and training loss/accuracy on each epoch has been defined in Table II.



a. Encoder architecture

b. Decoder architecture

FIGURE 2: Auto-encoder architecture

c. SOM

MiniSOM implementation in python has been used for SOM classification. X and Y dimensions of the SOM has been defined as 28. Input length has been defined as 28 due to the feature set. Sigma has been defined as 1.0. The learning rate has been adjusted as 0.005. Neighborhood function has been defined as Gaussian and the random seed has been defined as 5. Training iteration has been specified as 5000. The training process has been repeated with different parameter values and the best result has been achieved with the parameters' values described above.

IV. RESULTS AND DISCUSSIONS

The precision and recall curve for auto-encoder and SOM have been shown in Fig. 3 and 4. Similarly, Recall Vs precision have ben shown in Fig. 5 and 6. In Fig. 3 and Fig. 4 precision and recall have been plotted with different threshold values for Auto encoder and SOM respectively. It can be observed that precision and recall are the tradeoff in data science. At some point you have to determine a threshold. From both Fig. 3 and 4 it can be seen that 0.8 is the threshold value.

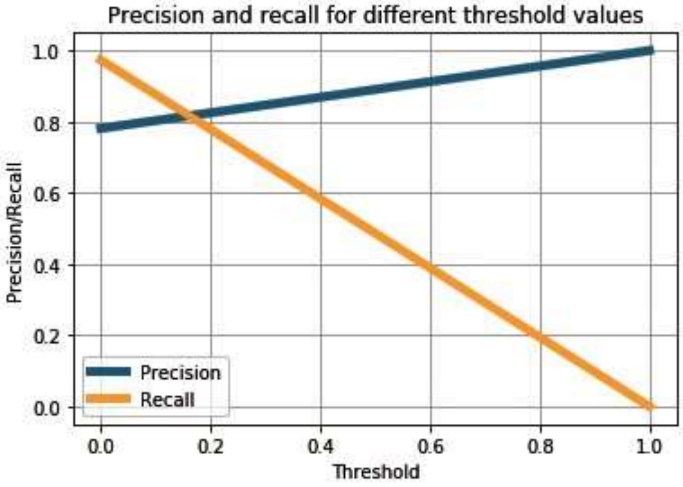


FIGURE 3: Precision Vs Recall for Autoencoder

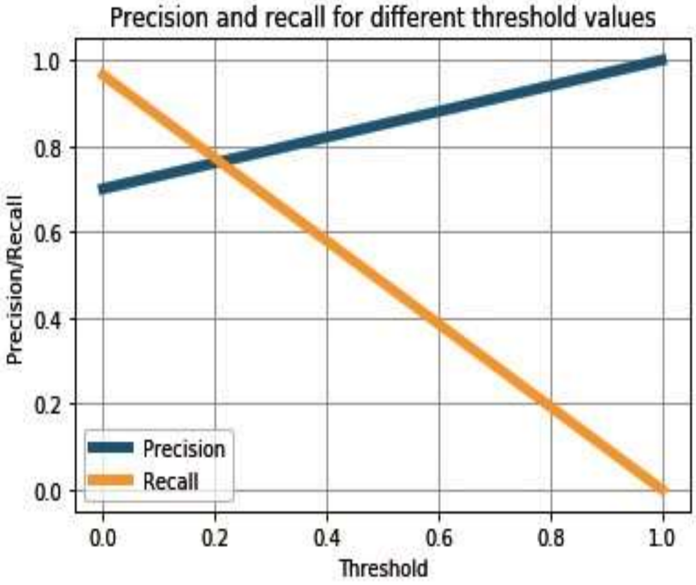


FIGURE 4: Precision Vs Recall for SOM

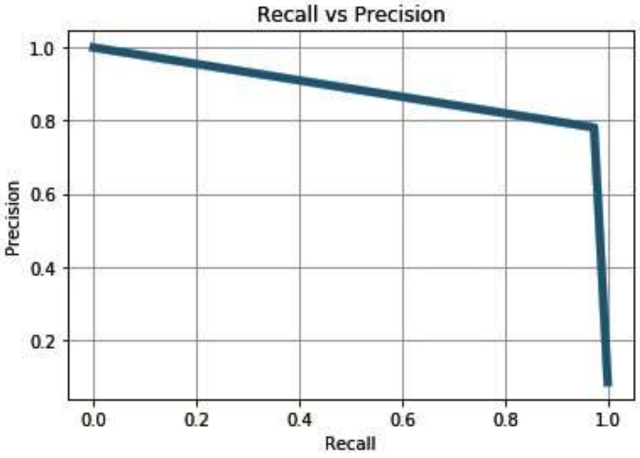


FIGURE 5: Recall vs Precision for Autoencoder

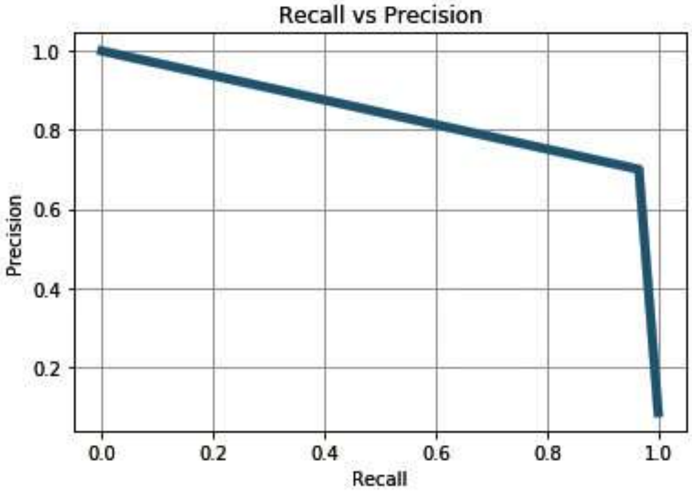


FIGURE 6: Recall vs Precision for SOM

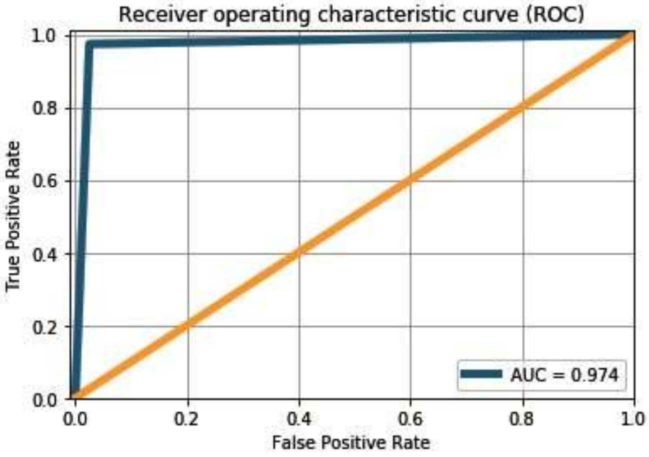


FIGURE. 7: ROC for Auto-encoder

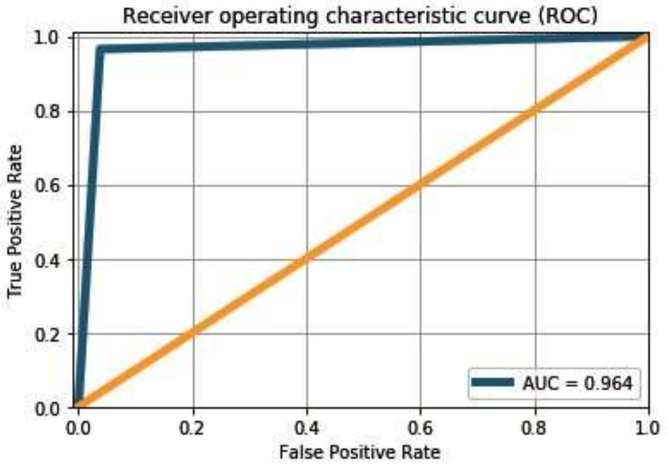


FIGURE. 8: ROC for SOM

It has been observed that state of the art results have been achieved during simulation of our proposed un- supervised deep learning techniques. 97% accuracy has been observed for Auto Encoder scheme and 96% accuracy has been achieved from MiniSOM simulation. Accuracy can be defined as the proportion of the values correctly classified. The comparison of accuracies can be visualized in Fig. 8. ROC curves for both techniques can be seen in Fig. 7 and 8. Figure 7 is graphical representation of ROC curve for Auto encoder. 0.947 AUC has been achieved. Figure 8 is ROC curve for SOM with 0.964 AUC. Recall with Precision visualization has been seen in Fig. 5 and Fig. 6 for both Auto encoder and SOM respectively. It can be observed that our both models performed very well. Auto encoder performed slightly better than SOM shown in Fig. 9.

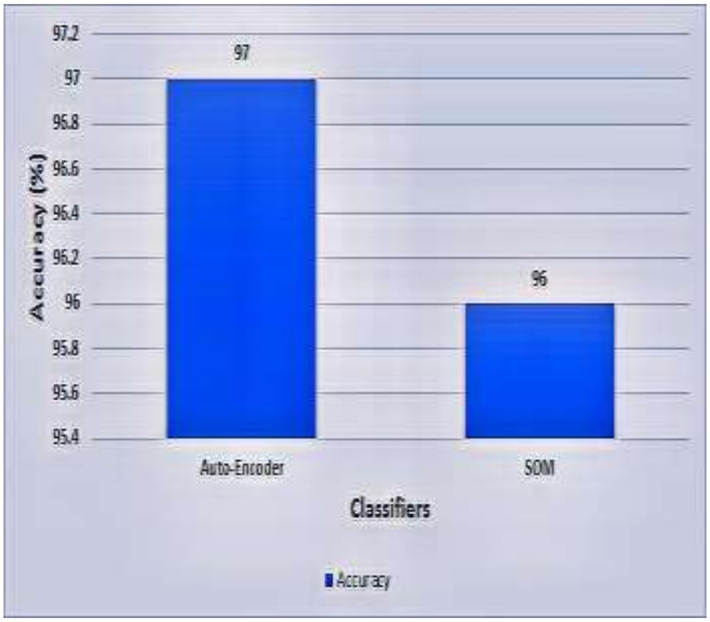


FIGURE. 9: Overall accuracy of AE and SOM

TABLE II

TESTING AND TRAINING LOSS AND ACCURACY ON EACH EPOCH

Epoch	Test Loss	Test Accuracy (%)	Validation Loss	Validation Accuracy (%)
1	0.6	57	7.97	50.01
2	0.58	59	5.26	55.1
3	0.58	60.1	4.35	58.9
4	0.44	65	3.03	61.3
5	0.49	63.2	3.59	59.3
6	0.5	62.4	3.13	60.25
7	0.42	66.2	2.45	65.4
8	0.48	63.8	2.15	70.12
9	0.5	62.5	2.05	74.2
10	0.4	72	1.89	78.2
11	0.38	73	1.58	81
12	0.3	78.2	1.23	87.4
13	0.25	87	1.45	85.6
14	0.26	86.1	1.02	90.21
15	0.29	84.3	1.1	90.35
16	0.27	85.96	0.95	95.4
17	0.2	90.2	0.95	96.5
18	0.2	92.5	0.96	95.89
19	0.15	95.6	0.94	97.52
20	0.14	97.53	0.92	98.52

Table III described the precision, recall and F score of each class for both classification techniques Auto encoder and SOM respectively. As per table both models performed extremely well for class 0 (Benign) but misclassified many instances of class 1 (Negative) that's why precision is lower than precision of class.

TABLE III

PRECISION, RECALL AND F1 SCORE OF EACH CLASS FOR BOTH AUTO ENCODER AND SOM TECHNIQUES.

SCHEME	CLASS	PRECISION (%)	RECALL (%)	F1 SCORE (%)
Auto-	0	100	98	99
	1	78	97	87
SOM	0	100	96	98
	1	70	97	81

Table IV is the tabular representation of confusion matrix of Auto-encoder. For class 0 (normal), Out of 34720, the 887 instances have been misclassified whereas for class 1 (Conflict), Out of 3186, the 101 instances have been misclassified. Confusion matrix for SOM can be visualized in table 4. The table described that for class 0, Algorithm with specified parameters misclassified 1358 instances out of 34249 and for class 1, 114 instances out of 3173 have been misclassified. It has been observed that unsupervised deep learning algorithms are performing extremely well for genetic variant classification although the dataset is imbalance. Similar results of around 98% AUC has been reported in [19]. Also, [18] has also shown similar results with an accuracy of around 97%.

TABLE IV

CONFUSION MATRIX FOR AUTOENCODER AND SOM

		Predicted Class		
	Classifiers	Classes	Normal	Conflict
Actual Class	**Autoencoder**	Normal	34720	887
		Conflict	101	3186
	SOM	Normal	34249	1358
		Conflict	114	3173

V. CONCLUSION

The paper proposed two unsupervised machine learning schemes for clinical classification of genetic variants. After extensive deliberation on related literature, it has been found that very few researches have employed unsupervised or semi-supervised approach to genetic variant classification. The paper have implemented self organizing maps and auto-encoder architecture for classification. The details are mentioned in section 3 and the results are reported in section 4. Following are the conclusions derived from results:

- Both schemes i.e. auto-encoders and self organizing maps have performed well, and state-of-the-art results have been achieved.
- The result showed better performance of autoencoder, where its accuracy is 97% (on test data).
- SOM has an accuracy of 96% (on test data)
- Similar results have been reported in 28, 29 with AUROC around 98%
- The proposed approach can save precious lives of many people via timely detection of genetic variants.

The purpose of this classification and framework is to improve the utilization of machine learning techniques to maximize the opportunity to study more about variants for the benefits of families and to reduce the risk of incorrect classification of variants in the clinical settings. In future studies, techniques based on deep learning such as convolutional neural networks, recurrent neural networks or other similar techniques can be tested. Also techniques based on ensemble of classifiers such as bagging and boosting can also be used.

REFERENCES

[1] Alipanahi, Babak, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. "Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning." *Nature biotechnology,* vol. 33, no. 8, pp. 831-838, 2015.

[2] Angermueller, Christof, Heather J. Lee, Wolf Reik, and Oliver Stegle. "DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning." *Genome biology,* vol. 18, no. 1, pp. 1-13, 2017.

[3] Carroll, Emma L., Mike W. Bruford, J. Andrew DeWoody, Gregoire Leroy, Alan Strand, Lisette Waits, and Jinliang Wang. "Genetic and genomic monitoring with minimally invasive sampling methods." *Evolutionary applications,* vol. 11, no. 7, pp. 1094-1119, 2018.

[4] Shameer, Khader, Kipp W. Johnson, Benjamin S. Glicksberg, Joel T. Dudley, and Partho P. Sengupta. "Machine learning in cardiovascular medicine: are we there yet?." *Heart,* vol. 104, no. 14, pp. 1156-1164, 2018.

[5] Xiong, Hui Y., Babak Alipanahi, Leo J. Lee, Hannes Bretschneider, Daniele Merico, Ryan KC Yuen, Yimin Hua et al. "The human splicing code reveals new insights into the genetic determinants of disease." *Science,* vol. 347, no. 6218, pp. 1254806, 2015.

[6] Yang, Yang, Katherine E. Niehaus, Timothy M. Walker, Zamin Iqbal, A. Sarah Walker, Daniel J. Wilson, Tim EA Peto et al. "Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data." *Bioinformatics,* vol. 34, no. 10, pp. 1666-1671, 2018.

[7] Zeng, Haoyang, Matthew D. Edwards, Ge Liu, and David K. Gifford. "Convolutional neural network architectures for predicting DNA–protein binding." *Bioinformatics,* vol. 32, no. 12, pp. i121-i127, 2016.

[8] Zhou, Jian, Chandra L. Theesfeld, Kevin Yao, Kathleen M. Chen, Aaron K. Wong, and Olga G. Troyanskaya. "Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk." *Nature genetics,* vol. 50, no. 8, pp. 1171-1179, 2018.

[9] Zhou, Jian, and Olga G. Troyanskaya. "Predicting effects of noncoding variants with deep learning–based sequence model." *Nature methods,* vol. 12, no. 10, pp. 931-934, 2015.

[10] Qiao, Peng, Di Zhang, Song Zeng, Yicun Wang, Biao Wang, and Xiaopeng Hu. "Using machine learning method to identify MYLK as a novel marker to predict biochemical recurrence in prostate cancer." *Biomarkers in Medicine,* vol. 15, no. 1, pp. 29-41, 2020.

[11] Arias, Pablo Millán, Fatemeh Alipour, K. Hill, and Lila Kari. "DeLUCS: deep learning for unsupervised classification of DNA sequences." *PLoS ONE,* 2021.

[12] Pham, Dinh-Minh, and Yu-Yen Ou. "An extensive examination of discovering 5-Methylcytosine Sites in Genome-Wide DNA Promoters using machine learning based approaches." *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 19, no. 1, pp. 87-94, 2021.

[13] Abid, A., Abid, A., Afshan, Z. and Anwar, S., "A Novel Machine Learning based Multiple-user Hand Recognition Approach," *Pakistan Journal of Engineering and Technology,* vol. 4, no. 1, pp.60-65, 2021.

[14] Mustafa, S., Khan, A., Hussain, S., Jhandir, M.Z., Kazmi, R. and Bajwa, I.S., "Automatic Speech Emotion Recognition using Mel Frequency Cepstrum Co-efficient and Machine Learning Technique," *Pakistan Journal of Engineering and Technology*, vol. 4, no. 1, pp.124-130, 2021.

[15] Nicora, Giovanna, Susanna Zucca, Ivan Limongelli, Riccardo Bellazzi, and Paolo Magni. "A machine learning approach based on ACMG/AMP guidelines for genomic variant classification and prioritization." *Scientific reports,* vol. 12, no. 1, pp. 1-12, 2022.

[16] Lai, C., Zimmer, A.D., O'Connor, R., Kim, S., Chan, R., van den Akker, J., Zhou, A.Y., Topper, S. and Mishne, G., "LEAP: Using machine learning to support variant classification in a clinical setting," *Human mutation*, vol. 41, no. 6, pp.1079-1090, 2020.

[17] Nicora, G., Marini, S., Limongelli, I., Rizzo, E., Montoli, S., Tricomi, F.F. and Bellazzi, R., "A semi-supervised learning approach for pan-cancer somatic genomic variant classification. In Conference on Artificial Intelligence in Medicine in Europe (pp. 42-46). Springer, Cham, 2019.

[18] Hussain, F., Saeed, U., Muhammad, G., Islam, N. and Sheikh, G.S., "Classifying cancer patients based on DNA sequences using machine learning," *Journal of Medical Imaging and Health Informatics*, vol. 9, no. 3, pp.436-443, 2019.

[19] Kumar, K., Saeed, U., Rai, A., Islam, N., Shaikh, G.M. and Qayoom, A., 2020. IDC Breast Cancer Detection Using Deep Learning Schemes. Advances in Data Science and Adaptive Analysis, vol. 12, no. 2, pp. 2041002, 2020.