## A SEMINAR REPORT

**on**

## Deep Embedded Clustering

Submitted in partial fulfilment for the requirements of the award of degree

**BACHELOR OF ENGINEERING**

**in INFORMATION TECHNOLOGY**

by

**Saikrishna Yeluri (160120733115) Under the guidance of**

**Dr.P Ramesh Babu(Associate Professor, IT Dept,CBIT) Mr. V Santosh(Associate Professor, IT Dept,CBIT)**

**Mr. R Govardhan Reddy(Associate Professor, IT Dept,CBIT)**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY(A)**

**(Affliated to Osmania University, Accerdited by NBA (AICTE) and NAAC (UGC), ISO Certified 9001:2015),Kokapet(V)**

**GANDIPET(M), HYDERABAD - 500 075**

**WEBSITE:** [**www.cbit.ac.in**](http://www.cbit.ac.in/)

**2022-2023**

## CERTIFICATE

Certified that seminar work entitled **“Deep Embedded Clustering”** is a bonafide work carried out in the seventh semester by “Saikrishna Yeluri (160120733115)” in partial fulfillment for the award of Bachelor of Engineering in Computer Science and Engineering from Chaitanya Bharathi Institute of Technology(A), Gandipet during the academic year 2022-2023.

**Seminar Cordinators Mentor**

**Mr B Govardhan Reddy Mr P Ramesh Babu**

Assistant Professor, Dept. of IT, CBIT, Assistant Professor, Dept. of, Hyderabad. Hyderabad.

**Mr V. Santosh**

Assistant Professor, Dept. of IT, CBIT, Hyderabad.

**Mr P Ramesh Babu**

Assistant Professor, Dept. of IT, CBIT, Hyderabad.

# Contents

[ABSTRACT 4](#_bookmark0)

1. INTRODUCTION 5
   1. [Scope and Research Problems . . . . . . . . . . . . . . 5](#_bookmark1)

[1.2 Objective and Goals . . . . . . . . . . . . . . . . . . . 5](#_bookmark2)

[1.3 Motivation . . . . . . . . . . . . . . . . . . . . . . . . . 5](#_bookmark3)

1. LITERATURE SURVEY 6
   1. [Deep Embedded Clustering Framework for Mixed Data 6](#_bookmark4)
   2. A Deep Embedded Clustering Algorithm for the Bin-

ning of Metagenomic Sequences . . . . . . . . . . . . . 7

* 1. Deep Clustering Bearing Fault Diagnosis Method Based on Local Manifold Learning of an Autoencoded Em-

bedding . . . . . . . . . . . . . . . . . . . . . . . . . . 8

1. SUMMARY AND CHALLENGES 11
2. CONCLUSION 13

[REFERENCES 14](#_bookmark5)

# ABSTRACT

This report provides an in-depth exploration of Deep Embedded Clustering (DEC), a cutting-edge unsupervised learning technique that leverages deep neural networks and traditional clustering algorithms to uncover latent data structures. DEC’s core principles and applications are discussed, with a particular emphasis on its versatil- ity across different data types. In addition to its general applications, this report covers DEC’s usage for mixed data, demonstrating its adaptability to heterogeneous datasets. It also delves into DEC’s application in the context of binning metage- nomic sequences, illustrating its potential in the field of genomics. Furthermore, the report highlights DEC’s effectiveness in bearings fault detection, showcasing its practical utility in predictive maintenance and fault diagnosis. The report offers practical insights into DEC’s implementation, including preprocessing and hyper- parameter tuning. DEC stands as a potent tool for revealing hidden patterns in complex datasets and demonstrates promise in diverse applications, ranging from mixed data analysis to genomics and industrial fault detection.

# INTRODUCTION

Deep Embedded Clustering (DEC) combines deep neural networks with clustering to uncover hidden data structures. This report ex- plores DEC’s principles, applications, and adaptability across diverse domains.

## Scope and Research Problems

This report investigates DEC’s applicability to mixed data analysis, metagenomic sequence binning, and bearings fault detection, addressing challenges related to het- erogeneity and real-world data complexities.

## Objective and Goals

Our objective is to explore and showcase DEC’s adaptability across these diverse domains, offering practical insights for effective implementation, and demonstrating its potential in solving real-world data-driven problems.

## Motivation

The motivation for this study arises from the need to leverage DEC’s power in ad- dressing practical challenges across various fields, such as genomics, industrial main- tenance, and mixed data analytics, ultimately contributing to data-driven decision- making and problem-solving solutions.

# LITERATURE SURVEY

## Deep Embedded Clustering Framework for Mixed Data

### Abstract:

Deep Embedded Clustering (DEC) is a powerful algorithm in the clustering domain, driven by deep learning. However, its applicability is limited to numerical data, and convergence issues can arise when soft assignment and target values differ signifi- cantly. To address these limitations, they introduced a novel deep embedded cluster- ing framework that accommodates mixed data and enhances convergence stability through soft-target updates, inspired by reinforcement learning techniques. Empiri- cal evaluations on diverse benchmark datasets containing mixed data showcase their framework’s superiority, outperforming existing clustering methods in standard met- rics. This work stands out as a state-of-the-art solution in the field of mixed data clustering.

### Introduction:

In the era of big data, researchers across various fields grapple with the challenge of extracting valuable insights from complex datasets. This study addresses two critical issues in clustering analysis. First, it extends the conventional clustering algorithms to handle mixed data types, combining numerical and categorical features. Second, it enhances clustering algorithm stability by employing a soft-target update tech- nique. The proposed approach demonstrates superior performance on UCI public datasets, outperforming existing methods. The study’s contributions lie in enabling clustering for mixed data and improving convergence stability. The paper covers related research, methodology, experiments, and concludes with a discussion of the proposed algorithm. **Methodology:**

In this section, the network architecture and training process for clustering has been outlined. The network employs an autoencoder structure based on a deep neural network for non-linear mapping of data into a latent space. The latent space facili- tates clustering, addressing the challenges of mixed data types and high dimension-

ality. The training process encompasses pre-training and clustering steps, focusingon mixed data representation and soft assignment methods for optimal performance. The architecture includes an embedding layer for categorical data, an autoencoderfor data compression, and trainable parameters for cluster centers. Novel techniques, such as simultaneous loss minimization and a soft-target network update, enhance clustering stability and performance.

## A Deep Embedded Clustering Algorithm for the Binning of Metagenomic Sequences

### Abstract:

Understanding microbial communities through metagenomic sequencing is crucial. Sequence classification into organisms, known as binning, is a key step in metage- nomic projects. Deep learning holds promise for accurate classification, but often lacks the necessary reference databases. Consequently, unsupervised learning is in- creasingly used, and MetaDEC presents a solution. This algorithm employs deep unsupervised learning, utilizing a two-phase approach for sequence grouping and cluster classification using an adversarial deep embedded clustering method. Experi- mental results demonstrate MetaDEC’s competitive performance on both simulated and real metagenomic data, addressing the challenges in unsupervised deep learning for binning.

### Introduction:

MetaDEC is an innovative unsupervised binning method tailored for metagenomic sequences, aiming to classify these sequences into distinct microbial organisms. This approach adopts a two-phase strategy, starting with sequence grouping based on over- lap information and then leveraging ADEC, an unsupervised deep learning method, for classification. ADEC employs an autoencoder architecture with a latent space interpolation factor to enhance data representation, optimizing it for clustering pur- poses. MetaDEC stands out as a unique and promising solution for metagenomic binning, offering a fresh perspective on sequence classification. In subsequent sec- tions, the report delves into experimental results, providing comprehensive compar- isons with existing binning techniques, showcasing the effectiveness and potential of MetaDEC in addressing the challenges of metagenomic sequence classification.

### Methodology:

MetaDEC employs a two-phase paradigm to overcome the challenge of lacking phy- logenetic information in short sequences during the clustering process. In Phase 1, sequences are grouped based on overlapping information, and cluster features are extracted. Utilizing a graph-based approach and multilevel partitioning, the method forms sequence groups, enhancing the efficiency and accuracy of subsequent clus-

tering. Phase 2 employs the ADEC (Adversarial Deep Embedded Clustering) deep learning method, classifying the sequence groups into clusters of closely related or- ganisms. Autoencoders are used to learn the latent space representation of the data, and k-means initializes the cluster centroids. The process further optimizes clus- tering through soft assignment and auxiliary target distribution. A discriminator network ensures the quality of the latent space, resulting in a robust and accurate binning method for metagenomic sequences.

## Deep Clustering Bearing Fault Diagnosis Method Based on Local Manifold Learning of an Autoencoded Embed- ding

### Abstract:

The ”Deep Clustering Bearing Fault Diagnosis Method” leverages an autoencoded embedding and local manifold learning to diagnose faults in bearings. This approach combines deep learning techniques with clustering methods to uncover hidden pat- terns in bearing data. By encoding the data into a latent space and performing local manifold learning, it becomes possible to detect and diagnose faults more effectively. This method holds promise for improving the accuracy and efficiency of bearing fault diagnosis, benefiting industries that rely on machinery and equipment with rotating parts.

### Introduction:

In recent years, deep learning has seen significant success in data-driven fault diag- nosis, primarily through classification models, which rely on labeled vibration signal data. These applications are typically categorized into scenarios where training andtest data share the same distribution or those where the distributions differ. However, in many practical fault diagnosis cases, acquiring labeled data is resource-intensive and often unattainable. This underscores the importance of developing unsupervised fault diagnosis models based on deep clustering techniques, which utilize deep neural networks for clustering.

Deep clustering methods can be divided into two categories: those with two stages, involving representation learning followed by clustering, and those employing joint optimization for feature learning and clustering. Although deep clustering has demon- strated remarkable progress in various applications, its application in unsupervised bearing fault diagnosis is underexplored.

This study proposes a novel unsupervised fault diagnosis method, E2LMC (Local Manifold learning of an auto Encoded Embedding). E2LMC leverages the frequency domain information as input, eliminating the need for time-consuming manual fea-

ture extraction. It re-embeds autoencoded representations using local manifold learn- ing to find a clusterable manifold, which is then clustered using a shallow clustering algorithm. Experimental results on the Case Western Reserve University (CWRU) bearing datasets demonstrate the effectiveness and efficiency of the proposed method. Moreover, E2LMC reduces the reliance on labeled data, making it suitable for real- world industrial applications.

### Methodology:

The presented approach involves three key stages for unsupervised fault diagnosis. First, original mechanical vibration acceleration signals are collected and converted into frequency domain data using the Fast Fourier Transform (FFT). In the pre- training stage, a multilayer autoencoder neural network is employed to learn low- dimensional representations of unlabeled data in an unsupervised manner. This enables meaningful feature extraction without the need for manual feature engineer- ing.

In the second stage, a local manifold learning method called UMAP (Uniform Mani- fold Approximation and Projection) is applied to identify clustering manifold struc-tures based on the representations learned in the pretraining phase. UMAP accu-rately captures local and global data structures while preserving relationships be- tween data points.

Finally, a Gaussian Mixture Model (GMM) is used for shallow clustering in the third stage. GMM is a versatile method for identifying different distribution patterns within a dataset, making it suitable for complex and varying data distributions. The combination of autoencoders, UMAP, and GMM allows for effective unsupervised fault diagnosis, enabling the identification of clusters in vibration signal data with-out the need for manual labeling, thus reducing the resource-intensive data labeling process.

|  |  |  |  |
| --- | --- | --- | --- |
| Research Paper | Concepts Used | Limitations | Insights/how can be improved |
| Deep Embedded Clustering Framework for Mixed Data | Deep Embedded Clustering, Embedding Layer,  One Hot Encoding | Sensitivity to Hyperparams, Bad encoding of mixed data | Enhance Data Collection,  Continuous Model Updates, Ensemble Approaches |
| A Deep Embedded Clustering  Algorithm for the Binning of Metagenomic Sequences | Deep Embedded Clustering, Adversarial Network, | Hyperparameter Tuning, Interpretability | Hyperparameter Optimization, Model Interpretability Techniques,  Regularization Techniques |
| Deep Clustering Bearing Fault  Diagnosis Method Based on Local Manifold Learning of an Autoencoded Embedding | DEC,  UMAP Manifold Learning, GMM | Imbalanced Data Impact, Performance Evaluation | Resampling Techniques, Evaluation Metrics  Ensemble of Balanced Models |

# SUMMARY AND CHALLENGES

DEC, or Deep Embedded Clustering, is a powerful technique in unsupervised ma-chine learning that involves jointly learning feature representations and clustering assignments. It has versatile applications across various domains, including mixed data analysis, binning of metagenomic sequencing data, and bearing fault detection.

Mixed Data Analysis: DEC can be applied to mixed data types, such as numerical and categorical data. By embedding the data into a continuous feature space, DEC enables the clustering of mixed data, making it useful in customer segmentation, anomaly detection, and more.

Metagenomic Sequencing Binning: In metagenomic sequencing, DEC helps with the binning of DNA fragments into taxonomically related groups. This aids in un- derstanding microbial communities and their functions within a sample, a crucial task in environmental and biomedical research.

Bearing Fault Detection: DEC’s application in bearing fault detection involves analyzing vibration or acoustic signals to identify potential faults in machinery or mechanical systems. By embedding sensor data into a feature space and performing clustering, DEC can effectively diagnose faults without the need for labeled training data.

DEC’s adaptability and effectiveness make it a valuable tool in scenarios where traditional supervised learning approaches are infeasible due to the absence of la- beled data, or in cases where insights need to be extracted from complex, mixed, or unstructured datasets.

Deep Embedded Clustering (DEC) is a powerful technique, but it comes with several challenges:

### Hyperparameter Tuning

One of the challenges in DEC is selecting the right hyperparameters, such as the number of clusters (K) and the learning rate. The choice of these parameters can significantly impact the clustering performance, and finding optimal values often requires experimentation.

### Initialization Sensitivity

DEC’s performance can be sensitive to the initialization of network weights and cluster centroids. The choice of initializations can lead to different local optima. Therefore, selecting suitable initializations is crucial for achieving good clustering results.

### Convergence Issues

Ensuring that the algorithm converges to a stable solution can be challenging. In some cases, the optimization process may get stuck in suboptimal solutions, and addressing convergence issues is essential for reliable clustering results.

### Scalability

DEC can be computationally expensive, especially for large datasets. Scaling the algorithm to handle big data efficiently is a significant challenge. Opti- mization techniques and parallelization methods may be required to make DECapplicable to real-world, large-scale problems.

# CONCLUSION

Deep Embedded Clustering (DEC) is a versatile technique with applications span- ning multiple domains.

In mixed data analysis, it efficiently categorizes and clusters data, enabling tar- geted marketing and fraud detection by integrating structured and unstructured data sources. In metagenomic sequencing, DEC facilitates the clustering of DNA sequences, aiding in the identification of microorganisms in various environments, benefiting environmental science and pharmaceutical research. For bearing fault de- tection, it learns patterns in vibration signals, distinguishing between healthy and malfunctioning machinery components, essential for manufacturing and transporta- tion.

DEC’s adaptability extends to document clustering for information retrieval and network security, where it identifies and isolates network anomalies. However, ad- dressing challenges like hyperparameter tuning, initialization sensitivity, convergence issues, and scalability is crucial to harness the full potential of DEC across these di- verse applications and beyond.

# REFERENCES

1. Jing An et al. “Deep clustering bearing fault diagnosis method based on local manifold learning of an autoencoded embedding”. In: *IEEE Access* 9 (2021), pp. 30154–30168.
2. Huynh Quang Bao, Tran Van Hoai, et al. “A deep embedded clustering algorithm for the binning of metagenomic sequences”. In: *IEEE Access* 10 (2022), pp. 54348–54357.
3. Xifeng Guo et al. “Improved deep embedded clustering with local structure preservation.” In:

*Ijcai*. Vol. 17. 2017, pp. 1753–1759.

1. Yonggu Lee, Chulwung Park, and Shinjin Kang. “Deep Embedded Clustering Framework for Mixed Data”. In: *IEEE Access* 11 (2022), pp. 33–40.
2. Junyuan Xie, Ross Girshick, and Ali Farhadi. “Unsupervised deep embedding for clustering analysis”. In: *International conference on machine learning*. PMLR. 2016, pp. 478–487.