CHAPTER 4

VORONAI-CLUSTERING SPARSE AUTO ENCODER DEEP LEARNING ALGORITHM

4.1 VORONAI-CLUSTERING

A popular tool in machine learning, clustering has attracted a lot of interest from researchers. The ability of Voronoi diagrams to naturally divide space into sub-regions, which helps with data clustering, is one of their main advantages. This method uses the largest empty Voronoi circles to locate neighboring points, which are Voronoi vertices called cluster prototypes. By repeatedly creating new Voronoi diagrams, these prototypes a in the efficient fusion of points and eventually form the desired clusters.

4.2 PROPOSED METHODOLOGY

The Spatial and temporal features are key elements to predict the air quality system. They are determined by location and features of site. Equation (4.1) provides the latitude and longitude coordinates at point lti, which are represented by pi and qi, respectively.

$$Lt\text{i} = (Lt\text{i}, P\text{i}, \text{q}I) \qquad \in \text{l},\text{P},\text{q} \tag{4.1}$$

where Lt_i is the denoted as latitude and longitude of location site co-ordinate points. Where p_i and q_i are the latitude and longitude of site.

4.2.1 Voronoi Clustering

Voronoi clustering, also known as Voronoi tessellation or Voronoi diagram, is a geometric method used for partitioning a space into regions based on proximity to a specified set of points. In this method, each point in the space is associated with a region consisting of all the points that are closer to it than to any other specified point.

1. Initialization: Begin with a set of points called seeds or generators. These points are typically randomly chosen or selected based on certain criteria.
2. The Voronoi Diagram Building: Create a Voronoi diagram centered on these starting points. Every seed point establishes a region in the space that is made up of all points that are closer to it than to any other seed point. This area is referred to as a Voronoi region or cell.

3. Clustering: Assign each data point in the space to the Voronoi cell corresponding to the nearest seed point. This effectively partitions the space into clusters, with each cluster represented by a Voronoi cell.

A cell is the region that is connected to a Voronoi point cj. Since each point in a cell is closer to its centroid than it is to another centroid, the boundaries of the cells precisely locate in the middle of the two centroids. Figure 4.1 depict the six independent cluster in the Voronai diagram.

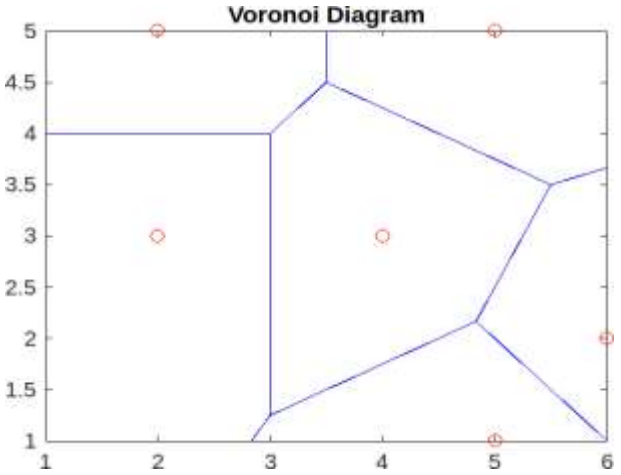

Figure 4.1 Voronoi Diagram for Six Clusters

Algorithm for Voronoi Clustering:

function (S, max)

Create the Voronoi diagram for $L = \{l_1, l_2, \ldots l_n\}$;
Estimate the Voronoi cell volumes and rank the instances. Consider the resultant rank $l_1, l_2, \ldots l_n$;
$(i = 1{:}n)$
If R_i *volume related to* s_i *is maximum*
Merge R_i with an adjacent cluster with the smallest class number; If one exists, or else allocate a new class number to the cell;

else
Allocate R_i to the nearest adjacent cluster;
end if
end for
end function

4.2.2 Improved Artificial Neural Network

The following techniques are used to train an air pollution dataset for the Improved Artificial Neural Network (IANN):

Perceptron:

Multi-Layer Perceptron, or MLPs, are Artificial Neural Networks (ANNs) composed of multiple interconnected layers of nodes, or neurons. The layers consist of an output layer, an input layer, and one or more hidden layers. Information can move forward through the network when each neuron in a layer of an MLP receives input from all the neurons in the layer before it and transmits its output to all the neurons in the layer after it. During the training phase, a weight is assigned to each neuronal connection in order to maximize the network's performance on a specific task.

Bayesian normalized:

Bayesian normalized doesn't seem to be a specific term or concept in the context provided. It's possible that you're referring to Bayesian normalization, which could involve using Bayesian methods for normalization in data analysis or machine learning tasks .In Bayesian statistics, normalization often refers to calculating posterior probabilities or likelihoods such that the sum of probabilities across all possible outcomes equals one. This ensures that the probabilities represent a valid probability distribution.

Scaled Conjugate Gradient:

Scaled Conjugate Gradient (SCG) is an optimization algorithm commonly used for training artificial neural networks, particularly in the context of supervised learning. It is an iterative optimization technique that aims to minimize a cost function, typically associated with training a neural network. Unlike traditional gradient descent methods, SCG dynamically adjusts the step size for each parameter update based on the curvature of the cost function. This adaptive step size helps accelerate convergence, especially in cases where the cost function has different curvatures along different dimensions. SCG is known for its efficiency in training neural networks, as it often requires fewer iterations compared to other optimization algorithms like gradient descent or conjugate gradient without scaling. This efficiency makes it particularly useful for training deep neural networks or networks with large numbers of parameters.

Overall, Scaled Conjugate Gradient is a powerful optimization algorithm that can significantly speed up the training process for neural networks, leading to faster convergence and improved performance on various tasks.

Algorithm of IANN Model: Consider a as the training rate and weight initialization as the procedures used in the proposed IANN model.

- Perform the forward stage
- Perform the backward stage
- Combine the individual gradients
- Merge the separate gradients for all I/O pairs to obtain the overall gradient
- Modify the weights: Based on a and utilizing, the weights are updated.

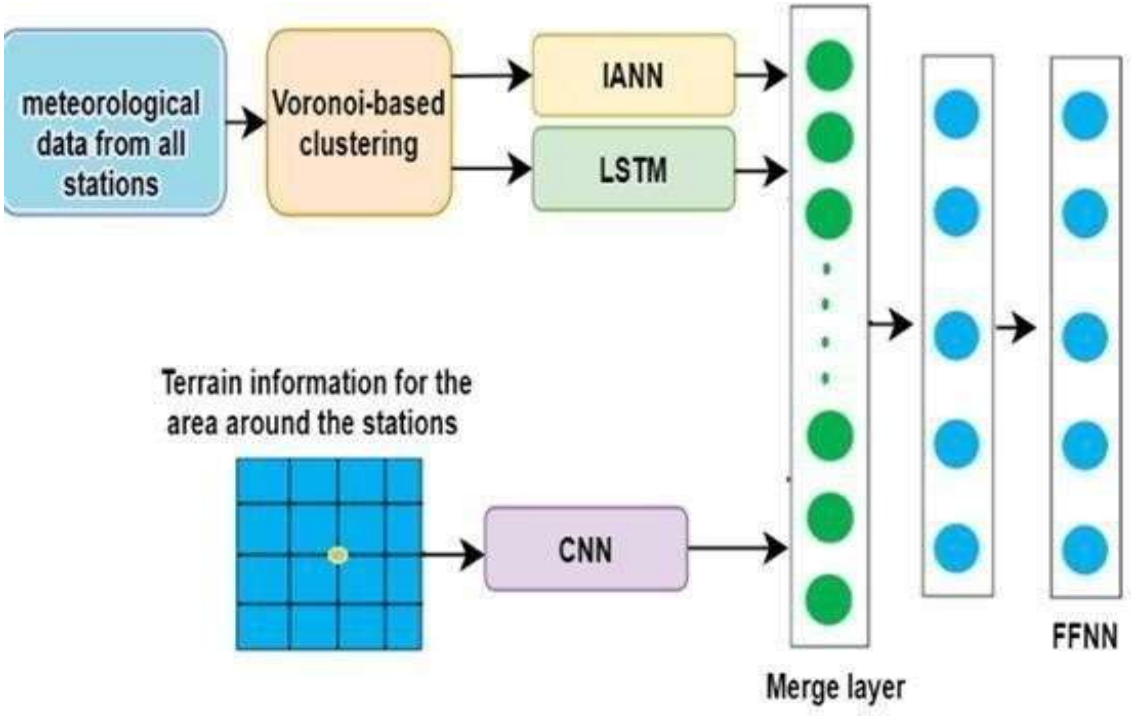

Figure 4.2 Architecture of VCSAE-DL

The integration of Voronoi-based clustering with a sparse autoencoder offers a robust framework for analyzing high-dimensional data. Sparse autoencoders are neural networks designed to learn compressed data representations by minimizing reconstruction loss while enforcing sparsity constraints. The computational complexity of training a sparse autoencoder is influenced by the number of data points (m), input dimensionality (d), hidden layer size (h), and the number of training epochs (e). Specifically, the complexity scales as $O(e \cdot m \cdot (d \cdot h + h^2))$, making the use of mini-batch training crucial for managing large datasets. The construction of Voronoi diagrams incurs significant computational costs, especially with increasing cluster numbers (n) or latent space dimensions (d′). Voronoi clustering in the latent space incurs $O(n^{\lceil d'/2 \rceil})$ for diagram construction and $O(m \cdot n)$ for point assignments.

The scalability of the Voronoi-based clustering sparse autoencoder model for air quality prediction is enhanced by dimensionality reduction through the sparse autoencoder, which reduces computational complexity by lowering the input data dimensions. This enables efficient clustering in the latent space. The model scales well with moderate data sizes, and techniques like mini-batch training and parallelization (e.g., using GPUs) allow it to handle larger datasets. However, as the number of clusters increases or the dataset grows significantly, approximations for Voronoi clustering and optimized nearest-neighbor searches are needed to maintain efficiency. The model is well-suited for real-time applications by balancing predictive accuracy and computational efficiency.

4.3 EXPERIMENTAL RESULTS

The experimental outcomes of the proposed VCSAE-DL system are compared with techniques such as ST-SVR and ISAE-DL

4.3.1 Accuracy

The accuracy achieved by the proposed method VCSAE-DL is compared with the two methods, ST-SVR (Spatial Temporal -Support Vector Regression) and ISAE-DL for different iterations and shown in Figure 4.3.

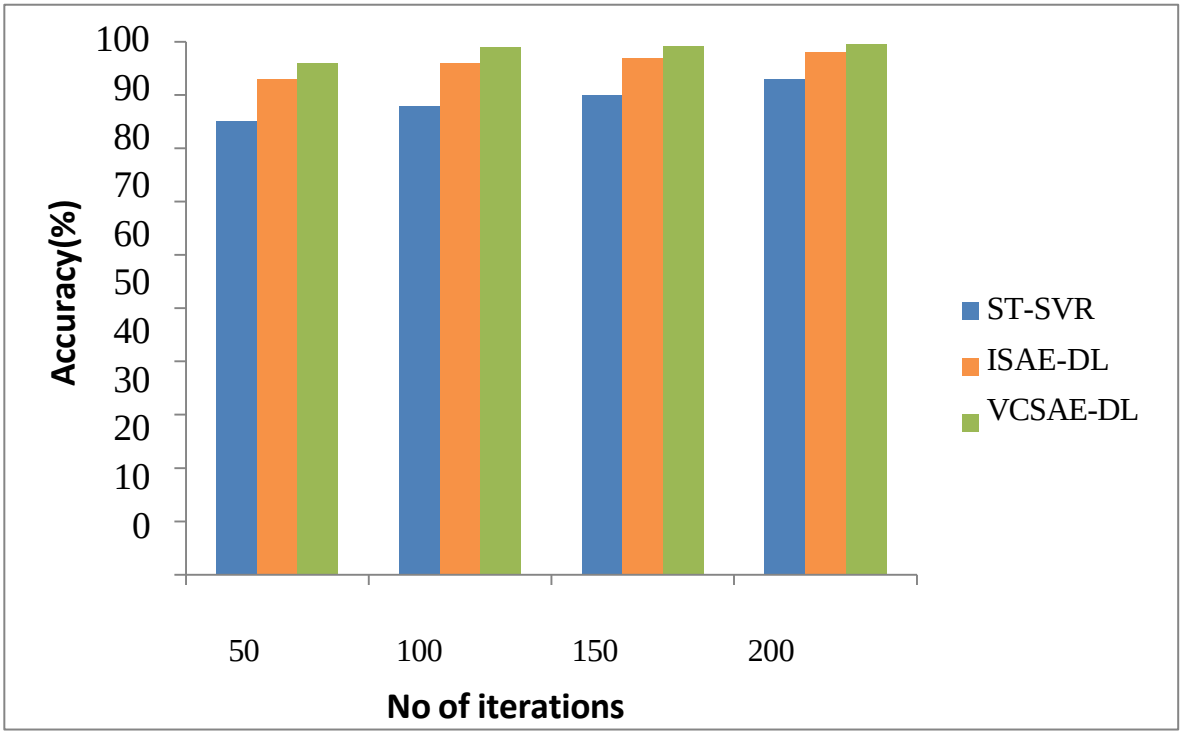

Figure 4.3 Comparison of accuracy

The findings reveal that the accuracy is ST-SVR is 93%, ISAE-DL is 98%, and the proposed method VCSAE-DL achieves an accuracy of 98.5%. This represents 0.5% enhancement in the accuracy. for VCSAE-DL.ST-SVR and ISAE-DL exhibit lower accuracy values, while VCSAE-DL demonstrates higher accuracy values.

4.3.2 Precision

Figure 4.4 presents a comparison of precision of the proposed method VCSAE-DL with ST-SVR and ISAE-DL for different iterations.

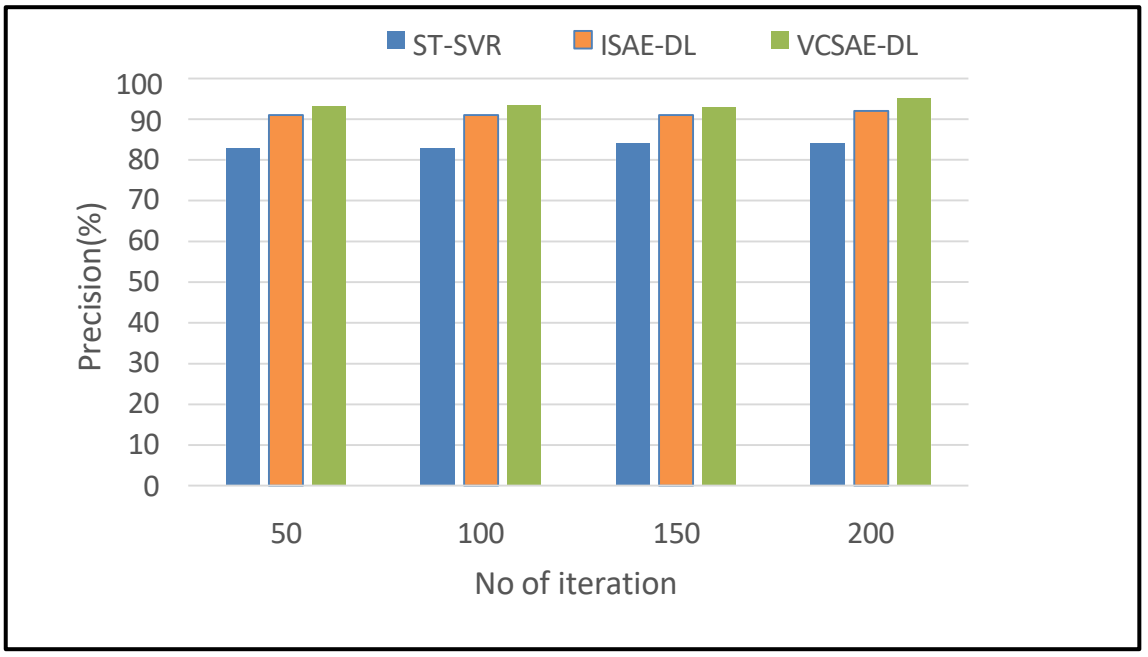

Figure 4.4 Comparison of precision

The results indicate that the Precision for ST-SVR is 84.2%, ISAE-DL is 92%, and VCSAE-DL achieves 95.2%. This signifies an improvement of 0.8% and 0.3% in the performance of VCSAE-DL with respect to precision value. These results indicate that VCSAE-DL outperforms ST-SVR and ISAE-DL.

4.3.3 Specificity

The Specificity of the proposed method VCSAE-DL for different iterations is compared with ST-SVR and ISAE-DL and is given in Figure 4.5. The experimental results indicate that the specificity for ST-SVR is 80%, ISAE-DL is 89%, and VCSAE-DL achieves 94%. This indicate that there is an improvement in the performance of VCSAE- DL by 0.9% and 0.5% compared to ST-SVR and ISAE-DL. These findings demonstrate that VCSAE-DL outperforms ST-SVR and ISAE-DL methods on comparing the precision value.

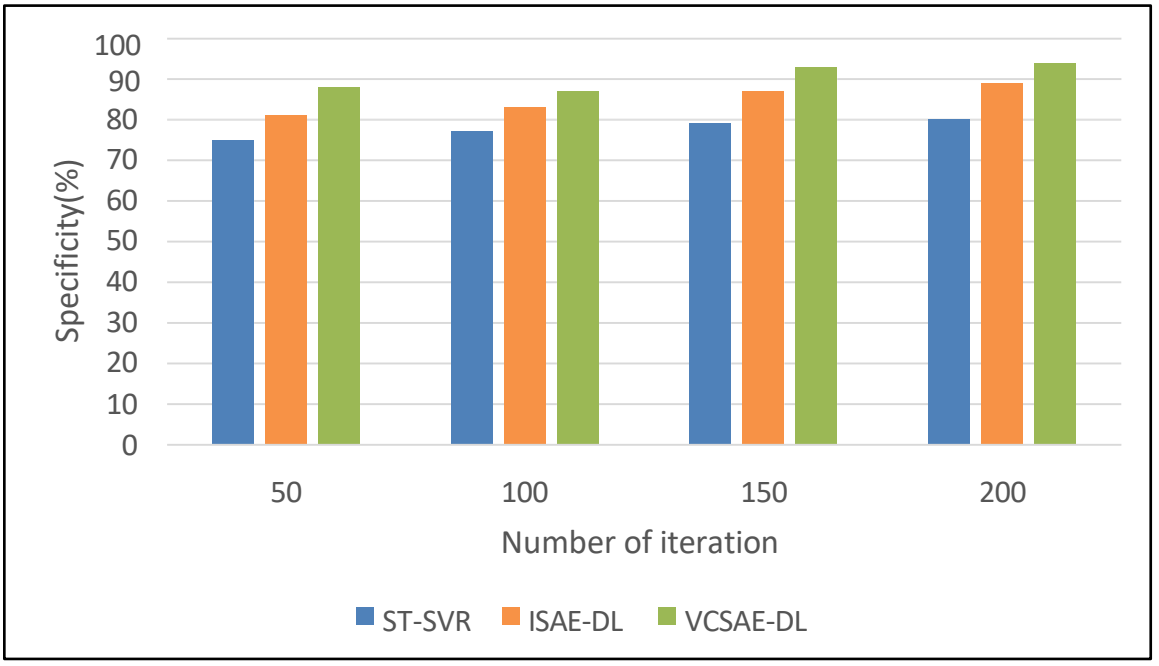

Figure 4.5 Comparison of Specificity

4.3.3 Sensitivity

The Sensitivity value obtained by the proposed method VCSAE-DL is compared with ST-SVR and ISAE-DL and is shown in Figure 4.6.

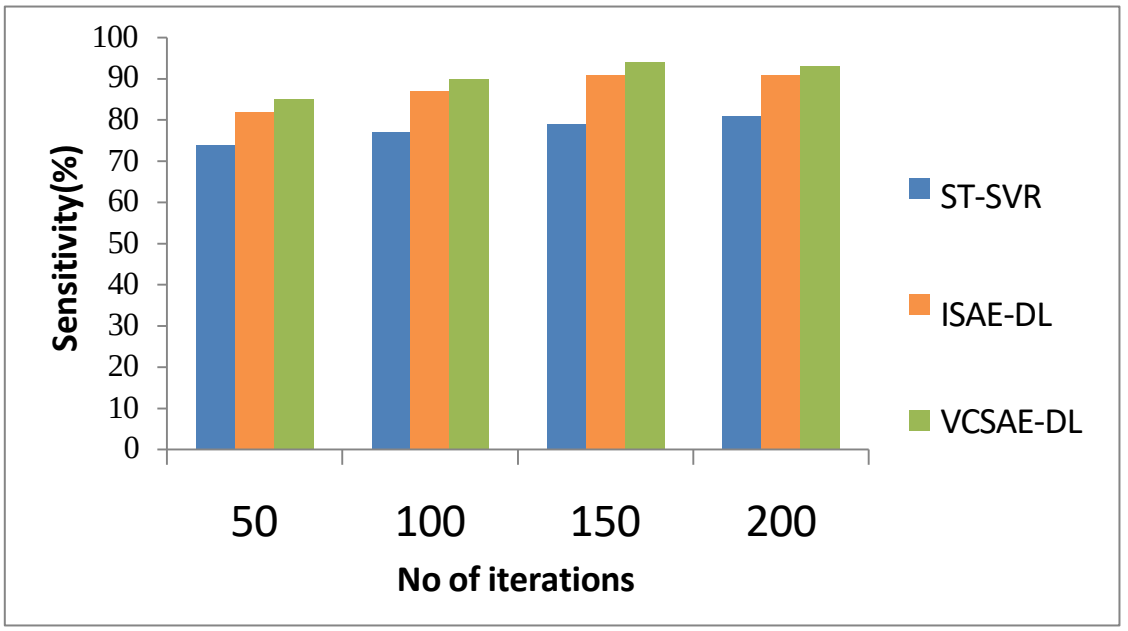

Figure 4.6 Comparison of Sensitivity

The results indicate that the sensitivity of ST-SVR is 81% and ISAE-DL is 91.5%, whereas the proposed method VCSAE-DL achieves 93%. This signifies an improvement of 0.2% and 0.12% in the performance of VCSAE-DL. These findings highlight that VCSAE-DL outperforms other current methods. Specifically, compared to ST-SVR and ISAE-DL, which demonstrate low sensitivity values, VCSAE-DL showcases high sensitivity values.

4.3.4 Area under Curve (AUC)

Figure 4.7 compares the Area under the Curve (AUC) of the proposed method VCSAE-DL with ST-SVR and ISAE-DL for different iterations.

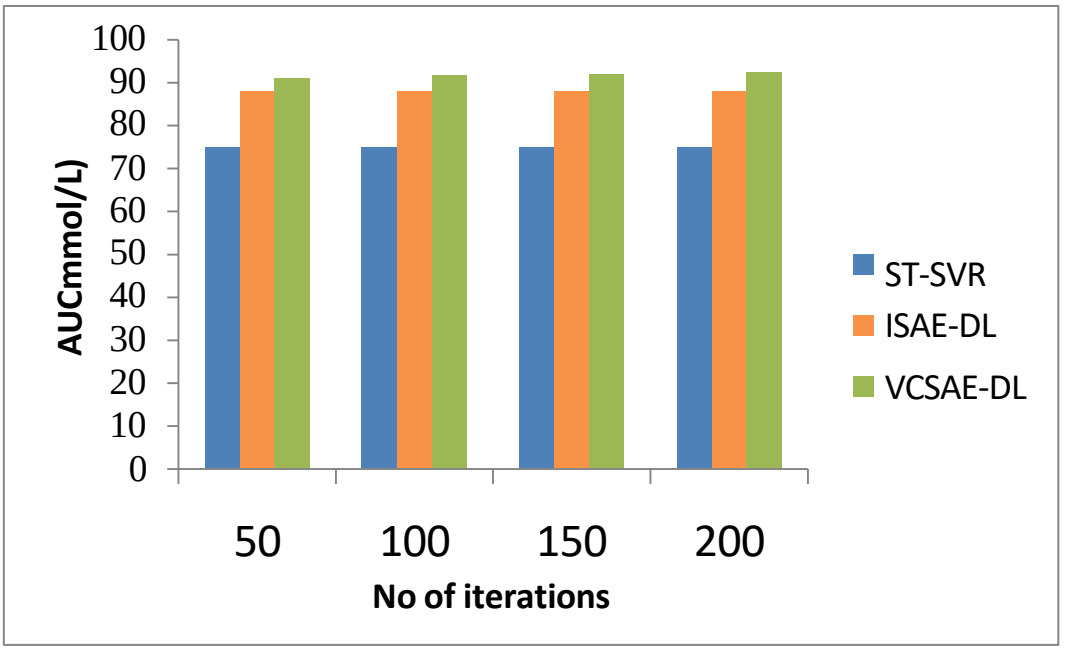

Figure 4.7 Comparison of AUC

The results given in the Figure 4.7 reveal that the AUC of VCSAE-DL has improved by 5.6% and 19.4% compared to ST-SVR and ISAE-DL in the prediction of air quality system. These results indicate that the proposed method VCSAE-DL outperforms other current methods as the existing models struggle to differentiate between the positive and negative classes. VCSAE-DL indicates perfect discrimination, meaning the model perfectly distinguishes between positive and negative classes.

4.3.5 Matthew's Correlation Coefficient (MCC)

Figure 4.8 compares the Mathew's Correlation Coefficient (MCC) of the proposed method VCSAE-DL with ST-SVR and ISAE-DL for different iterations. The experimental results indicate that MCC of VCSAE-DL has improved by 0.17% and 0.4% in the air quality prediction system. These findings demonstrate that VCSAE-DL exhibits higher MCC values compared to ST-SVR and ISAE-DL.

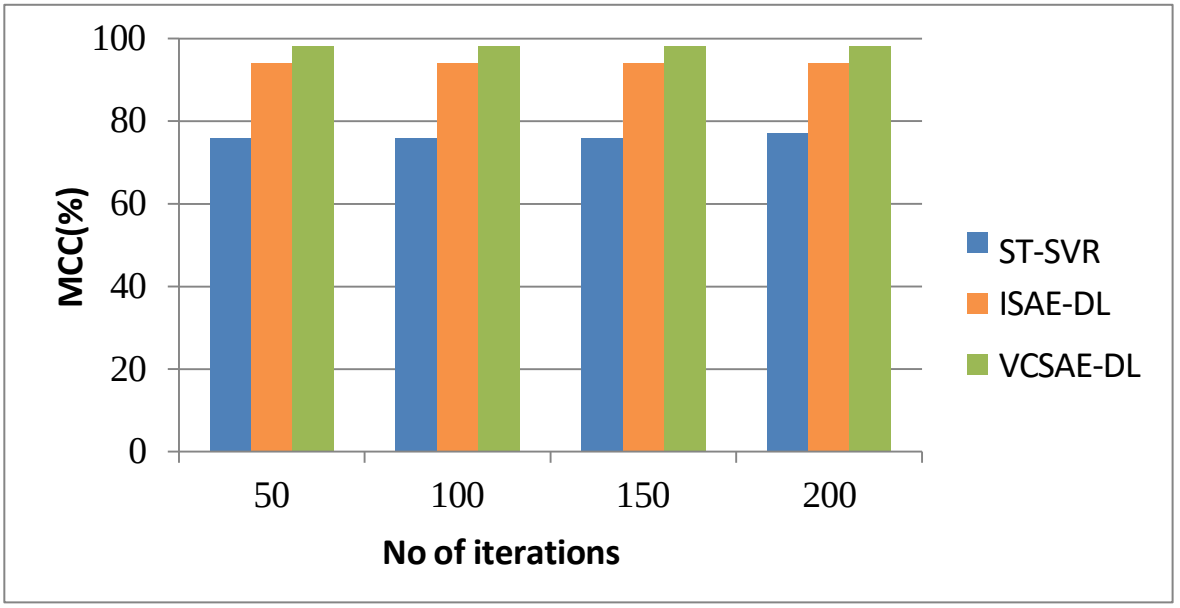

Figure 4.8 Comparison of MCC

4.3.6 F-Measure

Figure 4.9 compares the F-Measures of the proposed method VCSAE-DL with ST-SVR and ISAE-DL for different iterations. The experimental results indicate that F-measure of VCSAE-DL has improved by 11.97% and 5.28 % in the air quality prediction system. These findings demonstrate that VCSAE-DL exhibits higher F-measure values compared to ST-SVR and ISAE-DL

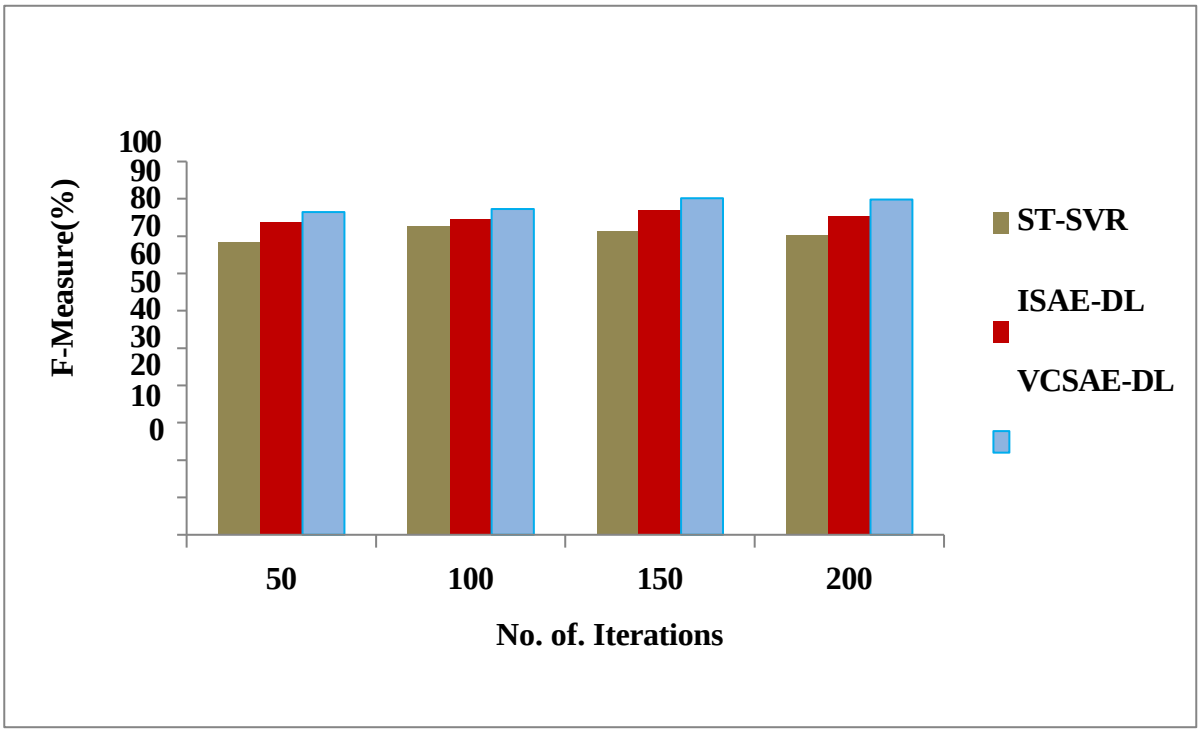

Figure 4.9 Comparison of F-Measure