BRAUDE
College of Engineering, karmiel

Software Engineering Department Braude College

Capstone Project Phase A

24-2-R-4

Citation Classification Using Graph Diffusion Models

Students:

Roei Nizri – RoeiNizri1@gmail.com, 318260742

Ran Polac – RanPolac@gmail.com, 318265808

Supervisors: Dr. Renata Avros, Prof. Zeev Volkovich

Table of Contents

Abstract

Anomaly detection on graphs presents significant challenges, primarily due to label imbalance and data insufficiency. These factors make learning and finding real anomalies in complex graph data hard. To address this, the study introduces a diffusion model-based graph generator capable of simultaneously producing graph structures and node features in a multitask fashion, focusing on generating synthetic anomalous examples. By integrating these synthetic anomalies into the training dataset, we achieve a more balanced dataset, significantly enhancing the training and performance of downstream anomaly detection models. The approach addresses the challenges of label imbalance and data insufficiency in graph anomaly detection, providing a robust framework for generating synthetic anomalies and improving the effectiveness of anomaly detection models.

Keywords: Anomaly Detection, Graph Neural Networks, Diffusion Models, citation classification.
GitHub: https://github.com/RanAndRoei/FinalProject

1. Introduction and Background

Anomaly detection has been a critical task in the realm of machine learning and data mining applications, due to its significance in identifying irregular patterns that deviate from the norm in large datasets. The main issue is the difficulty in identifying anomalous nodes within graph-structured data due to the scarcity of anomaly labels compared to normal data. This imbalance makes it hard for detection models to learn effectively, leading to suboptimal performance in detecting true anomalies. Several approaches have been proposed for anomaly detection on graphs. Methods like Graph Convolutional Networks [2] (GCNs) and Graph Attention Networks (GATs) [4] learn node embeddings and identify nodes that deviate from learned patterns. However, these methods often struggle with label imbalance, making it difficult to identify rare anomalies accurately. Random Walk Algorithms use random walks to capture proximity and structural patterns in the graph, identifying anomalies based on deviations in the random walk distributions. However, they may be computationally expensive and less effective in large or highly connected graphs. Factorizing the adjacency matrix of the graph to uncover latent features, with anomalies detected based on outliers in the latent space. This method may not fully capture the intricate relationships and dependencies within the graph, leading to suboptimal anomaly detection. To address the limitations of existing methods, we propose a diffusion model-based graph generator. This generator is capable of simultaneously producing graph structures and node features in a multitask fashion, with a particular focus on generating synthetic anomalous examples. By integrating these synthetic anomalies into the training dataset, we achieve a more balanced dataset, significantly improving the training and performance of downstream anomaly detection models. The solution is designed to benefit the following stakeholders.
1. **Data Scientists and Machine Learning Engineers**: These professionals are constantly seeking advanced techniques to improve the accuracy and efficiency of anomaly detection systems.

2. **Cybersecurity Experts:** Cybersecurity professionals use anomaly detection to identify suspicious activities and potential threats within network traffic or user behavior.

3. **Financial Analysts and Fraud Detection Specialists:** These individuals are interested in detecting fraudulent activities, such as abnormal transaction patterns or insider trading, which can be represented as graphs.

The proposed model enhances the detection of fraudulent transactions and anomalies in graph-structured data by generating synthetic anomalies and addressing label imbalance. This approach leads to more accurate and reliable data, improving overall security measures. In the following sections, we will delve deeper into the methodology of the diffusion model-based graph generator, discuss its advantages over existing methods, and present experimental results demonstrating its efficacy in anomaly detection tasks.

2. Related Work (Literature Review)

Anomaly detection on graphs is growing quickly because of the unique challenges with graph data. Unlike regular tabular data, graphs show complex relationships [13] and dependencies between entities, making anomaly detection harder. Early methods, such as clustering and proximity-based techniques, were designed to handle simpler graph structures but struggled with more complex, dynamic graphs. For example, proximity-based methods identify anomalies by measuring the similarity between nodes, which works well for small, static graphs but fails in larger, evolving graphs with diverse relationships. To address these challenges, researchers have developed various algorithms that leverage the structural properties of graphs. **Graph-based Clustering Algorithm** [1] algorithms segment the graph into clusters based on node similarity. Nodes that do not fit well into any cluster are considered anomalies. However, these methods struggle with high-dimensional and dynamic graphs. **Graph autoencoders** [2] learn low-dimensional representations of nodes and reconstruct the graph structure. Anomalies are identified as nodes with high reconstruction errors. While effective, autoencoders require careful tuning and may not generalize well to different types of graphs.

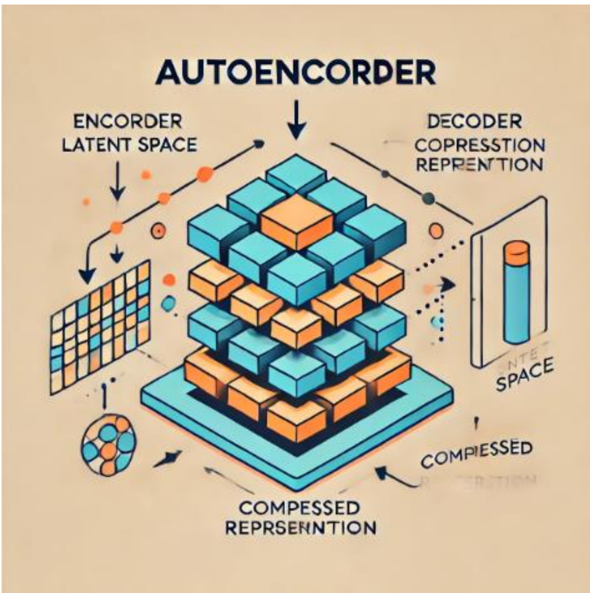


Figure 1: The architecture of a the Autoencoder

Generative Adversarial Networks [3] create synthetic graph data to improve the training set and fix label imbalance. Although promising, GANs [11] require a lot of computing power and can be hard to train.

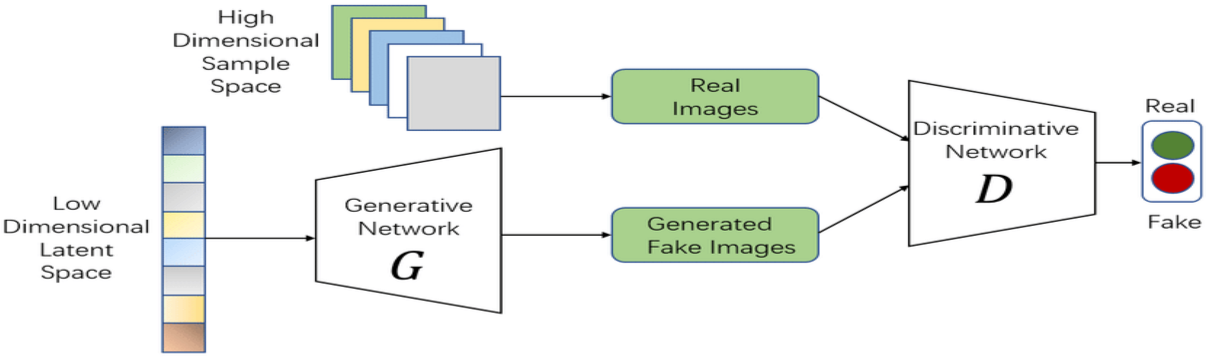


Figure 2: The architecture of a basic GAN

2.1 Graph Neural Networks (GNNs)

Graph Neural Networks (GNNs) [1] have gained prominence due to their ability to learn complex patterns and dependencies within graph data. GNNs can capture both local and global structures, making them effective for detecting anomalies in various applications such as fraud detection, network security, and social network analysis. One notable technique is the Graph Convolutional Network (GCN) [17], which extends the concept of convolutional neural networks to graph-structured data. GCNs aggregate information from neighboring nodes, allowing them to learn node representations that incorporate the graph's structure. This method has been successful in identifying anomalies in social networks and citation networks where the relationships between entities are crucial for detecting irregular patterns. However, GCNs often require large, labeled datasets for training, which can be a limitation in scenarios with scarce anomaly labels. Another advanced GNN technique is the Graph Attention Network (GAT), which introduces an attention mechanism to weigh the importance of different neighbors during the aggregation process. This allows GATs to focus on the most relevant parts of the graph, improving the detection of subtle anomalies. Despite their effectiveness, GATs can be computationally intensive and may still struggle with highly imbalanced datasets.

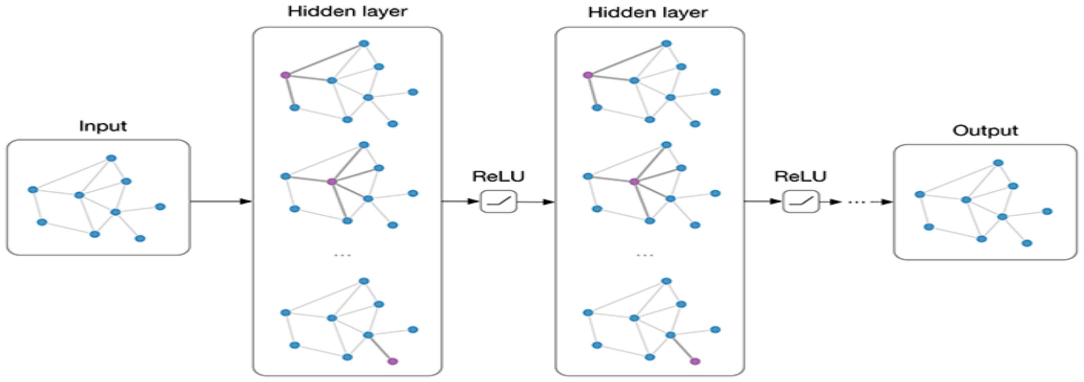


Figure 3: Multi-layer Graph Convolutional Network (GCN) with first-order filters

2.2 Encoding with Graph Autoencoder

Graph autoencoders [2] are used to map anomaly detection graphs into a latent space, simplifying the graph data while retaining its most salient features. The autoencoder consists of two main parts:

Encoder: Transforms input graph data into a lower-dimensional latent space by compressing the data.

Decoder: Reconstructs the input data from the latent representation, ensuring that the learned representations are meaningful.

Graph autoencoders [2] have been effective in identifying anomalies as nodes with high reconstruction errors. However, they require careful tuning and may not generalize well to different types of graphs. Variational Graph Autoencoders (VGAEs) [4] [15] extend this approach by incorporating a probabilistic framework, which improves generalization by predicting a distribution over the latent space. This allows VGAEs to better handle variability in graph data and generate more realistic reconstructions.

2.3 Generating Graph Structure

A traditional autoencoder is first applied to generate node attributes. This autoencoder learns efficient representations of the data, capturing the essential characteristics while reducing dimensionality. The process involves compressing the input data through the encoder and reconstructing it via the decoder.

Training: The model adjusts weights to minimize the difference between the original input and its reconstruction using a loss function Mean Squared Error (MSE). Through this process, the autoencoder learns to prioritize the most significant features in the input data, effectively learning a compressed but informative representation. Variational Autoencoder (VAE): Enhanced Autoencoder: Introduces a probabilistic approach to improve generalization.

Process: The encoder predicts two parameters (μ, µ) and standard deviation (σ) for a Gaussian distribution. During the forward pass, a sample is drawn from this distribution and decoded to reconstruct the input data.

Loss Function: Combines the reconstruction term (MSE) between input and reconstructed output) and a regularization term (Kullback-Leibler (KL)) [9] divergence between the learned distribution and a standard Gaussian). The VAE improves by predicting a Gaussian distribution's parameters (mean and standard deviation) and using a combined reconstruction and regularization loss function.

Conditional Generation for Nodes with Positive Labels controls over generating different node types (anomalous and normal). The methodology for generating graph structures employs the Variational Graph Autoencoder (VGAE), which captures both topological and feature information into a latent space. The VGAE model uses a shared Graph Neural Network (GNN) encoder for multi-task learning, adapting the VAE architecture to graph data. To handle heterogeneous graphs with multiple node and edge types, the standard VGAE encoder is replaced with a Heterogeneous Graph Transformer (HGT) [7]. The HGT can process diverse node and edge types, allowing for the accurate generation of adjacency matrices specific to each edge type. This approach enhances the model's ability to capture complex relationships in heterogeneous graphs, making it more effective for anomaly detection in such data. Different decoders are used for different edge types, allowing accurate generation of adjacency matrices specific to each edge. The transition to matrices allows a more accurate comparison between the weights of the edges and nodes.

2.4 Timestamp Generation

A temporal generator designed to generate timestamps parallels the structure of the feature generator but targets singular dimensional output. The generated timestamps are integrated into the diffusion model's latent space, ensuring that the synthetic data generated (including anomalies) maintains realistic temporal characteristics. Timestamps help capture the evolution of the graph over time, enabling the detection of anomalies that occur due to changes in the graph structure or node attributes. This temporal aspect is crucial for applications like network intrusion detection and fraud detection, where the timing of events can provide important clues about anomalous behavior.

Purpose: Generate timestamps as part of the graph data.

Architecture: Like the feature generator but focused on generating a single continuous variable representing time.

Loss Function: Uses MSE to minimize the average squared difference between estimated and actual timestamps, ensuring precision and reliability in the generated temporal data.

2.5 Diffusion models

Diffusion models, also called diffusion probabilistic models or score-based generative models, gradually transform simple distributions, like Gaussian noise, into the desired data distribution. This transformation is achieved through a series of small, reversible steps, allowing the model to learn the underlying data structure effectively. Recent advances in generative models, especially diffusion models, offer a promising approach to addressing the challenges of anomaly detection on graphs. Initially developed for image synthesis, these models have shown exceptional capability in capturing complex data distributions through a process of gradual noising and denoising. Although their application to graph data is relatively new, it holds significant potential. One key advantage is their ability to generate realistic graph structures and node features, which can be used to create synthetic anomalies. Consequently, this approach helps mitigate the label imbalance problem by enriching the training datasets with synthetic anomaly nodes. By generating synthetic graphs that mirror the intricacies of real-world graphs, these models provide a robust foundation for training anomaly detection algorithms. Moreover, they excel at capturing the complex, high-dimensional relationships inherent in graph data, making them well-suited for generating realistic synthetic anomalies. By learning the normal patterns within the graph data, the model can effectively identify anomalies as deviations from these learned patterns. This process involves adapting the diffusion process to work with the complex relationships and structures inherent in graph data. In the context of anomaly detection, the Denoising Diffusion Probabilistic Model (DDPM) [12] enhances the quality of synthetic anomaly generation. DDPM operates by adding noise to data points and then learning to reverse this process, effectively generating data points from noise. This process involves adapting the diffusion process to work with the complex relationships and structures inherent in graph data. In summary, by leveraging their unique capabilities, researchers can create enriched datasets that improve the robustness and accuracy of anomaly detection algorithms, making these models a powerful tool in the ongoing effort to enhance anomaly detection on graphs.

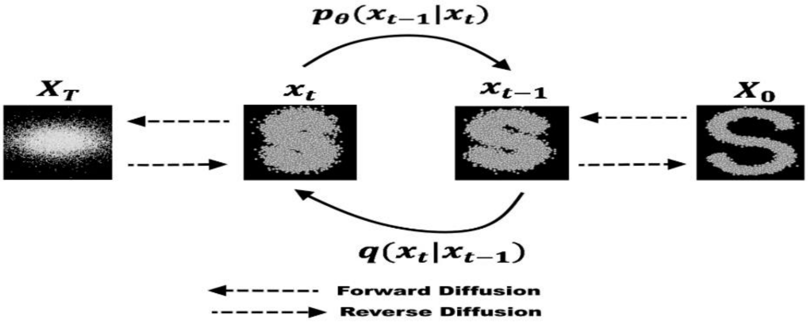


Figure 4: The main idea of diffusion model. By adding noise in the diffusion process the image became more cleaner and by imply reverse diffusion process the image return to the origin.

2.5.1 Diffusion Models in Latent space

Diffusion models in latent space offer a sophisticated approach to generating synthetic data, including anomalies, for anomaly detection tasks. These models operate by transforming input data into a latent representation, adding controlled noise through a diffusion process, and then applying a denoising process to reconstruct the data. This methodology allows the generation of realistic graph structures and node features that closely resemble real-world data while incorporating synthetic anomalies essential for robust training of anomaly detection models. The process begins with encoding the input graph data, which includes graph nodes and their features, into a latent space. The encoding phase utilizes techniques such as autoencoders or variational autoencoders (VAEs) to compress the high-dimensional graph data into a lower-dimensional latent representation. This latent representation retains the most significant features and structures of the original data, making it more manageable and suitable for further processing. Once the data is encoded into the latent space, the diffusion process introduces noise incrementally. The diffusion process can be mathematically described as a series of steps where Gaussian noise is added to the latent representation based on a predefined variance schedule. This gradual introduction of noise transforms the data into a more chaotic state, which the model learns to reverse. The key idea is to train the model to understand the underlying data distribution by learning how to effectively add and remove noise. The denoising process follows, where the model learns to reverse the added noise, reconstructing the original data from the noisy latent representations. This process involves training a neural network to predict the clean data from its noisy version at each step. By iteratively applying this denoising process, the model generates data that closely mirrors the original input, including the generation of synthetic anomalies.

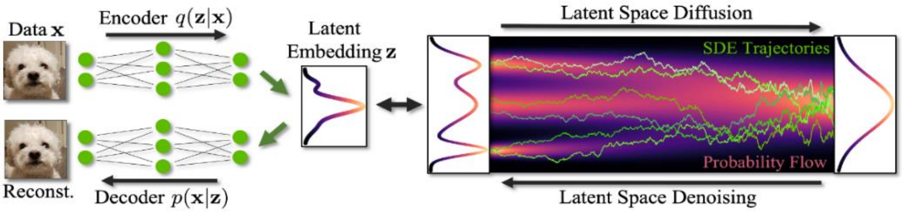


Figure 5: The process of the model, the data (citations) are encoded into a latent space, where a diffusion process introduces controlled noise, generating synthetic anomalies. The denoised data is then decoded back, allowing for robust anomaly detection.

2.6 DeepWalk and Node2Vec

DeepWalk [5] and Node2Vec [6] are algorithms that generate node embeddings by simulating random walks on the graph. These embeddings capture the structural relationships between nodes, allowing for the identification of anomalies as nodes with embeddings that deviate significantly from the norm. DeepWalk performs truncated random walks on the graph and treats these walks as sentences to train a Skip-Gram model, like word2vec in natural language processing. Node2Vec extends this by using a biased random walk that balances between breadth-first and depth-first strategies, allowing for more flexible graph exploration. While effective for homogeneous graphs, these methods struggle with heterogeneous graphs and may miss complex relationships.

2.7 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) [3] generate synthetic graph data to augment the training set and address label imbalance. GANs consist of two neural networks: a generator that creates synthetic data and a discriminator that evaluates the authenticity of the data. [11] This adversarial process drives the generator to produce increasingly realistic data. While GANs have shown promise in generating high-quality synthetic data, they can be computationally intensive and challenging to train. The training process requires careful balancing between the generator and discriminator to prevent issues like mode collapse. Despite these challenges, GANs offer a powerful tool for generating synthetic anomalies and enriching the training datasets for anomaly detection models. In conclusion, the field of anomaly detection on graphs has seen significant advancements with the development of machine learning and generative models. Graph Neural Networks [1] and generative models, particularly diffusion models, offer powerful tools to address the unique challenges of graph-structured data. The proposed diffusion model-based graph generator represents a approach to mitigate label imbalance and enhance the training of anomaly detection algorithms. As the field continues to evolve, these advancements hold great promise for improving the detection of anomalies in complex, dynamic graph data.

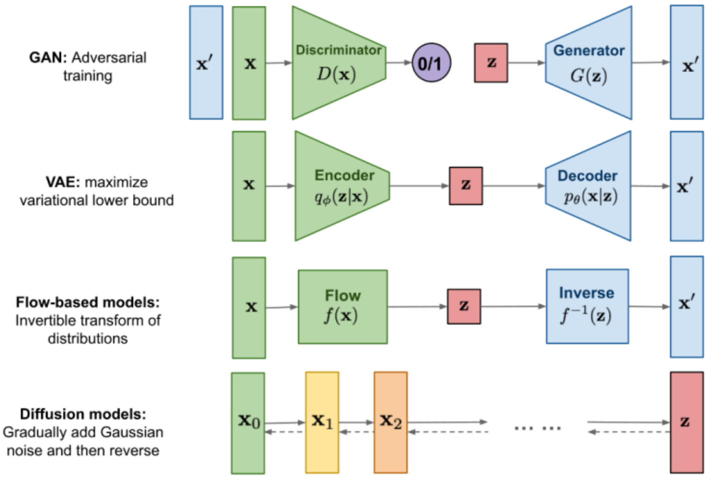

GAN: Adversarial training
x'
x
Discriminator
$D(\mathbf{x})$
0/1
z
Generator
$G(\mathbf{z})$
x'
VAE: maximize variational lower bound
x
Encoder
$q_\phi(\mathbf{z}|\mathbf{x})$
z
Decoder
$p_\theta(\mathbf{x}|\mathbf{z})$
x'
Flow-based models: Invertible transform of distributions
x
Flow
$f(\mathbf{x})$
z
Inverse
$f^{-1}(\mathbf{z})$
x'
Diffusion models: Gradually add Gaussian noise and then reverse
$\mathbf{x}_0$
$\mathbf{x}_1$
$\mathbf{x}_2$
... ...
z

Figure 6: Overview of different types of generative models.

3. Expected Achievements

This project tackles the problem of imbalanced data in graph anomaly detection by creating a strong framework for generating synthetic anomalies. This helps improve the performance of anomaly detection models. We will demonstrate the effectiveness of our approach through various tests and evaluations on different datasets related to graph compression, showcasing its performance and benefits. This approach aims to enhance the effectiveness of anomaly detection models through several key achievements.

Improved Detection Accuracy: By generating synthetic anomalies, the project aims to create more balanced training datasets, leading to improved detection accuracy of anomaly detection models. This enhancement will help in identifying true anomalies more effectively, reducing the rate of false positives and false negatives.

Enhanced Model Robustness: The introduction of synthetic anomalies is expected to improve the robustness of anomaly detection models. These models will be better equipped to handle variations in data and identify anomalies even in the presence of noise and other irregularities. Integration with Existing Systems: The framework is designed to be easily integrated with existing anomaly detection systems and workflows. This compatibility will enable organizations to enhance their current systems without the need for extensive modifications or overhauls, in several areas such as blockchain, image processing, disease detection, finding anomalies in citations, etc.

Identification of Suspected Citations: The project will *focus on finding suspected citations* in the data, which may indicate anomalies or outliers. By detecting and analyzing these citations, the model can gain deeper insights into the underlying patterns and relationships within the graph data. We aim to implement an advanced anomaly detection system for a large database of citations using the methodology outlined in the article on graph diffusion models for anomaly detection. The approach involves converting the dataset of citations into a graph format where nodes represent individual citations and edges represent relationships such as shared authorship or semantic similarity. We will utilize a graph autoencoder to encode this graph into a latent space, capturing the structural and feature information of the citations. A diffusion model will then be applied in the latent space to generate synthetic anomalies, addressing the label imbalance issue common in anomaly detection tasks. By adding conditional generation and temporal features, the model will improve the training dataset, making anomaly detection more robust and accurate. This system will help us identify suspected citations in graph-structured data, ensuring data integrity and providing insights.

4. Engineering Process

1. **Prepare Dataset**
Gather a large dataset of citations, including metadata such as authors, publication dates, and sources. Represent the dataset as a graph where: nodes represent individual citations. Edges represent relationships between citations, such as being from the same author, same publication, or similar content.
Input: A collection of citations with associated metadata.
Output: A graph representation of the dataset with nodes as citations and edges as relationships between citations.
2. **Graph Construction**
Each node (quote) should have features such as text embeddings, author information, and publication details. Create edges based on relationships like shared authorship, publication proximity, or semantic similarity.
Input: Dataset with node features and edge criteria.
Output: A fully constructed graph with nodes and edges defined.
3. **Graph Autoencoder implementation**
We use a graph neural network (GNN) to encode the graph into a latent space, capturing the structural and feature information. Train the autoencoder to reconstruct the original graph from the latent representation.
Input: Constructed graph with node features and edges.
Output: Latent space representation of the graph and reconstructed graph from the autoencoder.
4. **Diffusion Model application**
Implement a diffusion model in the latent space to generate synthetic anomalies. This involves iterative noise addition and removal to create synthetic nodes (citations) that resemble anomalies. We ensure the model can generate nodes with specific labels (anomalous or normal) based on the citation's characteristics.
Input: Latent space representation of the graph.
Output: Synthetic anomalous nodes in the latent space and their corresponding graph representations.
5. **Anomaly Detection**
We use the synthetic anomalies along with real data to train an anomaly detection model. Incorporate both real and synthetic data to address label imbalance.
Assess the model's performance using metrics like AUROC and AUPRC to ensure it effectively identifies anomalies.
Input: Real and synthetic graph data with labels.
Output: Anomaly detection model capable of identifying anomalous citations with performance metrics.

4.1 Pre-processing

The purpose goal is to enrich the dataset with synthetic anomaly nodes for robust training of anomaly detection systems. These synthetic nodes include four key components:
Node Feature: Represents various attributes associated with the node, such as metrics or characteristics that describe its properties and behavior within the graph.
Node Timestamp: A temporal marker indicating when a specific event or interaction involving the node occurred, crucial for analyzing the temporal dynamics of the network.
Positive Label: A binary indicator denoting the anomaly nature of the node, used to differentiate between normal and anomalous behavior for training and validating detection models.
Heterogeneous Graph Structure: Different types of nodes and edge connections.

4.1.1 Graph Autoencoder

The encoding process involves using a graph autoencoder to map anomaly detection graphs into a latent space. The autoencoder consists of two main parts:

Encoder: In the implementation for identifying suspected citations in an article, the encoder processes each citation (node) and its relationships (edges) by:

Embedding Text Features, converting the textual content of citations into dense vector representations using techniques like word embeddings or transformer models.

Incorporating Metadata, adding author information, publication details, and other relevant metadata to the node features. Using graph convolutional layers to capture the relationships between citations, like shared authorship or semantic similarity, and transforming these relationships into a compact form that keeps the important patterns and connections.

Decoder: Reconstructs text features and attempts to regenerate the textual content of citations from their latent representations.

Restores Metadata and Relationships, reconstructs the author information, publication details, and relationships between citations based on the latent space representations.

Validates Reconstruction, ensures that the reconstructed graph closely resembles the original input graph, maintaining the key features and structures identified by the encoder.

4.1.2 Generating Node Attributes

To control the creation of node types, especially anomalous ones, conditional variables are added to the VAE. During training, labels indicating node types are concatenated with feature vectors and fed into the encoder. During generation, the decoder uses these labels to create feature vectors that match the specified labels. This ensures that when the label indicates 'fraudulent' (anomalous), the generated features are characteristic of fraudulent nodes.

The benefit is to allow a more controlled generation of nodes, improving the model's ability to differentiate and generate distinct node types, which is crucial for applications like anomaly detection. In the implementation for identifying suspected citations, the process begins with preparing a graph where nodes represent citations and edges represent relationships such as shared authorship or semantic similarity. The citations have attributes including text embeddings, author information, and publication details. The autoencoder's encoder processes each citation and its relationships, compressing this information into a lower-dimensional latent space while retaining essential features. The decoder then reconstructs the original quote features from the latent representation, minimizing the reconstruction error (MSE) and ensuring the latent space captures the critical attributes of the citations. The VAE encoder predicts the mean and standard deviation for the Gaussian distribution representing each citation in the latent space. A sample is drawn from this distribution and decoded to reconstruct the citation, optimizing both the reconstruction accuracy and the regularization term (KL divergence). During training, node-type labels (normal or anomalous) are combined with feature vectors and processed by the encoder. This ensures the latent representation includes information about the node type. In the generation phase, the decoder uses these labels to produce feature vectors corresponding to the specified node type, enabling controlled generation of normal and anomalous citations. The VAE generates synthetic citations labeled as anomalous by using conditional variables. These synthetic anomalies imitate suspected citations that differ from usual patterns. The anomaly detection model is then trained on both real and synthetic data, using the synthetic anomalies to address label imbalance, enhancing the model's ability to identify suspected citations. This approach improves feature extraction, allows for precise generation of anomalous citations, and enhances the anomaly detection model's ability to detect suspected citations, ensuring robust performance.

4.1.3 Generating Heterogeneous Graph Structure

Following the previous step, to handle heterogeneous graphs with multiple node and edge types, the standard VGAE encoder is replaced with a Heterogeneous Graph Transformer (HGT) [16]. The HGT can process diverse node and edge types, allowing for the accurate generation of adjacency matrices specific to each edge type. This approach enhances the model's ability to capture complex relationships in heterogeneous graphs, making it more effective for anomaly detection in such data. Separate decoders are used for different edge types, allowing accurate generation of adjacency matrices specific to each edge. The transition to matrices allows a more accurate comparison between the weights of the edges and nodes. In the implementation for identifying suspected citations, the HGT begins by preparing a graph where nodes represent citations and edges represent relationships such as shared authorship or semantic similarity. The HGT processes each citation and its relationships, capturing the diverse and complex interactions within the graph. This step ensures the model can accurately generate adjacency matrices for different edge types, enhancing its ability to detect unusual patterns.

4.1.4 Timestamp Generation

The anomaly detection system adds synthetic anomaly nodes to datasets, using a temporal generator that works like the feature generator but focuses on a single output dimension. This temporal data significantly expands the space of time within the dataset, allowing for a more comprehensive analysis of temporal patterns and trends. By generating accurate and reliable timestamps, the system can monitor the progression and evolution of data points over time, enhancing the ability to detect anomalies that may arise due to unusual temporal behaviors. Furthermore, this step allows us to monitor the learning process of anomaly detection models. By adding accurate timestamps, we can check how well the models identify temporal anomalies. This ensures they detect anomalies based on both static features and event timing. This dual focus improves the system's ability to identify suspected citations, leading to more reliable detection outcomes.

4.2 Diffusion Model in Latent Space

Diffusion models are used to generate new data by adding noise to data points and then learning how to reverse this process, effectively generating data points from noise. This process can be particularly useful in generating synthetic anomalies for anomaly detection. The diffusion model in latent space is a crucial technique used to improve the generation quality of synthetic anomalies in graph data, which is essential for effective anomaly detection. In the implementation for identifying suspected citations, we adopt the Denoising Diffusion Probabilistic Model (DDPM), which stands for "Diffusion models with Denoising Priors." This type of generative model simulates the gradual diffusion (or spreading out) of noise over time. To explain this further, imagine that each citation in the dataset is a point in a complex graph. By introducing noise, we can perturb these points slightly and then train the model to learn how to revert these changes, effectively teaching it the underlying structure of normal and anomalous citations. In practice, this means we first add noise to the data points in the citation graph and then use the DDPM to denoise or reverse this process. This helps the model learn to differentiate between normal and anomalous patterns. For instance, if a citation suddenly deviates significantly from the established pattern of a research domain, the model, having learned from the noise reversal process, can identify this as a potential anomaly. In the context of graph anomaly detection, the divergence between the prior distribution and data distribution can be substantial due to the complexity of the graph structure. This divergence often results in suboptimal performance in downstream anomaly detection tasks. To address this issue, the DDPM improves the quality of synthetic anomaly generation, enhancing anomaly detection performance. By applying the diffusion model in the latent space, we effectively address the challenge of identifying suspected citations. This approach allows the system to generate high-quality synthetic anomalies, providing a richer and more varied set of examples for training. As a result, the anomaly detection system becomes more robust and accurate. Furthermore, by generating synthetic anomalies that mimic real-world deviations, we can fine-tune the detection system to recognize even the most nuanced irregularities. This comprehensive approach ensures the system can effectively identify and manage suspected citations, enhancing the overall reliability and effectiveness of anomaly detection.

4.2.1 Diffusion Model Application

In practice, we add noise to the citation data to create variations, which the model then learns to denoise. This training process helps the model identify what constitutes a typical citation pattern versus an anomaly. For example, if a citation pattern deviates significantly from established norms, such as citing unrelated topics or sources, the model can detect this irregularity based on its learned experience. Applying the diffusion model in the latent space helps address the challenge of divergence between the prior distribution and data distribution, which is common in complex graph structures like citation networks. The DDPM helps mitigate this issue by improving the quality of synthetic anomaly generation, thereby enhancing the anomaly detection system′s performance. This allows the model to generate high-quality synthetic anomalies that closely mimic real-world anomalies, providing a richer dataset for training. Moreover, generating synthetic anomalies allows us to test and refine the detection system continuously. By creating various scenarios that represent potential citation anomalies, we can evaluate the system′s responsiveness and accuracy, ensuring it remains effective in identifying suspected patterns. By leveraging these synthetic anomalies, the system becomes more adept at recognizing suspected citations. For instance, if a citation suddenly includes references from an unrelated field or anomalous patterns not seen in normal data, the system can flag these as potential anomalies. This approach ensures the model is finely tuned to detect even subtle irregularities, making the anomaly detection process more robust and accurate.

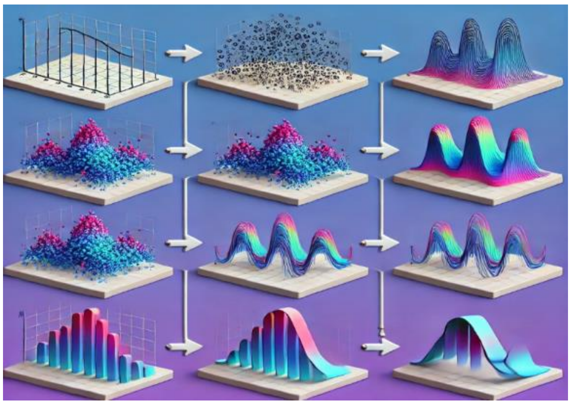


Figure 7: The generative process is visualized as a sequence of transformations. Initial random noise is iteratively refined through various stages to produce a structured, realistic distribution, illustrating the principle of diffusion models in generating high-quality data representations.

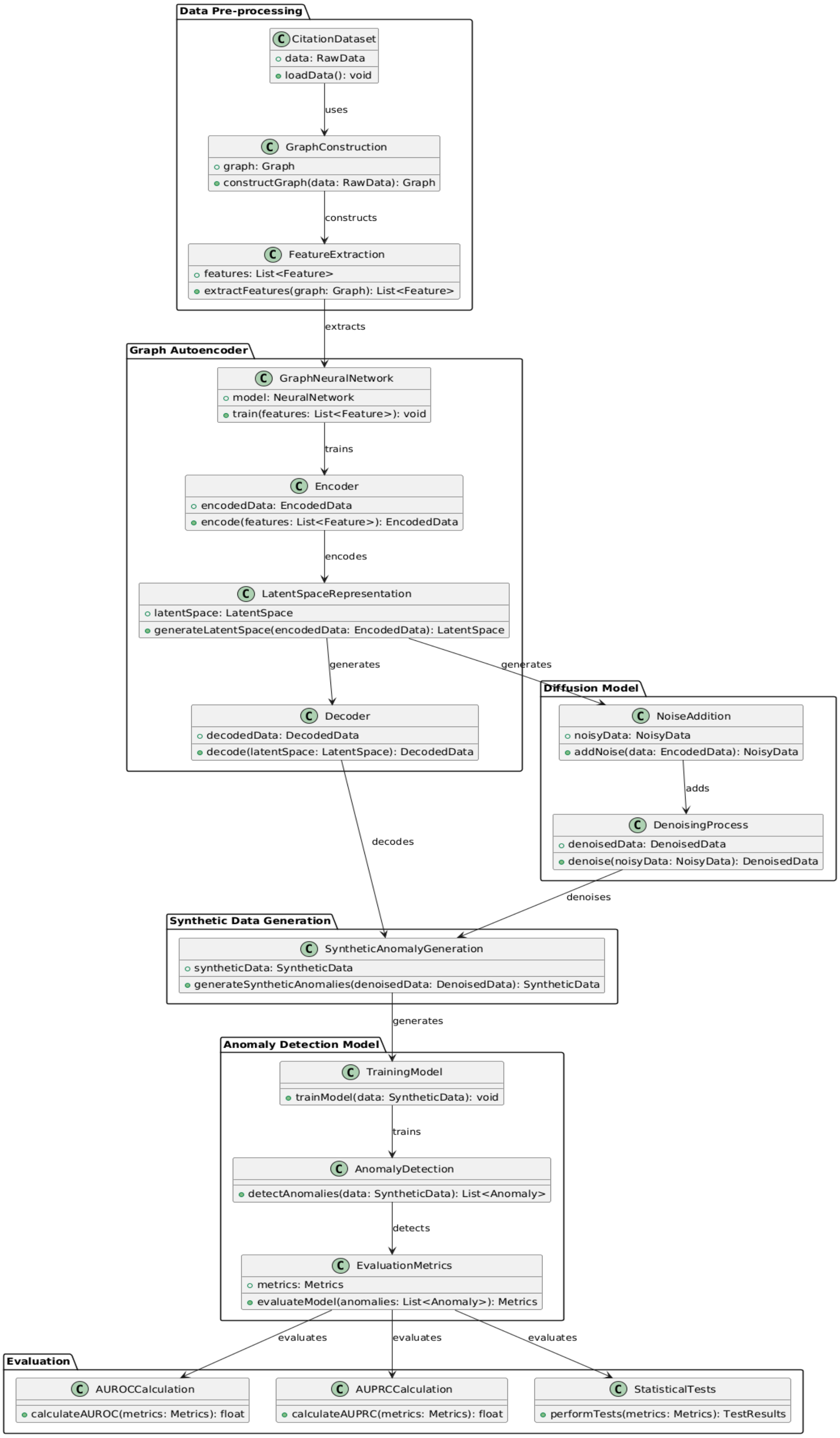

Data Pre-processing
CitationDataset
data: RawData
loadData(): void
uses
GraphConstruction
graph: Graph
constructGraph(data: RawData): Graph
constructs
FeatureExtraction
features: List<Feature>
extractFeatures(graph: Graph): List<Feature>
extracts
Graph Autoencoder
GraphNeuralNetwork
model: NeuralNetwork
train(features: List<Feature>): void
trains
Encoder
encodedData: EncodedData
encode(features: List<Feature>): EncodedData
encodes
LatentSpaceRepresentation
latentSpace: LatentSpace
generateLatentSpace(encodedData: EncodedData): LatentSpace
generates
generates
Decoder
decodedData: DecodedData
decode(latentSpace: LatentSpace): DecodedData
Diffusion Model
NoiseAddition
noisyData: NoisyData
addNoise(data: EncodedData): NoisyData
adds
DenoisingProcess
denoisedData: DenoisedData
denoise(noisyData: NoisyData): DenoisedData
decodes
denoises
Synthetic Data Generation
SyntheticAnomalyGeneration
syntheticData: SyntheticData
generateSyntheticAnomalies(denoisedData: DenoisedData): SyntheticData
generates
Anomaly Detection Model
TrainingModel
trainModel(data: SyntheticData): void
trains
AnomalyDetection
detectAnomalies(data: SyntheticData): List<Anomaly>
detects
EvaluationMetrics
metrics: Metrics
evaluateModel(anomalies: List<Anomaly>): Metrics
evaluates
evaluates
evaluates
Evaluation
AUROCCalculation
calculateAUROC(metrics: Metrics): float
AUPRCCalculation
calculateAUPRC(metrics: Metrics): float
StatisticalTests
performTests(metrics: Metrics): TestResults

Figure 8: The following class diagram represents the structure of our anomaly detection system. It highlights the main components involved in data preprocessing, graph autoencoding, diffusion modeling, synthetic data generation, anomaly detection, and evaluation.

4.2.2 Mathematical Formulation and Explanation

Latent Variables (z_0, z_1, ..., z_T)**:**
These are hidden variables that the model learns. z_0 is the original data point, and z_T is a highly noisy version of it.
Markov Chain:
A sequence of random variables (in this case, the latent variables) where each variable depends only on the previous one, starting from a standard Gaussian distribution [8].
$p(z_T) = N(z_T: 0, I)$.
Forwarding Diffusion Process:
This process adds noise to the data in steps, gradually making it more and more noisy.
Forward Process Definition:

$$q(z_1 : T|z_0) = \prod_{t=1}^{T} q(z_t|z_t - 1)$$

This means we're applying a series of steps to go from $z_0$to z_T.

Noise Addition:
The forward diffusion process, which incrementally introduces Gaussian noise into the data based on a variance schedule β_1, ..., β_T.

$$q(z_t|z_t - 1) = N(z_t; \sqrt{1 - \beta_t} z_t - 1, \beta_t I)$$

z_t is generated by adding noise to $z_t{-}1$. The noise level is controlled by βt.

Reversing Diffusion Process:
This process learns to reverse the noise addition, effectively generating new data points from noise. The process evolves with learned Gaussian transitions.
Reverse Process Definition:

$$p_\theta(z_0 : T) = p(z_T) \prod_{t=1}^{T} p_\theta(z_t - 1|z_t)$$

This means we start with z_t (pure noise) and step by step, generate less noisy versions until we get z_0.
Noise Removal:

$$p_\theta(z_t - 1|z_t) = N(z_t - 1; \mu_\theta(z_t, t), \sum\nolimits_\theta (z_t, t))$$

$z_t{-}1$ is generated by removing noise from z_t. The parameters μ_θ and Σ_θ are learned by the model.
Training:
The training objective of diffusion models involves optimizing the variational bound on the negative log likelihood: The goal is to train the model to be good at reversing the noise addition process. This is done by minimizing a loss function that measures how well the model can reverse the diffusion process.
Loss Function:

$$L = E[-\log p_\theta(z_0)] \leq E_q[-\log p(z_T) - \sum_{t \geq 1} \log \frac{p_\theta(z_t - 1|z_t)}{q(z_t|z_t - 1)}$$

This function measures the difference between the generated data and the real data. The goal is to minimize this difference.

Kullback-Leibler (KL) Divergence

This is a measure of how one probability distribution differs from another. In this context, it is used to compare the reverse process learned by the model with the actual forward process.

KL Divergence in Loss:

$$\mathrm{Eq}\{\ \mathrm{KL}[q(z_T|z_0)||p(z_T)] + \sum_{t>1} \mathrm{KL}[q(z_{t-1}|z_t, z_0)||p_\theta(z_{t-1}|z_t)] - \log p_\theta(z_0|z_1)\}$$

This breaks down the overall loss into components comparing different stages of the forward and reverse processes. This measures how well the model's reverse process matches the true posterior at each step. Each KL divergence term compares Gaussian distributions, which allows for efficient computation using closed-form expressions rather than high-variance Monte Carlo estimates. This optimization process aims to minimize the reconstruction error between the input data and the output from the reverse diffusion process.

Posterior Distributions

In the context of diffusion models, the posterior distributions help us to approximate the reverse process by considering the original data $z_0$and the noisy data z_T.

$$q(z_t - 1)|z_t, z_0) = N(z_t - 1; \mu_{\tilde{t}}(z_t, z_0), \beta_{\tilde{t}} I)$$

This means that the distribution of zt−1 given z_T and z0 is a Gaussian distribution with mean and variance .

$\mu_{\tilde{t}}(z_t, z_0)$ and variance $\beta_{\tilde{t}}$.

Mean of the Posterior ($\mu_{\tilde{t}}$)

The mean of this posterior distribution is given by:

$$\mu_{\tilde{t}}(z_t, z_0) = \sqrt{a_{\bar{t}-1}\beta_t} \backslash (1 - \alpha_{\bar{t}})z_0 + \sqrt{}\alpha_t((1 - \alpha_{\bar{t}-1})\ )\ /\ (1 - \alpha_{\bar{t}})\ z_t$$

$a_{\bar{t}-1}$ and $a_{\bar{t}}$ are cumulative products of α values up to t -1 and t, respectively $\alpha_t = 1 - \beta_t$ where β_t controls the amount of noise added at step t .

z_0 is the original data point and z_t is the noisy data at step t.

Variance of the Posterior ($\beta_{\tilde{t}}$)

The variance of this posterior distribution is given by:

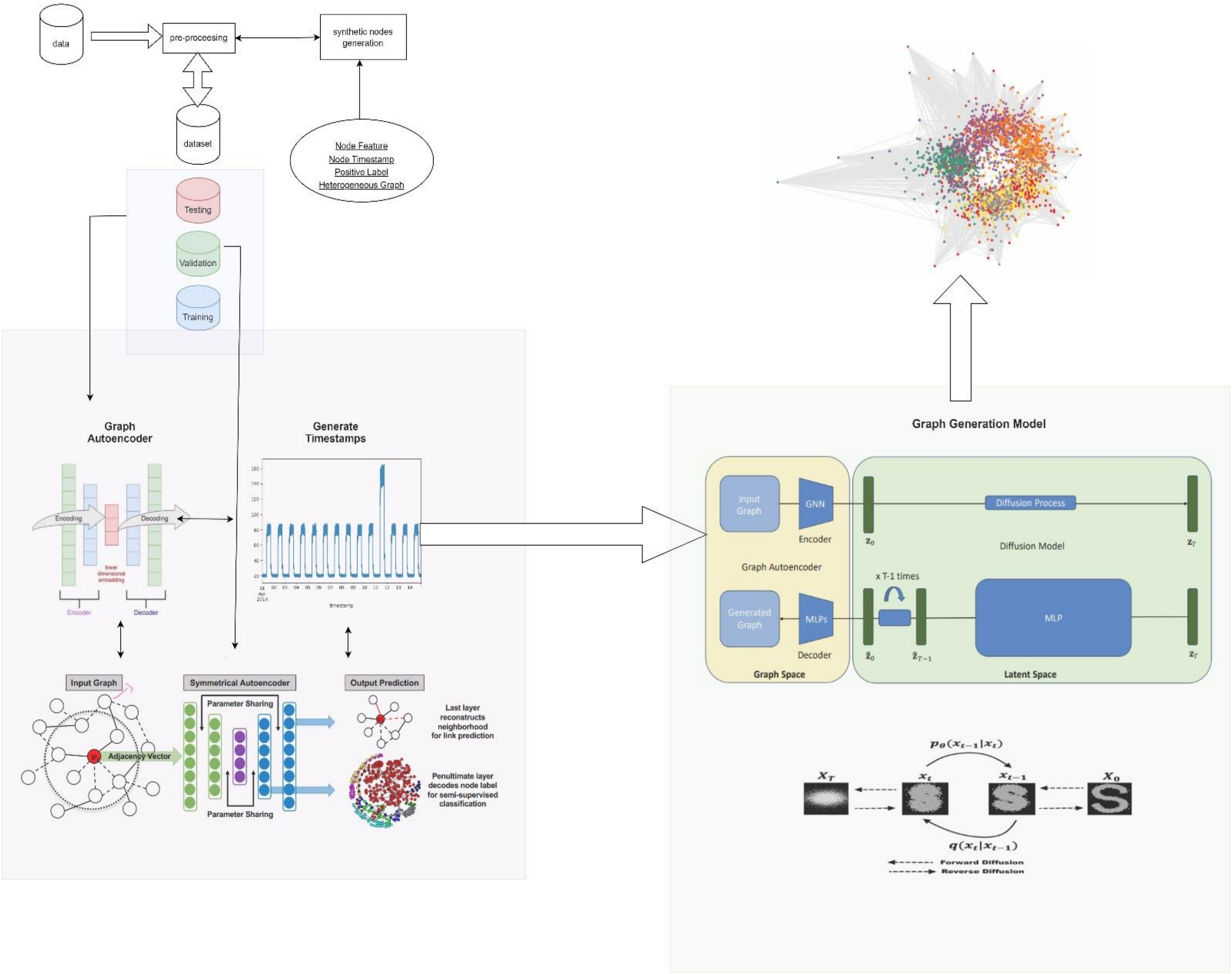
$$\beta_{\tilde{t}} = \frac{(1 - \alpha_{\bar{t}-1})}{(1 - \alpha_{\bar{t}-1})\beta_t}$$

This equation adjusts the noise level βt based on the cumulative noise up to steps t−1 and t.

5. Product

The product is an interface that will implement our research model, using a diffusion model and graph autoencoder for the generation of synthetic anomalies and the detection of anomalies in graph-structured data. Below is the flow chart of the research model.

Anomaly Detection Using Synthetic Nodes, Graph Autoencoders, and Diffusion Techniques



6. Testing and Evaluation

In evaluating the proposed graph diffusion model for anomaly detection, various metrics, and methods are utilized to gauge model performance. Metrics such as Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision Recall Curve, and (AUPRC) are employed to assess model efficiency, especially in handling label imbalance. AUROC measures the ability of the model to distinguish between normal and anomalous nodes, with a higher AUROC indicating better performance. AUPRC evaluates the precision-recall trade-off, which is crucial in scenarios with a high-class imbalance. For instance, the AUROC is calculated by plotting the true positive rate against the false positive rate at various threshold settings, offering a comprehensive view of the model's classification capabilities. A value closer to 1 indicates excellent performance, while a value around 0.5 suggests no better than random guessing. Statistical Tests, such as the t-test and chi-square test, can be employed to compare the performance of different models or to evaluate the significance of improvements observed. These tests help in determining whether the observed differences in performance metrics are statistically significant or due to random chance.

In this project, the diffusion model-based graph generator was evaluated against several baseline models, including GCN, GraphSAGE, GAT, and Graph Transformer. [7] The experimental results indicated that our model consistently outperformed these baselines across multiple datasets, such as Reddit, Weibo, Amazon, tFinance, and DGraph. For example, on the tFinance dataset, our model achieved an AUROC of 0.97, significantly higher than the best-performing baseline, GAT, which scored 0.94. Additionally, studies showed that the diffusion process is crucial for improving synthetic anomaly generation and detection performance. The model was trained on real data and tested on its ability to create realistic anomalies, which were then used to train anomaly detection systems. These generated anomalies balanced the dataset, addressing label imbalance and enhancing model robustness. Evaluations, including synthetic data generation and statistical tests, proved the model's effectiveness in generating high-quality synthetic anomalies, improving anomaly detection in graph-based data.

6.1 Testing plan

Test	Test Subject	Expected Result	Unit Test	Type
1	Graph Generation	The model will accurately generate graphs with both homogeneous and heterogeneous structures based on the input parameters.	TestGraphGeneration	JUnit (Java) / unittest (Python)
2	Anomaly Node Generation	The model will generate synthetic anomaly nodes with realistic features, timestamps, and positive labels.	TestAnomalyNodeGeneration	JUnit (Java) / unittest (Python)
3	Graph Processing	The model will correctly process the generated graphs, maintaining the integrity of both local and global relationships.	TestGraphProcessing	JUnit (Java) / unittest (Python)
4	Graph Output Validation	The model will produce valid output graphs that accurately represent the intended relationships and properties, including anomalies.	TestGraphOutputValidation	JUnit (Java) / unittest (Python)
5	Runtime Efficiency	The model will perform the graph generation and processing tasks within an acceptable time frame, demonstrating scalability.	TestRuntimeEfficiency	JUnit (Java) / unittest (Python)
6	Input Size/Graph Size	The model will handle larger graphs without significant performance degradation or memory issues.	TestInputSizeHandling	JUnit (Java) / unittest (Python)
7	Empty Graphs	If the input graph is empty, the model will return a specific error message indicating that the input graph cannot be empty and halt further processing.	TestEmptyGraphHandling	JUnit (Java) / unittest (Python)
8	Incomplete Graphs	If the input graph is incomplete (missing nodes or edges), the model will either return an error message or automatically fill in the missing information based on predefined rules or assumptions.	TestIncompleteGraphHandling	JUnit (Java) / unittest (Python)
9	Incorrect Graphs	If the input graph contains invalid nodes or edges, the model will return an error message indicating the nature of the error and halt further processing until a valid graph is provided.	TestIncorrectGraphHandling	JUnit (Java) / unittest (Python)
10	Incompatible Graphs	If the input graph is incompatible with the model (e.g., too large or unsupported structure), the model will return an error message indicating the incompatibility and halt further processing until a compatible graph is provided.	TestIncompatibleGraphHandling	JUnit (Java) / unittest (Python)

11	Generation Similarity Index	The generated graphs will achieve a similarity index greater than 85% compared to the real graphs, ensuring high fidelity in synthetic anomaly generation.	TestGenerationSimilarityIndex	JUnit (Java) / unittest (Python)
12	Cross-Validation Performance	The model will maintain consistent performance across different folds of cross-validation, demonstrating robustness and generalizability.	TestCrossValidationPerformance	JUnit (Java) / unittest (Python)
13	Synthetic Data Quality	The synthetic anomaly data will closely match the distribution of real anomaly data, validated using metrics like Jensen-Shannon Divergence.	TestSyntheticDataQuality	JUnit (Java) / unittest (Python)
14	Statistical Significance	Improvements in model performance will be statistically significant, validated through appropriate statistical tests like t-tests or chi-square tests.	TestStatisticalSignificance	t-test / Chi-square test
15	Label Imbalance Handling	The model will effectively address label imbalance, ensuring that anomaly detection performance does not degrade due to the imbalance.	TestLabelImbalanceHandling	JUnit (Java) / unittest (Python
16	Anomaly Detection Accuracy	The unit test will validate that the anomaly detection algorithm correctly identifies anomalies in a test graph, achieving a detection accuracy greater than 90%.	TestAnomalyDetectionAccuracy	JUnit (Java) / unittest (Python)

References

[1] Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2018). Graph Neural Networks: A Review of Methods and Applications. arXiv preprint arXiv: 1812.08434.

[2] Kipf, T. N., & Welling, M. (2016). Semi-Supervised Classification with Graph Convolutional Networks. arXiv preprint arXiv: 1609.02907.

[3] Hamilton, W. L., Ying, Z., & Leskovec, J. (2017). Inductive Representation Learning on Large Graphs. Advances in Neural Information Processing Systems (NeurIPS), 30, 1024-1034. Link to paper.

[4] Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph Attention Networks. arXiv preprint arXiv: 1710.10903.

[5] Grover, A., & Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 855-864. Link to paper.

[6] Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk: Online Learning of Social Representations. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 701-710. Link to paper.

[7] Wang, D., Cui, P., & Zhu, W. (2016). Structural Deep Network Embedding. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1225-1234. Link to paper.

[8] Xu, W., Liu, X., & Gong, Y. (2003). Document Clustering Based on Non-negative Matrix Factorization. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 267-273. Link to paper.

[9] Tran, D. T., Nguyen, H., & Phung, D. (2021). Graph Neural Networks for Anomaly Detection in Dynamic Graphs. IEEE Transactions on Neural Networks and Learning Systems, 32(10), 4533-4546. Link to paper.

[10] You, Y., Ying, Z., & Leskovec, J. (2020). Design Space for Graph Neural Networks. Advances in Neural Information Processing Systems (NeurIPS), 33, 17009-17021. Link to paper.

[11] Salha-Galvan, S., Hennequin, R., & Vazirgiannis, M. (2019). Simple Spectral Graph Convolution. arXiv preprint arXiv: 1902.07153.

[12] Wang, X., Yu, R., Zheng, K., & Sun, Y. (2019). Dynamic Network Embedding: An Extended Approach for Skip-gram based Network Embedding. Proceedings of the 28th International Joint Conference on Artificial Intelligence, 1427-1433. Link to paper.

[13] Xu, X., Yuruk, N., Feng, Z., & Schweiger, T. A. J. (2007). Scan: a structural clustering algorithm for networks. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 824-833. Link to paper.

[14] You, J., Ying, R., Ren, X., Hamilton, W., & Leskovec, J. (2018). Graphrnn: Generating realistic graphs with deep auto-regressive models. International Conference on Machine Learning. PMLR, 5708-5717. Link to paper.

[15] Kipf, T. N., & Welling, M. (2016). Variational graph auto-encoders. arXiv preprint arXiv: 1611.07308.

[16] Hu, Z., Dong, Y., Wang, K., & Sun, Y. (2020). Heterogeneous graph transformer. Proceedings of the Web Conference 2020, 2704-2710. Link to paper.

[17] Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv: 1609.02907.