

Identifying Data Quality Issues in Clinical Trials Using OpenAI Embeddings

Prepared by:
Ashwin Rai
Director of Data Science & Analytics
Evidera – PPD/ThermoFisher Scientific

Prepared for:
PHUSE Innovation Challenge 2024

February 27, 2024



Risk-based Monitoring



Risk-based monitoring (RBM) is a clinical trial monitoring approach that focuses on identifying and managing the risks associated with a study, with a particular emphasis on patient safety, data integrity, and the overall success of the trial.

Unlike traditional on-site monitoring, which involves frequent site visits, RBM uses a more targeted and risk-driven strategy:

1

Risk identification: Identification of potential risks to data quality and patient safety.

2

Centralized monitoring: Increased reliance on centralized monitoring using technology and data analytics.

3

Site focus: Monitoring activities tailored to the specific risks and complexities of each clinical trial site.

4

Data-driven approach: Use of data-driven methods to identify trends, outliers, and potential issues.



Importance of Data Quality in RBM



Patient safety:

The use of high-quality data ensures that patient safety is maintained throughout the trial by providing accurate information on adverse events (AEs) and patient responses to interventions.



Reliability of outcomes:

Reliable data are essential for the generation of accurate and meaningful outcomes. Ensuring data quality contributes to the credibility of trial results.



Regulatory compliance:

Regulatory bodies place a strong emphasis on data integrity. High-quality data are necessary for compliance with regulatory requirements and standards.



Risk mitigation:

Effective RBM relies on high-quality data for the identification of risks and the implementation of appropriate mitigation strategies.



Efficiency and cost savings:

High-quality data allow for more efficient monitoring, reducing the need for extensive on-site visits. This can result in cost savings and more streamlined trial processes.



Decision-making:

Stakeholders, including sponsors and investigators, rely on high-quality data for informed decision-making throughout the trial.



RBM for Outlier Detection

Outliers in clinical trial data represent deviations from the norm, and identifying and managing these outliers is crucial for ensuring data quality, patient safety, and the success of medical interventions.



Protocol deviations: Any significant deviations from the study protocol can be considered outliers. This includes deviations in the administration of investigational products, adherence to visit schedules, and protocol-specific procedures.



Subject enrollment discrepancies: Discrepancies in subject enrollment numbers or patterns across study sites can be indicative of potential issues. Unexpected variations in enrollment rates may signal problems in participant recruitment or retention.



AEs: Outliers in the frequency or severity of AEs may raise concerns. Unusual spikes in reported AEs or unexpected patterns can be considered outliers, necessitating an investigation into potential safety issues.



Data entry errors: Data entry errors, especially those that are inconsistent or frequent, can be flagged as outliers. These errors may include inaccuracies in patient demographics, medical histories, or other critical data points.



Monitoring visit findings: Findings during monitoring visits that deviate significantly from expected norms can be treated as outliers. This includes unexpected discrepancies in source data verification or deviations from the monitoring plan.



RBM for Outlier Detection

Outliers in clinical trial data represent deviations from the norm, and identifying and managing these outliers is crucial for ensuring data quality, patient safety, and the success of medical interventions.



Site performance disparities: Disparities in site performance metrics, such as data quality, adherence to the protocol, or completion of required documentation, can be considered outliers. Early identification of poorly performing sites is crucial for risk mitigation.



Data trends and patterns: Unusual trends or patterns in the collected data, such as unexpected variations in efficacy or safety outcomes, may be flagged as outliers. These patterns may indicate underlying issues that need to be investigated.



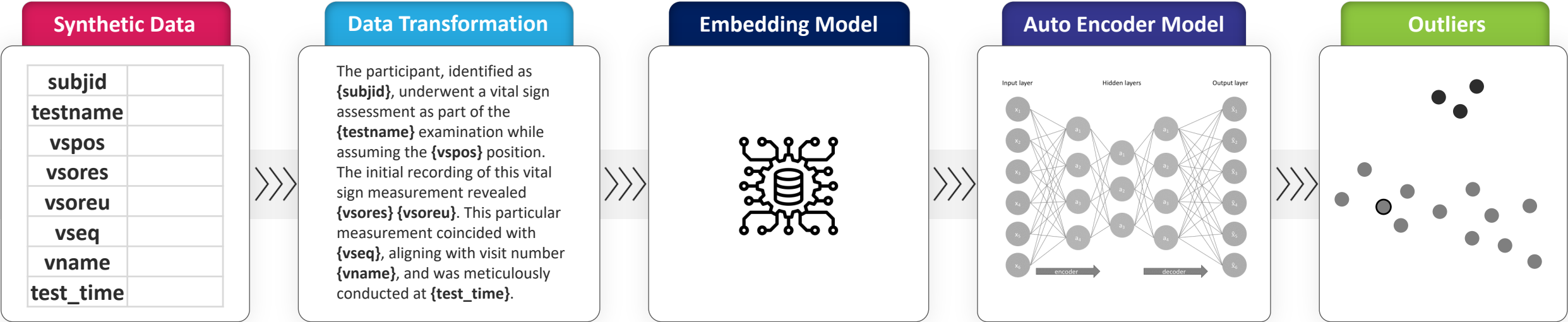
Noncompliance with regulatory requirements: Incidences of noncompliance with regulatory requirements, including deviations from Good Clinical Practice guidelines, may signal potential risks. Addressing these outliers is vital for maintaining regulatory compliance.



Monitoring alerts and triggers: RBM incorporates predefined monitoring alerts and triggers based on statistical algorithms and risk indicators. Events that trigger these alerts, such as unexpected deviations in data trends or safety signals, are considered outliers.



Site-specific challenges: Site-specific challenges, such as unexpected difficulties in patient recruitment, retention, or data collection, can be considered outliers. Identifying and addressing these challenges is essential for optimizing site performance.





subjid	
testname	
vspos	
vsores	
vsoreu	
vseq	
vname	
test_time	



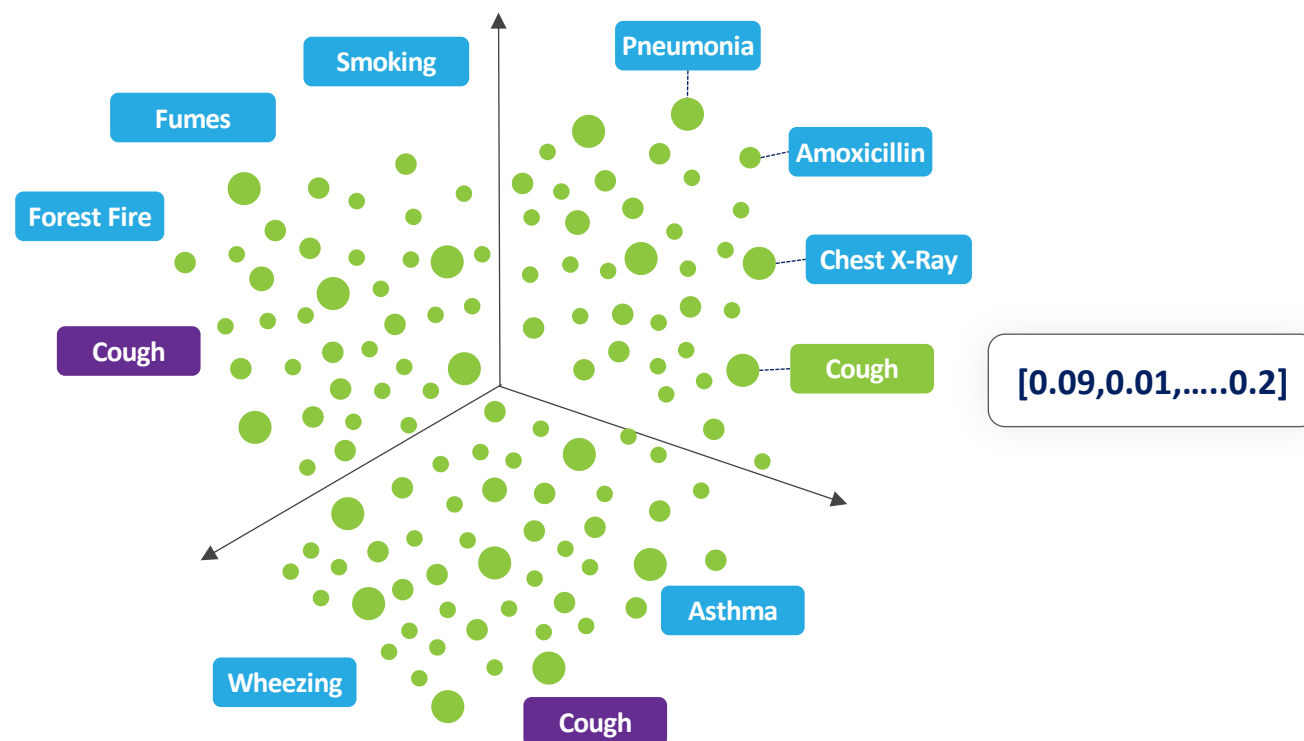
The participant, identified as {subjid}, underwent a vital sign assessment as part of the {testname} examination while assuming the {vspos} position. The initial recording of this vital sign measurement revealed {vsores} {vsoreu}. This particular measurement coincided with {vseq}, aligning with visit number {vname}, and was meticulously conducted at {test_time}.

- Real-world data challenges can be simulated with synthetic data to ensure the methodology is robust enough to handle complex clinical trial datasets.
- This example uses synthetic tabular data consisting of 20,000 data points, including 1,000 outlier data points. The dataset encompasses six variables—vspos, vsores, vsoreu, vseq, vname, and testtime—representing six respective tests: blood pressure, temperature, oxygen saturation, heart rate, pulse, and respiratory rate.
- The tabular clinical trial data can then be transformed into a text-like format, setting the stage for the encoding and embedding process.
- This example vividly illustrates how narrative elements such as test conditions, positions, and visit details seamlessly convert into textual elements, enriching the linguistic depth of the dataset.



Embedding

The patient presented with symptoms of **pneumonia**, including a persistent **cough**, and was prescribed a course of **amoxicillin**. A **chest X-ray** was recommended to assess the extent of lung involvement.



- Embedding vectors are numerical representations of words or entities in a multidimensional space, used to capture specific semantic or contextual information.
- Each dimension in an embedding vector represents a specific feature or aspect learned during training.
- Embeddings enable the measurement of semantic relatedness between words, reflecting their similarity in meaning or context.
- Embeddings capture contextual nuances to help models understand word usage variations.
- Pretrained embeddings can transfer knowledge across domains, enhancing performance with limited data.
- Embedding vector assist machine learning models and algorithms in understanding content relationships and performing tasks like clustering or retrieval.

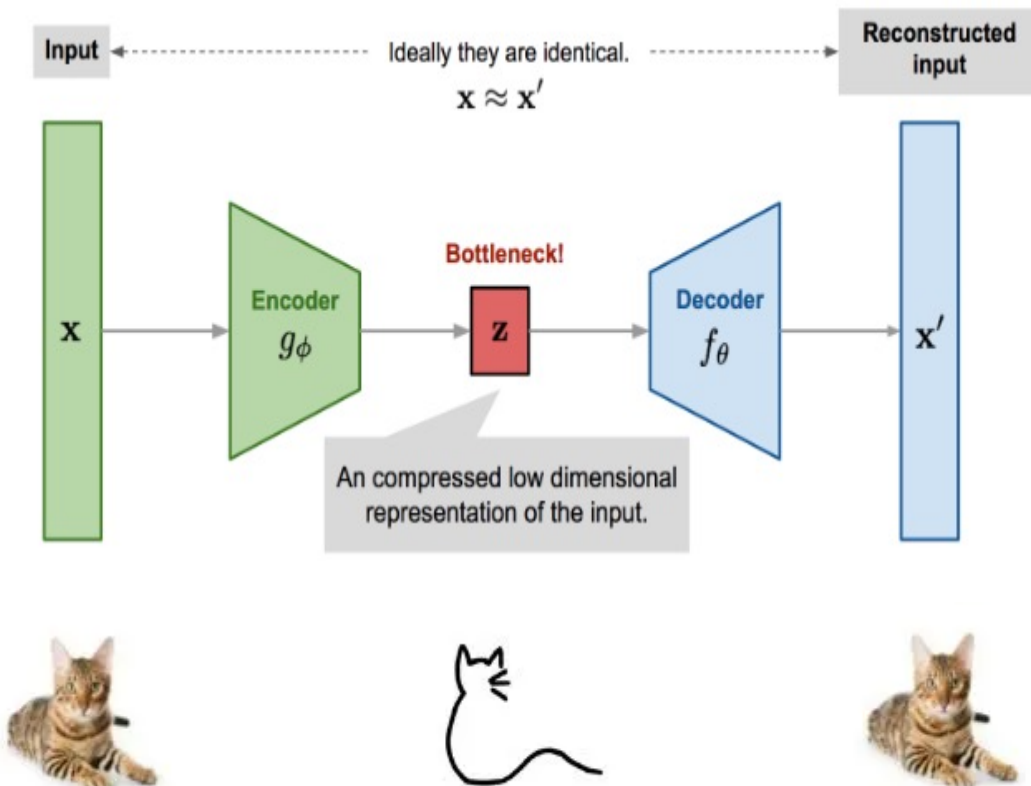


The participant, identified as {subjid}, underwent a vital sign assessment as part of the {testname} examination while assuming the {vspos} position. The initial recording of this vital sign measurement revealed {vsoreu} {vsoreu}. This particular measurement coincided with {vseq}, aligning with visit number {vname}, and was meticulously conducted at {test_time}.





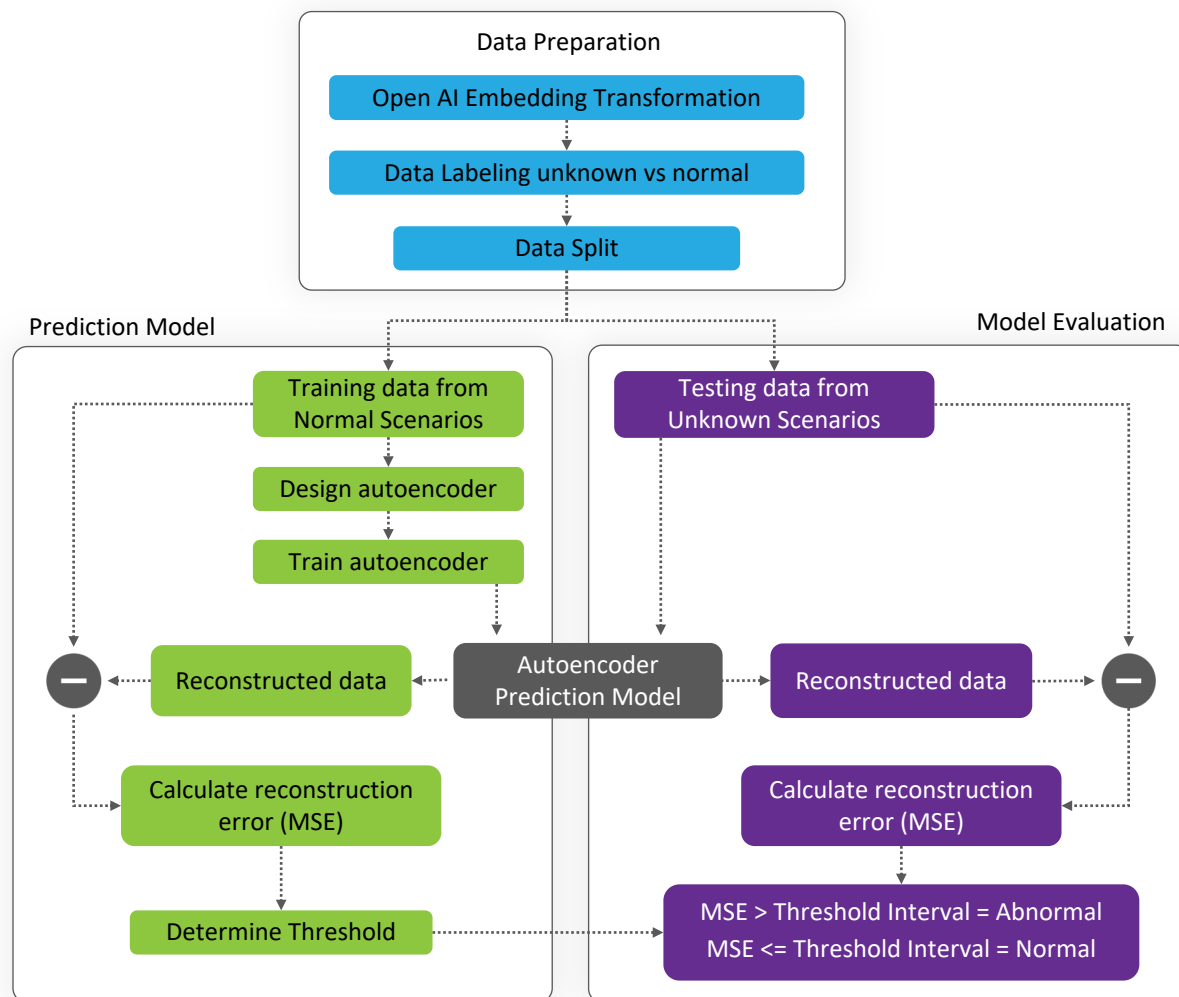
Auto Encoder



- An autoencoder is a neural network for unsupervised learning, consisting of an encoder to compress input data into a lower-dimensional representation and a decoder to reconstruct the input.
- Autoencoder minimizes the difference between input and reconstructed output and the lower-dimensional representation retained by the autoencoder encapsulates essential features, rendering it valuable for detecting outliers.
- Instances deviating significantly from this representation may indicate anomalies.
- During training, the model learns to accurately reconstruct normal instances, while outliers lead to a higher rate of reconstruction errors.
- The loss function quantifies this difference, with anomalies contributing to higher loss, aiding in their identification.

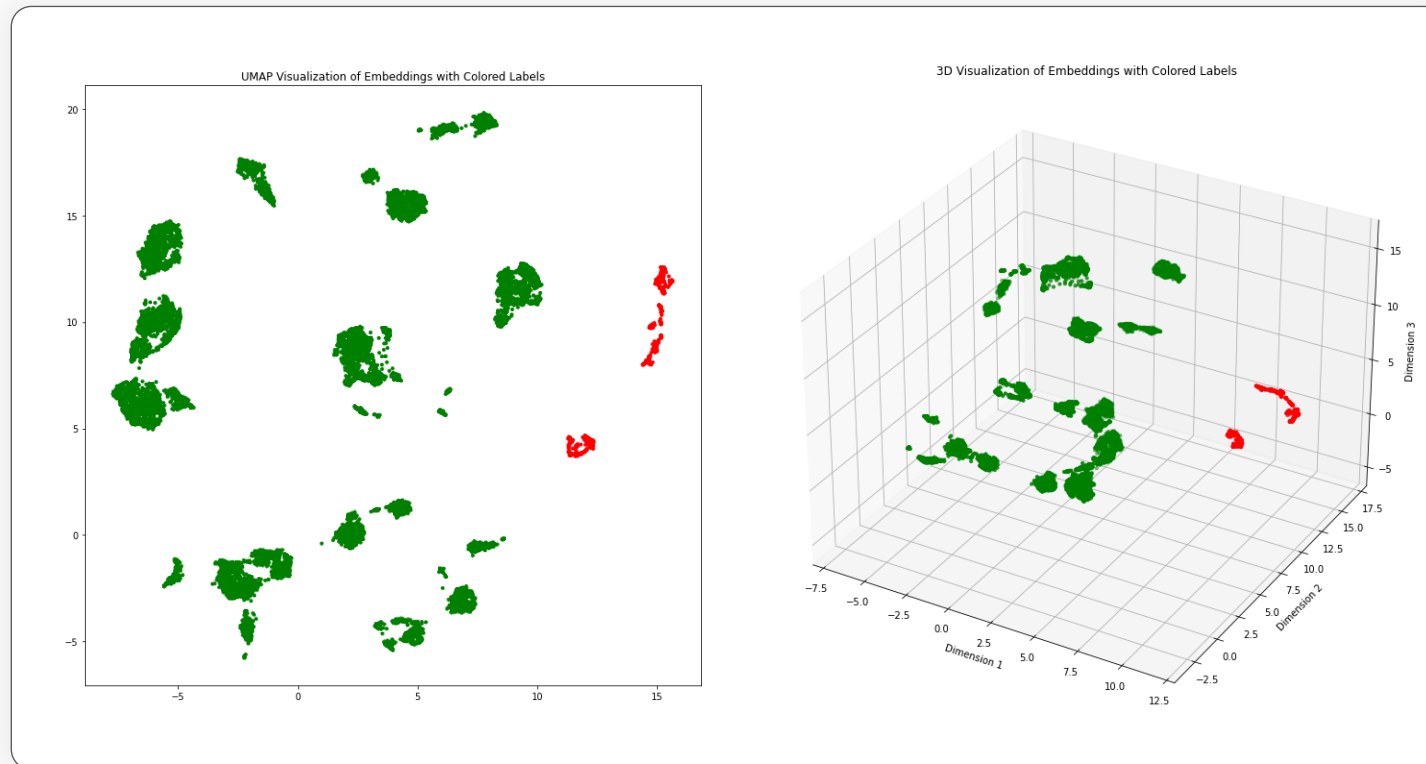


Outlier Detection Architecture



Abbreviations: AI = artificial intelligence; MSE = mean squared error

1. **Data preprocessing:** The dataset is split into training and test sets, consisting of embeddings.
2. **Training the autoencoder:** The autoencoder is trained unsupervised on normal, nonoutlier data points. It learns to compress input data into a lower-dimensional representation using an encoder, and to reconstruct the original input using a decoder.
3. **Calculating reconstruction error:** After training, the autoencoder reconstructs all data points and the difference between original and reconstructed data is measured using metrics such as mean squared error or binary cross-entropy.
4. **Threshold selection:** A threshold for reconstruction error is chosen based on dataset characteristics and the desired sensitivity to outliers.
5. **Identifying outliers:** Data points with reconstruction errors exceeding the threshold are labeled as outliers, indicating significant deviation from the normal patterns learned during training.

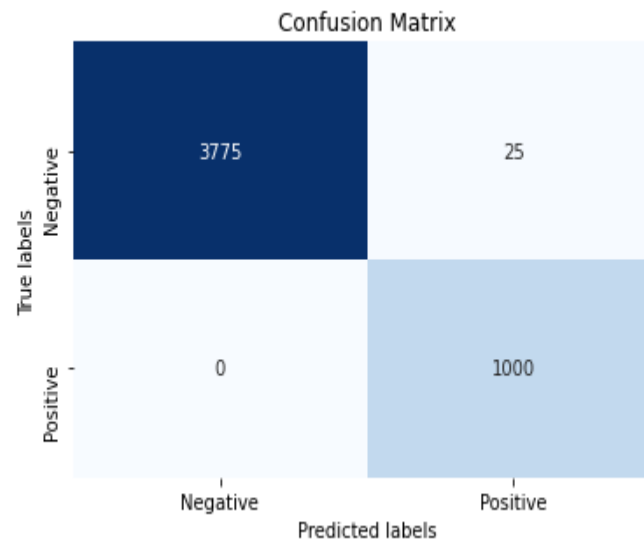
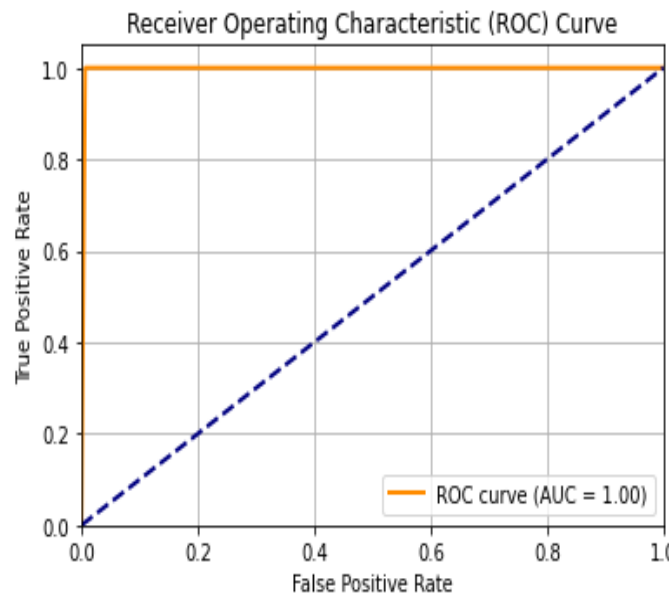


Abbreviations: UMAP = uniform manifold approximation and projection

- The detected outliers can be visualized and analyzed to gain insights into the anomalies present in the dataset.
- Visualization techniques, such as scatter plots or dimensionality reduction methods, can depict the distribution of normal and outlier data points.
- In the context of outlier visualization, uniform manifold approximation and projection (UMAP) is a valuable tool for identifying anomalies by projecting the data into a two- or three-dimensional space where outliers may be visually distinct from the rest of the data points.
- By preserving the local structure of the data, UMAP can reveal clusters and outliers that may not be apparent in the original high-dimensional space, providing insights into the distribution and relationships within the dataset.



Results



- The performance assessment of the autoencoder model further validated the robustness of this methodology, which demonstrated exceptional accuracy when applied to a test dataset comprising 4,800 records, 3,800 normal data points, and 1,536 embedding dimensions, including 1,000 deliberately introduced outlier data points.
- The methodology achieved an accuracy of 0.9948, with a true positive rate (recall) of 1 and a positive predictive value (precision) of 0.9756, highlighting its ability to successfully identify and recall the majority of outlier data points. Additionally, the impressively low false-positive rate of 0.0064 underscored the methodology's precision in minimizing false positives, enhancing its reliability for practical applications in healthcare data analysis.



Conclusion



The integration of OpenAI embeddings and autoencoder models presents a promising approach for outlier detection by leveraging semantic meaning and data reconstruction capabilities.



The proposed approach is unsupervised, allowing it to learn meaningful representations directly from the data. This enables effective detection of outliers, anomalies, or unusual patterns without the need for labeled training data, making it adaptable to diverse domains and robust to noise. This stands in contrast to traditional methods, which rely on predefined features.



The ability to reconstruct data from compressed representations enables a deeper understanding of outlier instances. However, the low-dimensional representations produced by this approach can be challenging to interpret compared with traditional feature-based methods.



Despite this limitation, the robustness and versatility of this approach make it suitable for identifying outliers across various fields and industries, especially when the requirement is to determine if it is an outlier rather than how it qualifies to be an outlier.



Further Work



Adapt the model to your own data.



Use Clinical embeddings such as Clinical Bert, Med-Bert, Bio-Bert instead of Open AI embeddings



OpenAI embeddings offer versatility across various domains. Retraining OpenAI embeddings on healthcare data will enable organizations to develop specialized language models that better capture the nuances and complexities of medical terminology, thereby improving performance on healthcare-specific tasks.



Code



`openai_autoencoder_outlier_detection_sample_code.py`



`openai_autoencoder_outlier_detection_sample_code.ipynb`



`openai_autoencoder_outlier_detection_sample_code.html`

Q&A