e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

# SAMPLING IN EEG BASED DEPRESSION DETECTION SYSTEM USING SMOTE

**Amal Krishna NM*1, Dr. Priya S*2, Dr. Ashok Kumar T*3**

*1Student, Department Of Computer Engineering, Model Engineering College, Kerala, India.

*2Professor, Department Of Computer Engineering, Model Engineering College, Kerala, India.

*3Associate Professor, Department Of Computer Engineering, Model Engineering College, Kerala, India.

DOI : https://www.doi.org/10.56726/IRJMETS42161

## ABSTRACT

Sampling in EEG-based Depression Detection System Using SMOTE is a project that addresses the challenge of class imbalance in depression detection datasets. This research explores the application of advanced sampling techniques, including SMOTE, Gaussian SMOTE, and Adaptive Synthetic Sampling, to enhance the performance of depression detection systems using EEG signals. The project leverages Graph Neural Networks (GNNs) for predictive modeling, trained on EEG data that has been balanced using the aforementioned techniques. By addressing the issue of data imbalance, this study aims to improve classification accuracy and ensure robust performance of the GNN model. Among the sampling methods investigated, Gaussian SMOTE demonstrated superior capability in generating realistic synthetic samples, leading to notable improvements in model accuracy. This research has the potential to advance the field of mental health diagnosis by enabling more accurate and reliable detection of depression, which can be instrumental in clinical settings and personalized treatment strategies. A concise conclusion highlights the findings and their implications for future research.

**Keywords:** EEG-based Depression Detection, Sampling Techniques, SMOTE, Gaussian SMOTE, Adaptive Synthetic Sampling, Graph Neural Networks (GNNs).

## I. INTRODUCTION

The growing prevalence of mental health disorders, particularly depression, has emphasized the critical need for innovative and efficient methods for early detection and intervention. EEG (Electroencephalogram)-based depression detection systems have emerged as a promising area of research, leveraging advancements in signal processing and machine learning. Traditional approaches in analyzing EEG data often face significant challenges, such as class imbalance in datasets, which can hinder accurate diagnosis.

This research focuses on addressing these challenges by incorporating advanced sampling techniques like SMOTE, Gaussian SMOTE, and Adaptive Synthetic Sampling. The project's objective is to develop a robust system that balances EEG data effectively, thereby improving the accuracy of depression detection using Graph Neural Networks (GNNs). By automating and enhancing the process of analyzing EEG signals, this system aims to support clinicians and researchers in making precise and timely diagnosis.

The study also highlights the labor-intensive and complex nature of traditional EEG data analysis, which requires domain expertise and significant manual effort. By integrating sampling techniques with GNN models, this project reduces the burden of manual data handling while offering a scalable and efficient solution for mental health research. The technology developed in this study can be further applied in clinical diagnosis, personalized treatment strategies, and mental health monitoring, setting the foundation for future advancements in this critical field.

## II. METHODOLOGY

### Data sources

In this study, the primary data sources utilized are the open-source MODMA (Mental Disorder and Multimodal Analysis) datasets. These datasets were specifically curated for research and analysis related to mental health disorders. The MODMA dataset includes resting-state data collected from individuals in two distinct groups: healthy participants and those diagnosed with depression. This setup enables researchers to compare and analyze the underlying neurological differences between the two populations, contributing to a deeper understanding of depression and other related conditions.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

The MODMA dataset, released by Lanzhou University, is a comprehensive multimodal dataset designed to support studies in mental health diagnostics and therapy. For the collection of EEG data, the research team used a sophisticated 128-channel HydroCel Geodesic Sensor Net. This setup, along with Net Station acquisition software, facilitated the recording of high-quality, continuous EEG signals. The advanced 128-channel system ensures that neural activity across a wide range of scalp regions is captured, providing a detailed representation of the brain's electrical activity.

During the EEG recording process, participants were instructed to remain in a resting state, avoiding unnecessary movement. They were asked to stay awake and minimize actions such as blinking or making eye movements to reduce noise and artifacts in the EEG data. This approach ensures the acquisition of clean and reliable resting-state EEG recordings, which are crucial for accurate analysis. As part of this experiment, the researchers successfully recorded 128 resting-state EEG datasets. These datasets provide valuable insights into the neural mechanisms underlying mental disorders, serving as a critical resource for future research. Table 1 in the study outlines the specific clinical information for each participant group, offering additional context about the dataset. Furthermore, Figure 1 provides sample data from the EEG recordings, illustrating the nature and quality of the signals captured in this experiment.

**Table 1:** Details of dataset

| | Depressed patients | Healthy control subjects |
|---|---|---|
| Number of people | 24 | 29 |
| Gender (male/female) | 13/11 | 20/9 |
| Age (years) | 16-56 | 18-55 |
| Sampling frequency | 250 Hz | |
| Reference electrode | Cz | |
| Single acquisition time | 5 minutes | |

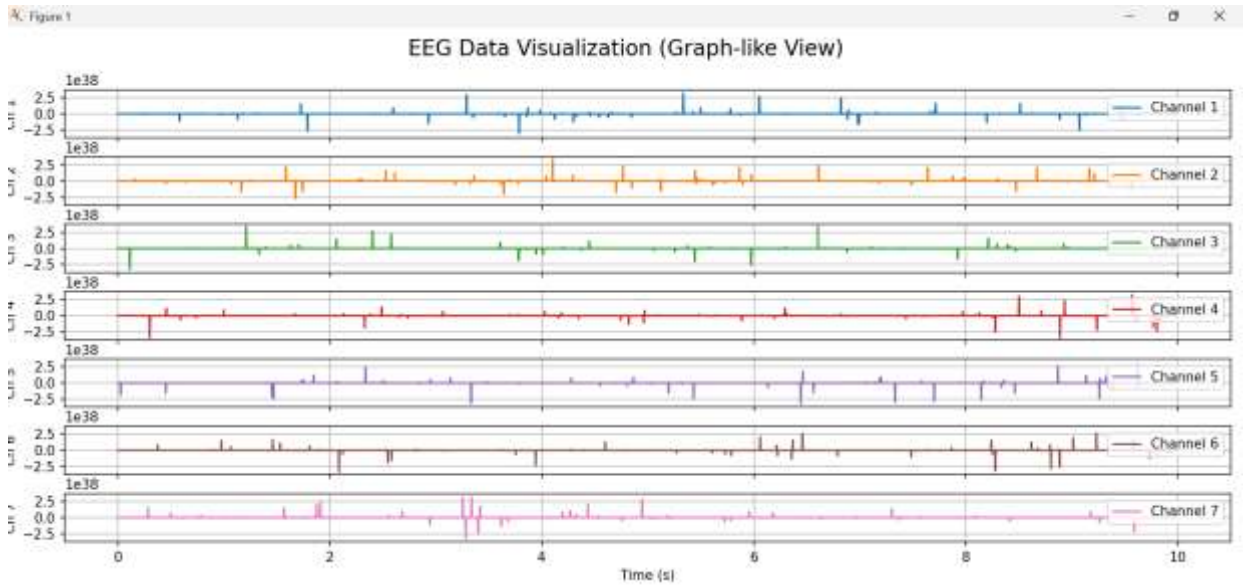


**Figure 1:** EEG sample data

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

**Data preprocessing**

The preprocessing pipeline for this research is designed with a meticulous approach to transform raw EEG data into a format that is optimized for effective depression detection using machine learning algorithms. The process begins by systematically loading the EEG data files from a specified directory. Using the mne library, which specializes in EEG and MEG data processing, each .raw file is read and stored as a raw object, allowing for efficient manipulation and access to the EEG signals. This initial step ensures that the raw data is appropriately handled and ready for subsequent preprocessing tasks.A critical component of this preprocessing pipeline is artifact removal through Independent Component Analysis (ICA). EEG data often contains artifacts caused by physiological processes such as eye blinks, muscle activity, and cardiac signals, which can distort the analysis. To prepare the data for ICA, it is first high-pass filtered to remove low-frequency noise, ensuring that the signal components of interest are preserved. Additionally, the data is downsampled to a manageable sampling frequency, reducing memory requirements and computational complexity while maintaining the integrity of the signal.

The ICA algorithm is then applied in chunks, a strategy that balances computational efficiency with the need for accurate artifact removal. This chunking approach allows the ICA to isolate and eliminate artifacts effectively, ensuring that the cleaned EEG signals accurately represent neural activity. After identifying artifact components through correlation with known artifact sources, these components are excluded, and the corrected EEG signals are reconstructed. Following artifact removal, frequency filtering is performed to isolate the frequency band of interest, typically between 1 Hz and 49 Hz, which contains critical information for depression detection. This step eliminates high-frequency noise and irrelevant low-frequency components, ensuring that the subsequent analysis focuses on the most informative signal features. The filtered EEG data is then converted into a numerical format using numpy for further processing.

To ensure the data's usability and consistency, missing or infinite values are handled by replacing them with zeros, a practical approach that maintains data integrity without introducing bias. Additionally, the number of EEG channels is standardized by padding signals where necessary, ensuring that all samples conform to the expected dimensions for input into the machine learning model. This step is particularly important for compatibility with Graph Neural Networks (GNNs), which require fixed input dimensions.

The final preprocessing step involves normalization, where the data is standardized by removing the mean and scaling by the standard deviation for each channel. This normalization step not only enhances the convergence of machine learning models but also ensures that variations in signal amplitude across different channels do not disproportionately affect the model's performance. A small epsilon value is added during normalization to prevent division by zero, further ensuring numerical stability.

The preprocessed EEG data is then saved as numpy arrays in a designated directory, making it readily available for subsequent training and evaluation phases. This pipeline effectively transforms raw EEG data into a clean, standardized, and normalized format, addressing key challenges in EEG signal processing and ensuring that the data is optimized for use in the depression detection system.

**System Architecture**

The system architecture for this project is designed to evaluate the impact of various sampling techniques, such as SMOTE, Gaussian SMOTE, and Adaptive Synthetic Sampling, on EEG-based depression detection. It begins with the Data Acquisition phase, where raw EEG signals are collected from recording devices and stored for further processing. In the Preprocessing stage, the raw data undergoes high-pass filtering to remove noise, Independent Component Analysis (ICA) to eliminate artifacts, and band-pass filtering to retain frequencies between 1–49 Hz. The data is then normalized to ensure consistency and padded to maintain uniform dimensions across samples. Following this, the Feature Engineering and Sampling phase converts the EEG signals into graph structures, treating EEG channels as nodes and defining adjacency through edges. Sampling techniques are then applied to balance the dataset, generating synthetic samples to address class imbalance and improve model performance.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

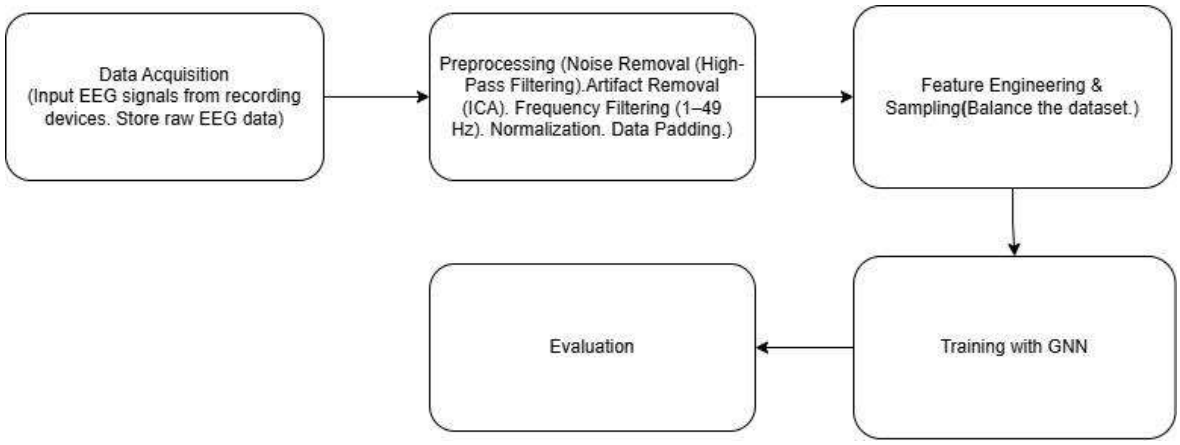


**Figure 2:** System Architecture

In the Training with GNN phase, the balanced dataset is used to train a Graph Neural Network (GNN) designed for binary classification. The GNN model leverages graph convolutional layers to extract spatial relationships from the data, with dropout layers incorporated for regularization. Finally, the Evaluation and Comparison stage measures the model's performance across different metrics, such as accuracy, precision, recall, and F1-score, comparing results achieved without sampling and with SMOTE, Gaussian SMOTE, and Adaptive Synthetic Sampling. This architecture enables a detailed analysis of how these techniques influence classification accuracy, emphasizing their role in improving the reliability of EEG-based depression detection systems.

**Sampling techniques**

In EEG-based depression detection, datasets often suffer from significant class imbalance, where one class (e.g., non-depressed subjects) is overrepresented compared to the other class (e.g., depressed subjects). This imbalance poses a critical challenge for machine learning models, as they tend to favor the majority class, leading to biased predictions and poor generalization for minority class instances. Consequently, achieving reliable and accurate detection of depression requires addressing this imbalance effectively.

To mitigate this issue, various sampling techniques have been employed, including Synthetic Minority Over-sampling Technique (SMOTE), Adaptive Synthetic Sampling, and Gaussian SMOTE. These methods generate synthetic samples for the minority class, balancing the dataset and improving model training. Here's an overview of the techniques:

**A. SMOTE**

SMOTE (Synthetic Minority Over-sampling Technique) is a widely used method to address class imbalance in datasets. It works by generating synthetic samples for the minority class, rather than duplicating existing samples, to ensure that the dataset is more balanced. This technique improves the performance of machine learning models by preventing them from being biased toward the majority class. SMOTE achieves this by creating synthetic data points along the line segments between a sample and its nearest neighbors in the feature space, thereby maintaining the diversity and variability of the minority class.

The mathematical representation of SMOTE can be expressed as follows: For a minority class sample Xi, and one of its k-nearest neighbors Xj, a synthetic sample Xnew is generated using the formula:

$$Xnew = Xi + \lambda \cdot (Xj - Xi),$$

where $\lambda$ is a random number drawn from a uniform distribution $\lambda \sim Uniform(0,1)$. This ensures that the synthetic data points are evenly distributed between the original samples and their neighbors.

**B. Gaussian SMOTE**

Gaussian SMOTE is an advanced variant of the Synthetic Minority Over-sampling Technique (SMOTE) designed to improve the diversity of synthetic samples by incorporating Gaussian noise. This approach generates synthetic samples not only along the line segments between existing minority samples and their nearest neighbors but also introduces a controlled randomness. By adding Gaussian-distributed noise, Gaussian SMOTE aims to create synthetic data points that better capture the natural variability of the minority class, which is particularly beneficial when dealing with non-linear relationships or complex feature spaces.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

The mathematical representation of Gaussian SMOTE is as follows: For a minority class sample Xi, a synthetic sample Xnew is generated using:

$$Xnew=Xi+N(\mu,\sigma2)$$

where N(μ,σ2) represents Gaussian noise with mean μ and variance σ2. This noise is added to Xi to produce synthetic samples that emulate the natural distribution of the minority class. By introducing Gaussian noise, this technique avoids the potential overfitting risks associated with traditional SMOTE and provides more realistic and representative synthetic samples. Gaussian SMOTE is particularly effective when the minority class exhibits complex, non-linear patterns, making it a valuable tool for handling imbalanced datasets.

**C. Adaptive Synthetic Sampling**

Adaptive Synthetic Sampling (ADASYN) is a dynamic approach to addressing class imbalance that focuses on generating synthetic samples for the minority class based on the difficulty of learning those samples. Unlike SMOTE, which generates synthetic samples evenly across the minority class, ADASYN prioritizes creating more samples for instances that are harder to classify, as determined by their proximity to the majority class in feature space. This targeted oversampling improves the model's ability to generalize and handle complex decision boundaries, making it particularly effective for imbalanced datasets with difficult or overlapping classes.

The mathematical representation of ADASYN is as follows: Let G be the total number of synthetic samples to be generated, distributed across the minority class samples Xi based on their learning difficulty Ri. The number of synthetic samples Gi generated for each sample Gi is defined as:

$$g_i = G \cdot \frac{r_i}{\sum_{k-1}^{n} r_k},$$

where $r_i$ is the difficulty score, calculated as:

$$r_i = \frac{\text{Number of majority class neighbors of } x_i}{k},$$

where k is the number of nearest neighbors considered.

For each minority sample xi, gi synthetic samples are generated as:

$$Xnew=Xi+\lambda\cdot(Xj-Xi),\lambda\sim Uniform(0,1),$$

where Xj is one of the k-nearest neighbors of xi within the minority class.

## III. RESULTS AND DISCUSSION

**Table 2:** Comparison of Different techniques

| Technique | Accuracy(%) |
|---|---|
| Without SMOTE | 54.55 |
| With SMOTE | 63.64 |
| Adaptive synthetic sampling | 63.64 |
| Gaussian sampling | 72.73 |

The results of this analysis, summarized in Table 2, showcase the influence of various sampling techniques on the model's accuracy for EEG-based depression detection. Without applying any sampling, the model achieves an accuracy of 54.55%, which highlights the challenges posed by class imbalance in the dataset. By incorporating SMOTE, the accuracy improves to 63.64%, showing the effectiveness of synthetic oversampling. Similarly, Adaptive Synthetic Sampling yields an identical accuracy of 63.64%, indicating its capability to

generate useful synthetic samples for balancing the dataset. Notably, Gaussian Sampling outperforms the other techniques, achieving the highest accuracy of 72.73%. This result underscores its strength in generating realistic and representative synthetic samples, which significantly enhance the model's performance. These findings, detailed in Table 2, demonstrate the critical role of sampling techniques in achieving accurate and reliable predictions for depression detection.

## IV. CONCLUSION

In this research, we thoroughly analyzed the effectiveness of various sampling techniques in addressing class imbalance for EEG-based depression detection. By incorporating SMOTE, Gaussian SMOTE, and Adaptive Synthetic Sampling, we demonstrated notable improvements in model performance compared to training without sampling. Specifically, Gaussian SMOTE emerged as the most effective method, achieving the highest accuracy of 72.73%, underscoring its ability to generate realistic and representative synthetic samples.

The proposed system leverages a Graph Neural Network (GNN) architecture for feature extraction and classification, achieving promising results in detecting depressive states. This study highlights the critical role of sampling techniques in enhancing the robustness and accuracy of depression detection models, especially in imbalanced datasets. Future work can explore further improvements by integrating feature-weighted fusion or incorporating domain-specific knowledge into the sampling process to refine performance even further.

## V. REFERENCES

[1] Z. Wan, J. Huang, H. Zhang, H. Zhou, J. Yang and N. Zhong,"HybridEEGNet: A Convolutional Neural Network for EEG Feature Learning and Depression Discrimination," in IEEE Access, vol. 8, pp.30332-30342, 2020, doi: 10.1109/ACCESS.2020.2971656.

[2] J. Shen et al., "Exploring the Intrinsic Features of EEG Signals via Empirical Mode Decomposition for Depression Recognition," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 356-365, 2023, doi: 10.1109/TNSRE.2022.3221962.

[3] Swaymprabha Alias Megha Mane, Arundhati Shinde, StressNet: Hybrid model of LSTM and CNN for stress detection from electroencephalogram signal (EEG), Results in Control and Optimization, Volume 11, 2023,100231, ISSN 2666-7207, https://doi.org/10.1016/j.rico.2023.100231.

[4] Maddirala AK, Veluvolu KC. ICA With CWT and k-means for Eye-Blink Artifact Removal From Fewer Channel EEG. IEEE Trans Neural Syst Rehabil Eng. 2022;30:1361-1373.
doi: 10.1109/TNSRE.2022.3176575.

[5] N. S. Amer and S. B. Belhaouari, "EEG Signal Processing for Medical Diagnosis, Healthcare, and Monitoring: A Comprehensive Review," in IEEE Access, vol. 11, pp. 143116-143142, 2023, doi:10.1109/ACCESS.2023.3341419.

[6] Y. Song, Q. Zheng, B. Liu and X. Gao, "EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 710-719, 2023, doi: 10.1109/TNSRE.2022.3230250.