Dimensionality Reduction using t-SNE Technique for Fundus Image Analysis in Diabetic Retinopathy

P.S Vijayalakshmi [1], M Jayakumar [2], Wilfred Blessing N.R [3]

[1] Assistant Professor, Department of Computer Applications
Dr. N.G.P Arts and Science College, Coimbatore-48, TN, India.

[2] Assistant Professor, Department of Computer Science,
Hindusthan College of Arts and Science, Coimbatore-28, TN, India.

[3] College of Computing and Information Sciences, University of Technology and Applied Sciences-Ibri, Sultanate of Oman.

Abstract: Dimensionality reduction states that the method of data transformation from a complex-dimensional space to one with fewer dimensions so that the implications drawn from the analysis of the reduced dataset are reasonably close to those drawn from the analysis of the original data set. A prime example is diabetic retinopathy (DR), which damages more than 50 percent of the total of all diabetes patients' eyes in some way. The symptoms of DR can range from blurry vision to total blindness, and most of the time, patients don't report any initial symptoms. In this study, we used the retinal Fundus Images study for the Diabetic Retinopathy Dataset dubbed CMCH to examine the unique method for dimension reduction with T-distributed method stochastic neighbor embedding (t-SNE).

***Keywords**: T-distributed method stochastic neighbor embedding (t-SNE), Diabetic Retinopathy (DR), Dimensionality reduction, High Dimensions*

I.INTRODUCTION

Adult blindness's most common complication is diabetic retinopathy (DR). Every year, more patients who suffer from DR-related vision loss [1] are diagnosed. Everyone is aware that DR becomes one of the most common consequences of diabetic mellitus (DM). Since there are no symptoms seen in the early days of DR and people hardly ever detect visual loss, treating DR is difficult [2]. The majority of them did not become aware that they had DR until the condition began to impairment vision, which happens in the last stages. As a result, they might not start therapy straight immediately.

For adults between the ages of 20 and 74, DR is the largest common problem of new incidences of visual impairment. Nearly all type 1 diabetes patients and more than 60% of type 2 diabetic patients develop retinopathy throughout the first 2 decades of the disease. Legal blindness was present in 3.6% of diabetes type -1 patients with a younger onset and 1.6% of type 2 diabetes mellitus patients with an older start in the Wisconsin Epidemiologic Survey of Diabetic Retinopathy (WESDR). 86% of blindness in the group with younger onset was caused by diabetic retinopathy. One-third of the instances of legal blindness in the older age group, when other eye disorders were prevalent, were brought on by diabetic retinopathy[3]. In this paper, we stated the related work of literature in section 2, section 3 shows the dataset selected, and section 4 shows Dimensionality reduction. Section 5 presents the Distributed Stochastic Neighbour Embedding, results and discussions are presented in the section 6, and finally, the conclusion in the section 7 finally.

II. RELATED WORK

Sundaram and Alli planned to employ Machine Learning (ML) as well as the Ensemble Classification approach to identify the properties of retina in DR detection using Machine Learning algorithms with Bagging Ensemble Classification (MLBEC). The applicable substances, which include "blood vessels, optic nerve, nerve cells, retinal nerve fiber rim, visual acuity size, thickness, and variance," were extracted from retinal pictures in the first stage of the suggested model. These objects were initially retrieved using ML's t-distributed method Stochastic Neighbor Embedding technique (t-SNE). Here, high-dimensional images were divided into comparable and dissimilar pairs using t-SNE to construct a probability distribution across them.[4]

Ever-expanding data sets with RNA overexpression for countless genes from as many as millions of cells are produced by single-cell transcriptomics. Dimensionality reduction is a common phase in data analysis pipelines that allows data to be seen in two dimensions. This step is often carried out to use t-distributed method of stochastic neighbor embedding (t-SNE). In high-dimensional data, it specializes in revealing local structures.[5]

To cut down on the overwhelming quantity of features, researchers have looked into methods that map an initial features field into a minimum feature space in the representation. Principal Component Analysis (PCA) was used as the dimension reduction technique by authors in [7] to propose a patient-dependent, semi-automated approach. The authors of [8] used Similar Spatial Patterns (CSP) to reduce the dimensionality of the feature space using autocorrelation to extract features. In [6], the authors retrieved 26 characteristics per channel and used PCA to whittle them down to 8. They used an LDA classifier to report the accuracy of seperate EEG channels. The distance links between data points at low dimensions cannot be properly preserved by PCA, even though it can minimize the more number of features. The number of individuals examined in [6] is insufficient to draw any firm conclusions [9].

III. DATASET SELECTED

For this method, we are using CMCH (Coimbatore Medical College Hospital, Coimbatore dataset. It contains 3670 images in rural and urban south India. The images in the dataset have 5 classes of Diabetic Retinopathy images. They are normal-class, mild -class, moderate- class2, severe-class and Proliferative Diabetic Retinopathy-class. We applied the t-SNE dimensionality reduction for these images. An improved grey wolf algorithm and convolutional neural networks are utilized for the diagnosis of Diabetic retinopathy [10]. Here region-based segmentation techniques are applied for the lesions segmentation. The modified ant colony optimization algorithm with convolutional neural network employed to give the better results in this model [11]. Various datasets like Kaggle and APTOS are used for that research.

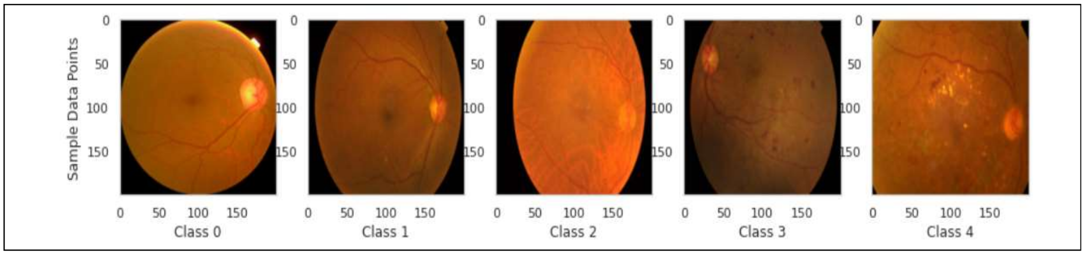

Figure 1: Sample Dataset

3.1 Grayscale Conversion

Grayscale images were created from RGB images for improved image improvement and viewing. It eliminates the challenges brought on by computational requirements and makes algorithm simplification easier. It makes learning easier for those who are not experienced with image processing. This is so because an image is reduced to its simplest pixel via grayscale compression. It makes straightforward visualizing easier.

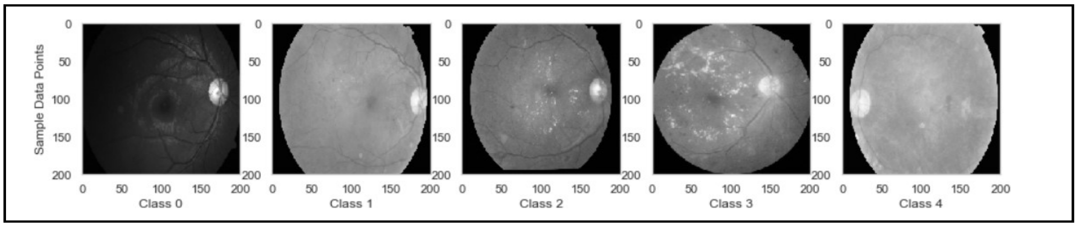

Figure 2: Grayscale images

IV.DIMENSIONALITY REDUCTION

In Machine Learning, having a huge number of characteristics introduces a lot of issues, the two most significant ones being that it makes training exceedingly sluggish and makes it challenging to come up with the best solution. This is regarded as the "curse of dimensionality," to put it simply, this dimensionality reduction is the method of narrowing the set of features to those that are more important. Data Compression, Noise Reduction, Data Classification, and Dimensionality Reduction are the most common applications. Data Visualization is one of the most important components of dimensionality reduction.

V. T-DISTRIBUTED STOCHASTIC NEIGHBOUR EMBEDDING

The dimensionality reduction method known as t-distributed method stochastic neighbor embedding, or t-SNE, developed in 2008 by Laurens van der Maaten with Geoffrey Hinton is especially well applied for the display of large datasets. With this utilization of t-SNE, a high-dimensional data collection can be transformed into less dimensional graphs while retaining the majority of the original data. It accomplishes this by assigning a location on a two- or three-dimensional map to each data point. This method locates clusters in the data, ensuring that an embedding maintains the significance of the data. While attempting to keep similar cases close together and different instances apart, t-SNE lowers dimensionality. [12]

A method for dimensionality reduction that works particularly well for the presentation of high dimensional data is the t-Distributed method of Stochastic Neighbor Embedding (t-SNE). By reducing the dimensions of the data, t-distributed Stochastic Neighbor Embedding (t-SNE) is mostly used for data visualization [13] A mathematical tool for data visualization is called t-SNE.

Stochastic Neighbor Embedding (SNE), which is the initial step in the t-SNE process, is used to transform the dataset's high-dimensional Euclidean distances towards conditional probabilities that represent similarities for each pair of data [13]. This conditional probability p $_{a|b}$, denoted in the equation below [12], represents how similar the data xa and xb are,

$$p_{a|b} = \frac{exp\frac{-\|x_b - x_a\|^2}{2\sigma^2}}{\sum_{a \neq k} \frac{-\|x_k - x_a\|^2}{2\sigma^2}}$$

______________________________(1)

Equation (1) calculates the distance between two data points x_a and x_b while taking into account a Gaussian distribution surrounding x_b with a specified variance of 2. Every data has a different variance, which is selected such that information in dense areas have a lower variance than data in sparse areas [16]. Then, to generate the two pairs of the probabilities (Q $_{a|b}$) inside the low-dimension space, a "Student t-distribution" with one level of freedom, comparable to the Cauchy distribution, is utilized [16]. The similarities among P a|b and Q $_{a|b}$ gets equal if indeed the high-dimension variables x_a and x_b are appropriately mapped to the low-dimension variables y_a and y_b. Therefore, from low dimension towards high dimension spaces, t-SNE reduces the difference between the two probabilities. The cost function () of the summing of Kullback-Leibler divergence is optimized to determine this difference, as given below [16]:

$$\phi = \sum_a \sum_b P_{a|b} log \frac{P_{a|b}}{Q_{a|b}}$$

______________________________(2)

Algorithm 1: t-SNE

```
Input: $X \in R^{n \times d}$
Output: $Y \in R^{n \times k}$
1: Apply SNE to X to calculate the conditional probabilities Pa|b and Qa|b
2: Map X to Y by minimizing the difference between Pa|b, and Qa|b based on the cost function $\phi$
```

5.1 Hyperparameter Tuning

The following 2 parameters in this technique can have a significant impact on the outcomes: a) The algorithm's iterative cycle count

b) perplexity This can be compared to the number of nearby points. t-SNE must think about.

VI. RESULTS AND DISCUSSION

For the CMCH images on different perplexities like 2,5,10,15,20,30,40,50, the unique t-SNE approach is used. The "perplexity" feature of t-SNE specifies, in general terms, how to manage attentiveness among local as well as global elements of the data. In a way, the parameters are an educated guess as to how many near neighbors each point has. The resultant images are complicated by the perplexity value. The concept of perplexity provided by Van der Maaten and Hinton[17] can be seen as a smooth measurement of the actual number of neighbors. Typical values range from 5 to 50, and the effectiveness of t-SNE method is pretty resistant to variations in the perplexity.

The diagrams do depict these clusters with perplexity numbers in the (5–50) range proposed the van der Maaten and also Hinton, albeit in quite different shapes. Things start to become strange outside of that range. Perplexity 2 is dominated by regional variances. The merged cluster image with perplexity 100 highlights a potential trap: for the process to work properly, the perplexity is to be below the number of points. Implementations sometimes result in unanticipated behavior.

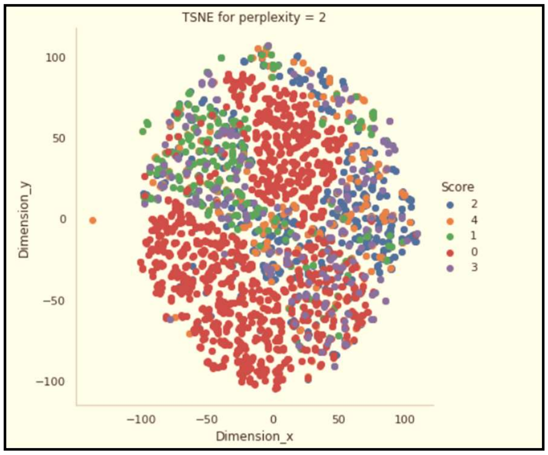

Figure 3: TNSE for perplexity=2

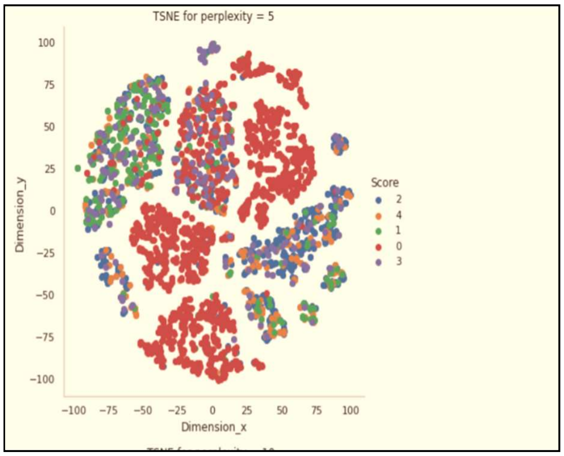

Figure 4: TNSE for perplexity=5

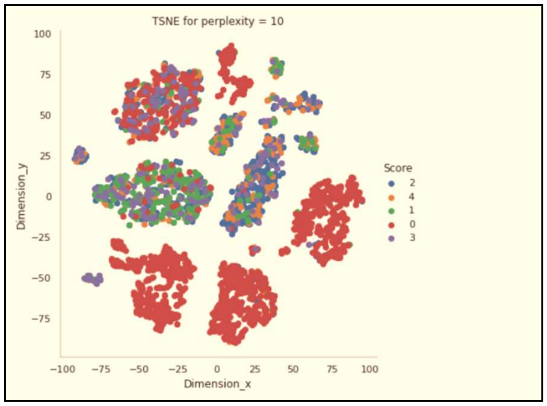

Figure 5: TNSE for perplexity=10

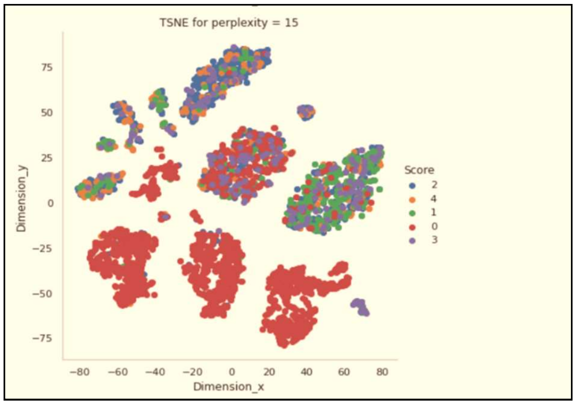

Figure 6: TNSE for perplexity=15

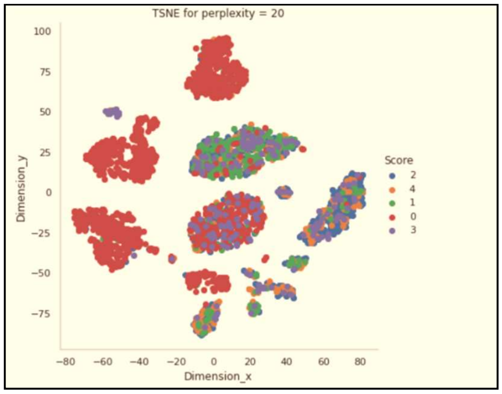

Figure 7: TNSE for perplexity=20

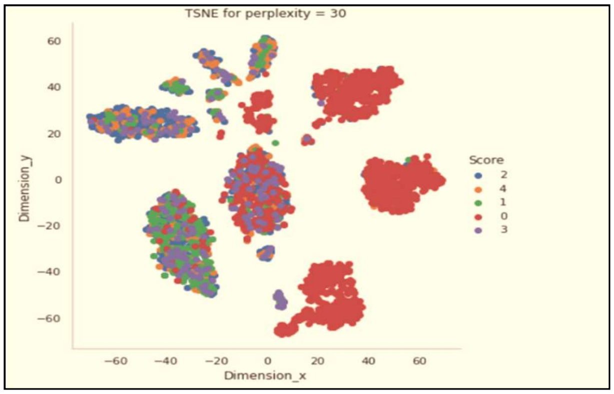

Figure 8: TNSE for perplexity=30

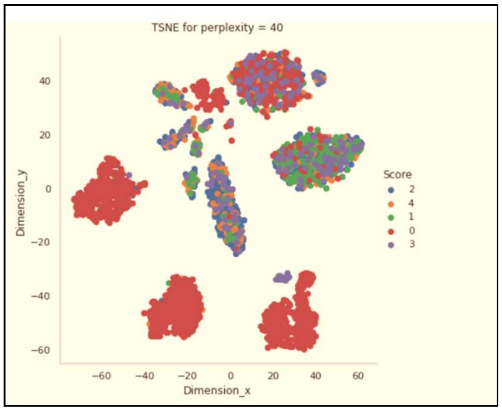

Figure 9: TNSE for perplexity=40

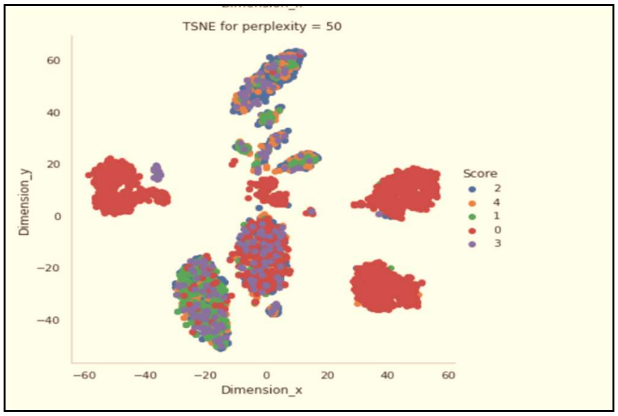

Figure 10: TNSE for perplexity=50

VII.CONCLUSION

The challenge of high-dimensional data visualization, which involves dealing with data with a large range of dimensionality, is significant in many different fields. In this study, we present a method for transforming a high-dimensional dataset together into a matrix of pairwise similarities. We also present a novel visualization method for the similarity data that we dub "t-SNE." The local of structure of both the high dimensional data can be captured by t-SNE to a large extent, and it can also disclose global of structure, such as the existence of clusters at various sizes. The memory and computational complexity of t-SNE is exponential in the number of datapoints, like many other visualization approaches. Because of this, it is impossible to use the usual t-SNE method on data sets that have considerably more points than, say, 10,000. Of course, utilizing t-SNE to display a random selection of the data points is doable. In this study, we expanded the scope of retinal fundus image analysis for Diabetic Retinopathy datasets by including t-SNE dimensionality reduction. Additionally, efforts are undertaken to calculate the performance of the t-SNE utilizing various perplexities. In multivariate data, important structures and unexpected correlations are intended to be revealed through dimensionality reduction techniques. The t-SNE outperformed other methods for reducing the dimensionality of fundus images, according to the results. To put it briefly, t-SNE is a form of machine learning that focuses on preserving the arrangement of neighbor points but consistently yielding marginally different conclusions on the same set of data. Additionally, we discovered that a non-linear dimensionality reduction strategy is better suited to this set of data than a linear one.

References

[1] D. K. Prasad, L. Vibha, and K. Venugopal, " Early detection of diabetic retinopathy from digital retinal fundus images," 2015 IEEE RecentAdvances in Intelligent Computational Systems (RAICS), 2015.

[2] S. Masood, T. Luthra, H. Sundriyal, and M. Ahmed, "Identification of diabetic retinopathy in eye images using transfer learning," 2017International Conference on Computing, Communication and Automation(ICCCA), 2017.

[3] https://diabetesjournals.org/care/article/27/suppl_1/s84/24669/Retinopathy-in-Diabetes

[4].Somasundaram SK, Alli P. A machine learning ensemble classifier for early prediction of diabetic retinopathy. J Med Syst. 2017;41:201

[5] Kobak, D., Berens, P. The art of using t-SNE for single-cell transcriptomics. Nat Commun 10, 5416 (2019). https://doi.org/10.1038/s41467-019-13056-x

[6] M. Yıldız, E. Bergil, The investigation of channel selection effects on the epileptic analysis of EEG signals.

[7] O. Smart, M. Chen, Semi-automated patient-specific scalp eeg seizure detection with un395 supervised machine learning, Computational Intelligence in Bioinformatics, and Computational Biology (CIBCB), 2015 IEEE Conference on, IEEE, 2015, pp. 1–7.

[8] S. Khanmohammadi, C.-A. Chou, A simple distance-based seizure onset detection algorithm using

common spatial patterns, in International Conference on Brain and Health Informatics, Springer, 2016, pp. 233–242.

[9] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) 215–220

[10] Vijayalakshmi, P. S. "An Efficient Deep Intelligent Based MACO-CNN Algorithm for Classification of Diabetic Retinopathy Disease from Retinal Fundus Images." Design Engineering (2021): 5182-5205.

[11] Vijayalakshmi, P. S., & Kumar, M. J. (2022). An improved grey wolf optimization algorithm (iGWO) for the detection of diabetic retinopathy using convnets and region based segmentation techniques. International Journal of Health Sciences, 6(S1), 13100–13118. https://doi.org/10.53730/ijhs.v6nS1.8330

[12] van der Maaten, L.J.P. t-Distributed Stochastic Neighbor Embedding

[13] L.V.D.Maaten, and. Hinton, "VisualizingDatausingt-SNELaurens, "J.Mach. Learn. Res., vol. 9, pp. 2579–2605, 2008.

[14] L.v.d. Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (1) (2008) 2579–2605.

[15] B.M. Devassy, S. George, Dimensionality reduction and visualization of hyperspectral ink data using t-SNE, Forensic Sci. Int. 1 (1) (2020) 1–9.

[16] F.H.M. Oliveira, A.R. Machado, A.O. Andrade, On the use of t-distributed stochastic neighbor embedding for data visualization and classification of individuals with Parkinson’s disease, Comput. Math. Methods Med. 1 (1) (2018) 1–18.

[17] L. Maaten, G. Hinton. Visualizing Data using t-SNE, 2008 http://www.jmlr.org/papers/volume9/ vandermaaten08a/vandermaaten08a.pdf