**Detecting Pay Outliers in the United States Labor Force with Context-Aware Attention-Enhanced Latent Diffusion Models**

by Juan R. Falck

B.S. in Computer Science, May 1992, United States Naval Academy
MBA, December 2002, INSEAD
M.S. in Data Science, December 2023, City University of New York

A Praxis submitted to

The Faculty of
The School of Engineering and Applied Science
of George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Engineering

January 9, 2026

Praxis directed by

Muhammad F. Islam
Professorial Lecturer in Engineering and Applied Science

The School of Engineering and Applied Science of The George Washington University certifies that Juan R. Falck has passed the final examination for the degree of Doctor of Engineering as of November 6, 2025. This is the final and approved form of the Praxis.

**Detecting Pay Outliers in the United States Labor Force with Context-Aware Attention-Enhanced Latent Diffusion Models**

Juan R. Falck

Praxis Research Committee:

Amir Etemadi, Adjunct Professor of Engineering and Applied Science, Praxis Chair

Muhammad F. Islam, Professorial Lecturer in Engineering and Applied Science, Praxis Advisor

Jeffrey Yu, Professorial Lecturer in Engineering and Applied Science, Committee Member

© Copyright 2025 by Juan R. Falck
All rights reserved.

## Dedication

I would first like to thank the most important people in my life: Saby, Daniella, and Megan, my incredible wife and two daughters. Your support was the driving force pushing me through the finish line. I will forever thank you for being part of my life.

To Liceo Militar del Norte in Honduras, which early in my life showed me the values of teamwork and hard work.

To the U.S. Naval Academy, for providing me with the discipline needed to succeed under pressure and navigate through difficult challenges and for being where I started on this life journey in computer science and data analytics.

To INSEAD, for broadening my horizon and introducing me to global business and collaboration.

To the City University of New York, for providing a challenging data science and machine learning curriculum, which pushed me to the next level in my knowledge of artificial intelligence.

Finally, to The George Washington University, for giving me the tools I needed throughout this arduous but rewarding process.

To every academic institution in my life, I can only extend my sincere gratitude as each has contributed to this journey.

## Acknowledgments

I would like to extend my deepest gratitude to Dr. Muhammad Islam, his mentorship and sincere interest in my research was pivotal for the completion of this work.

## Abstract of Praxis

### Detecting Pay Outliers in the United States Labor Force with Context-Aware Attention-Enhanced Latent Diffusion Models

Outlier or anomaly detection is a critical process in many industries. It is an important topic of research in artificial intelligence with applications in fields like fraud detection, cyber-security, and many others.

This research focuses on the process of anomaly detection using unsupervised methods (no labels used or available). Primary goal focuses on anomaly detection for compensation but rapidly expands into general anomaly detection.

For the initial goal of anomaly detection in compensation, it is recognized as a critical part of any operating businesses as compensation is closely regulated internally and externally. Early accurate detection becomes crucial to identifying potential underlying issues.

Traditional methods used like density estimation, inter-quartiles ranges, or simple fixed rules suffer from inefficiencies flagging outliers and generating a larger number of false positives and negatives (He et al., 2005; Pang, Cao, et al., 2021).

Most data which need to be reviewed for potential anomalies comes unlabeled or labelling would be too costly or impossible to do within the allowed timeframes, thus there is a need for accurate unsupervised anomaly detection systems.

This research introduces **CALDM**, which stands for **Context-Aware Latent Diffusion Model**. This method is an unsupervised method for anomaly detection (no labels used for training). This method offers advantages in cost and speed. Also, the introduction of attention mechanisms adds a contextual dimension to the anomaly

detection process, which is critical as context plays a fundamental role in determining if a sample is anomalous or not.

CALDM was tested for its anomaly detection role in compensation using a dataset from the Bureau of Labor Statistics (BLS). For the wider-scope and more general anomaly detection role, CALDM was tested using a publicly available group of datasets used for benchmarking outlier detection systems called ADBench (Han et al., 2022). These datasets were used to compare CALDM against a comprehensive list of commonly used unsupervised anomaly detection algorithms available publicly.

The test results showed that CALDM outperformed all other selected unsupervised anomaly detection algorithms. Its ability to add a contextual dimension to the process gave CALDM a leading edge over other systems tested. These results position CALDM as powerful general-purpose algorithm across industries outside the primary and original focus which was compensation.

## Table of Contents

Dedication ..... iv

Acknowledgments ..... v

Abstract of Praxis ..... vi

Table of Contents ..... viii

List of Figures ..... xiii

List of Tables ..... xiv

List of Equations ..... xv

List of Symbols ..... xvi

List of Abbreviations and Acronyms ..... xvii

Chapter 1—Introduction ..... 1

1.1 Background ..... 1

1.2 Research Motivation ..... 6

1.3 Problem Statement ..... 8

1.4 Thesis Statement ..... 9

1.5 Research Objectives ..... 9

1.6 Research Questions and Hypotheses ..... 10

1.7 Scope of Research ..... 11

1.8 Research Limitations ..... 12

1.9 Organization of Praxis ..... 13

Chapter 2—Literature Review ..... 15

2.1 Introduction ..... 15

2.2 Understanding Outliers: Definitions And Types ..... 15
2.2.1 Distinguishing Semantic Anomalies from Data Noise ..... 16
2.3 Outlier Detection in Tabular Data ..... 17
2.3.1 Taxonomy of Unsupervised Outlier Detection Methods ..... 17
2.3.2 Context and Motivation ..... 21
2.4 Generative Models for Tabular Data ..... 21
2.5 Latent Diffusion Models ..... 22
2.6 Transformer-Based Architectures ..... 23
2.7 Hybrid Models: VAE + Diffusion + Transformer ..... 24
2.8 Applications for Compensation Analytics and HR Data ..... 25
2.9 Summary and Conclusion ..... 26
Chapter 3—Methodology ..... 28
3.1 Introduction ..... 28
3.2 Summary of Methodology ..... 29
3.3 Datasets ..... 31
3.3.1 Bureau of Labor Statistics (BLS) Dataset ..... 31
3.3.1.1 General Characteristics of OEWS dataset ..... 32
3.3.1.2 Generation of individual employees from aggregate data ..... 33
3.3.1.3 Assignment of Performance Scores to each synthetically generated employee ..... 34
3.3.1.4 Assignment of Outlier labels to BLS data ..... 35
3.3.1.5 Synthetic MSAs for rare category testing ..... 36
3.3.1.6 Final characteristics of complete synthetic dataset ..... 37

3.3.1.7 Characteristics of Sub-Set used for Training and Testing ..... 39
3.3.1.8 Exploratory Data Analysis (Synthetic data derived from BLS aggregate data) ..... 39
3.3.2 ADBench Public Outlier Detection Datasets ..... 43
3.4 Model Development ..... 44
3.4.1 Baseline Models for Comparison ..... 45
3.4.2 CALDM – Architecture And Other Design Considerations ..... 46
3.4.2.1 Variable Auto-Encoder ..... 48
3.4.2.2 Diffusion Model ..... 50
3.4.2.3 Context Encoder ..... 51
3.4.2.4 CALDM Algorithm ..... 54
3.4.3 Advantages over a VAE model for outlier detection ..... 56
3.4.3.1 Limitations of VAEs for Outlier Detection ..... 56
3.4.3.2 The Benefits of Latent Diffusion Models ..... 57
3.4.4 Advantages of Latent Diffusion Models over pure Diffusion models ..... 57
3.4.4.1 Limitations of DDPMs ..... 58
3.4.4.2 Latent Diffusion Models' Compensation for the Limitations of Diffusion Models ..... 58
3.4.5 Conclusion of the advantages of a Latent Diffusion Model ..... 58
3.5 Experimental Setup ..... 59
3.6 Evaluation Metrics ..... 60
3.6.1 Average Precision (AP) ..... 60
3.6.2 Area Under the Receiver Operating Characteristic Curve ..... 61

3.6.3 Other Metrics Used ..... 62
3.7 Research Questions Operationalization ..... 62
3.8 Ethical Considerations ..... 63
3.9 Limitations of Research ..... 64
3.10 Summary ..... 65
Chapter 4—Results ..... 67
4.1 Introduction ..... 67
4.2 Datasets Overview ..... 67
4.2.1 Synthetic BLS Compensation Dataset ..... 67
4.2.2 Public Outlier Detection Datasets ..... 68
4.3 Experimental Results ..... 68
4.3.1 Hypothesis 1 ..... 69
4.3.1.1 Outlier Detection Datasets ..... 70
4.3.1.2 Synthetic BLS Dataset ..... 74
4.3.2 Hypothesis 2 ..... 79
4.3.3 Hypothesis 3 ..... 80
4.3.3.1 Summary of Rare Category Problem ..... 81
4.3.3.2 Results of Rare Category Test ..... 82
4.4 Characteristics of the Latent Space generated by CALDM ..... 84
4.5 Summary of Key Results ..... 86
Chapter 5—Discussion and Conclusions ..... 87
5.1 Discussion ..... 87
5.1.1 Alignment with Research Questions ..... 87

5.1.2 Interpretation of Key Findings ..... 88
5.1.3 Contrast with Traditional Methods ..... 90
5.1.4 Practical Implications ..... 91
5.1.5 Limitations and Considerations ..... 93
5.2 Conclusions ..... 94
5.3 Contributions to Body of Knowledge ..... 95
5.4 Recommendations For Future Research ..... 96
References ..... 100
Appendix A ..... 117
Appendix B ..... 120
Appendix C ..... 122
Appendix D ..... 123
Appendix E ..... 124
Appendix F ..... 125

## List of Figures

Figure 1- Example of raw data from BLS dataset ..... 33

Figure 2 - Example distribution of synthetically generated employees Wyoming..... 34

Figure 3 - Theoretical performance distribution (1–5) for given salaries..... 35

Figure 4- U.S. States by Average Salary and Employed Population..... 39

Figure 5- Boxplot of States by MEDIAN Salary..... 40

Figure 6- Density Plots of outlier and non-outlier population..... 41

Figure 7 - Average Salary by Performance Level ..... 42

Figure 8 - Overall architecture and data flow of CALDM. ..... 48

Figure 9 - CALDM Algorithm..... 54

Figure 10 - BLS dataset test - confusion matrix ..... 76

Figure 11 - AUROC for BLS dataset test ..... 77

Figure 12– Reconstruction Loss Histogram ..... 78

Figure 13 - Histogram Reconstruction Loss Normal vs Outliers. ..... 78

Figure 14- PCA dimensional reduction of test points..... 79

Figure 15- Rare category test confusion matrix..... 84

Figure 16 - Example of CALDM's VAE's latent space ..... 85

Figure 17– PCA approximation of Latent Space..... 85

## List of Tables

Table 1- BLS Dataset outlier types ..... 36

Table 2 - BLS Dataset columns ..... 37

Table 3- Composition of outliers in dataset ..... 38

Table 4- Description and count of categorical data in the salary dataset ..... 38

Table 5 - ADBench datasets utilized ..... 44

Table 6 - Unsupervised algorithms utilized for comparison and benchmarking ..... 45

Table 7- Overall ranking of CALDM by dataset tested ..... 70

Table 8 - Detailed AUROC results of all systems ..... 71

Table 9 - Detail AUROC rankings of all systems ..... 72

Table 10 - AP results of all systems ..... 73

Table 11- Detailed AP Rankings of all systems ..... 74

Table 12- Performance results for BLS dataset of all systems ..... 75

Table 13- Performance results BLS reduced-size training dataset ..... 80

Table 14- Rare category test - Sample characteristics ..... 83

Table 15- Rare category test - Performance results ..... 83

Table 16 - Rare category test, detailed results ..... 84

## List of Equations

Equation 1- Kolmogorov–Smirnov (KS) ..... 33

Equation 2- Marginal Likelihood of VAE ..... 48

Equation 3 - Lower Bound ..... 49

Equation 4 - ELBO Formula ..... 49

Equation 5 - VAE Loss function ..... 49

Equation 6 - Diffusion forward (noising) process ..... 50

Equation 7 - Diffusion backward (denoising) process ..... 50

Equation 8 - ELBO in Diffusion Models ..... 50

Equation 9 - DDPM ELBO as Sum of KL Divergences ..... 50

Equation 10 - Sinusoidal Positional Embedding ..... 51

Equation 11 - Diffusion Model loss function ..... 51

Equation 12- Attention Formula ..... 52

Equation 13 - CALDM Reconstruction Loss ..... 55

Equation 14 - Mean Squared Error ..... 55

Equation 15 - Average Precision (AP) ..... 61

Equation 16 - Precision ..... 61

Equation 17 - AUROC ..... 61

Equation 18 - Recall ..... 62

Equation 19- F1 Score ..... 62

Equation 20 - False Positive Rate (FPR) ..... 62

Equation 21- CALDM Simplified Reconstruction Loss ..... 77

## List of Symbols

| | |
|---|---|
| $X$ | Input data |
| $\tilde{X}$ | CALDM output data |
| $C$ | Context output from multi-head attention |
| $D_{KL}$ | Kullback-Leibler (KL) divergence |
| $E[x]$ | Expected value of x |
| $\mu$ | Mean of a distribution |
| $\sigma^2$ | Standard deviation of a distribution |
| $\mathcal{L}_{vAE}$ | Variational Autoencoder (VAE) reconstruction loss |
| $\mathcal{L}_{Diff}$ | Diffusion Model reconstruction loss |
| $\mathcal{L}_{REC}$ | CALDM reconstruction loss |
| p(x) | Marginal distribution of x |
| $p(z\|x)$ | Posterior distribution, distribution of z conditioned on x |
| z | Latent variables |

## List of Abbreviations and Acronyms

| | |
|---|---|
| AP | Average Precision or Under the Precision-Recall Curve |
| AUROC | Area Under the Receiver Operating Characteristic Curve |
| BLS | Bureau of Labor Statistics |
| CALDM | Context-Aware Latent Diffusion Model |
| CBLOF | Clustering-based Local Outlier Factor |
| COPOD | Copula-Based Outlier Detector |
| DDPM | Denoising Diffusion Probabilistic Model |
| ECOD | Empirical Cumulative Distribution-based Outlier Detector |
| ELBO | Evidence Lower Bound |
| HBOS | Histogram-based Outlier Detector |
| HR | Human Resources |
| IForest | Isolation Forest |
| KL | Kullback-Leibler divergence |
| KNN | K-Near Neighbors |
| KS | Kolmogorov–Smirnov measure of fitness |
| LDM | Latent Diffusion Model |
| LOF | Local Outlier Factor |
| MSA | Metropolitan Statistical Area |
| MSE | Mean Square Error |
| OCSVM | One-Class Support Vector Machines |
| PCA | Principal Component Analysis |
| VAE | Variational Autoencoder |

# Chapter 1—Introduction

## 1.1 Background

Outlier, anomaly, or out-of-distribution detection is a fundamental challenge in artificial intelligence and machine learning with a large presence in many fields. It includes application as anomaly identification in healthcare diagnostics and industrial systems, fraud detection in financial transactions, and network intrusion detection in cybersecurity, and many others.

Detecting anomalies accurately and timely is an important part of research within AI and machine learning specifically. There is ample literature and available applications ranging from simple regressions to more recent generative methods.

The list of machine learning methods includes supervised (trained with labels), unsupervised (no labels used), and semi-supervised, so there is no generally preferred approach to identify data points that deviate from the norm. Real-world implementations likely utilize a combination of several methods in their outlier identification process. However, these methods often struggle with high-dimensional data, sparse data, and rare category cases (He et al., 2005; Pang, Cao, et al., 2021).

Outliers are data points that deviate significantly from the expected values of the distribution of a dataset. They often indicate errors or rare but sometimes meaningful or interesting instances. The concept of outliers has been extensively studied in statistics and machine learning, with early definitions starting from important works in statistical analysis. For example Hawkings defines an outlier as an observation that deviates so much from the rest, that raises suspicion it was generated by a completely different model

or mechanism (Hawkins, 1980). Barnett also defined an outlier as values so extreme that appear to be produced by a different model (Barnett & Lewis, 1994).

Outlier detection has become even more critical in recent years as the volume of data flows to review has increased exponentially, and the reaction times required have decreased substantially. Nonetheless distinguishing between normal data volatility and true outliers of interest remains a challenge for all industries (Foorthuis, 2021) (Lavin & Ahmad, 2015; Pang, Shen, et al., 2021).

If a larger number of cases are flagged for further review, this may drive a high level of false positives with a potentially high cost to the organization performing the reviews. If the detection model is more conservative and flags a smaller number of cases, an increase in false negatives may occur, with the potential of high costs to the firm due to undetected anomalies.

In many fields, especially those that are data rich in categorical features, which may include categories with rare combinations (such as data used in people analytics), outliers must be understood in their full contextual environment. This understanding requires the use of advanced detection methods capable of handling high-dimensional and sparse data distributions.

Outlier detection methods can be broadly categorized the same way machine learning methods are supervised, semi-supervised, and unsupervised approaches, each with their own strengths and weaknesses.

Supervised methodologies depend on labeled datasets in which anomalies are explicitly tagged. These methods treat outlier detection as a classification problem, using algorithms such as logistic regression, random forests, and neural networks to distinguish

between normal data and outliers. Supervised methods have been shown to be effective when sufficient labeled data are available, yet supervised methods are not practical in many real-world scenarios where anomalies are rare, difficult or costly to label, or evolving at a rate that may be challenging or impossible to follow. Furthermore, supervised models are susceptible to biases and class imbalances, as datasets often contain significantly fewer outliers compared to normal instances.

Given the challenges (time and cost) of acquiring labeled data, the need to generalize across rare categories, and the constantly changing nature of real-life distributions, an **unsupervised** approach is often a more appropriate choice for some outlier detection use cases.

Pay equity has become an important goal for all organizations striving to maintain a fair environment within their workplace. Differences in pay can originate from systemic biases or oversight deficiencies.

Regulatory frameworks such as the Equal Pay Act (EPA); (USA, 1963), Gender Pay Gap Reporting Regulations (UK, 2017), and EU Pay Transparency Directives (EU, 2023) mandate fair pay practices, requiring organizations to review their compensation structures closely.

Besides compliance with these frameworks, addressing pay equity improves employee retention and overall workplace job satisfaction. In general, detecting pay inequities remains a difficult challenge to solve due to the rarity of certain job-role-location combinations.

Current more traditional methods rely heavily on statistical thresholds without considering contextual information in their detection methodology. Thus many of the

current methods struggle when facing high-dimensional and sparse categorical data (Talagala et al., 2021).

If machine learning models are used for anomaly detection, they may suffer from concept drift, where patterns are learned, and anomalies are detected within the learned pattern. The issue arises when distributions and patterns change with time, making the machine learning less capable in detecting new anomalies (Lu et al., 2019).

One struggle for anomaly detection systems is when the predictor variables (not necessarily the target itself) are rare, for example compensation data from a place or business where a company may have only a few employees. This research focuses also on this "rare category" challenge.

This praxis proposes an unsupervised methodology for anomaly detection that uses a latent diffusion model (LDMs); (Rombach et al., 2022), adds to it an attention mechanism (Vaswani et al., 2017) , in order to incorporate contextual information critical for anomaly detection. This proposed methodology has been named **Context Aware Latent Diffusion Model (CALDM)**.

Latent Diffusion Model (LDMs) have shown impressive results working with high dimensional data, mainly imaging. Their internal mechanics involve the reduction of high dimensional space into a low dimensional or latent space. The LDM also incorporates a process of adding gaussian noise to this latent space representation, learning how to reconstruct it back to the high dimensional representation. This reconstruction process makes them good candidates for anomaly detection systems.

For CALDM the addition of an attention mechanism provides the added capability of learning contextual relationships between categorical variables, such as job

function, geographical clusters, business groups, and organizational levels. This ensures that flagged anomalies are contextually meaningful rather than statistical discrepancies.

This research improved on the work done by Zamberg and team where an LDM named TabADM was used for outlier detection (Zamberg et al., 2023). CALDM expands on this work by adding a contextualizer through an embedded attention mechanism which gives an extra level of capability in context dependent outlier detection processes.

The CALDM methodology has several key advantages:

a) Unbiased Anomaly Detection: No labels mean, there is no human bias in determining what is and what is not an outlier. Even if patterns change, the unsupervised nature of the methodology adapts dynamically without any learned biases maintained.

b) Adaptability: Like unbiased features, if patterns and distributions change, CALDM will always learn what is extreme and what is not.

c) Handling of rare categories: The addition of an attention mechanisms ensures contextual interdependencies are learned and considered by CALDM in the outlier detection process.

d) Ability to learn complex relationships: LDMs in general learn very complex probability distributions as shown in the generation of images. This capability will be exploited by CALDM when facing complex and hierarchical structures.

The CALDM methodology could help organizations to implement and manage a more robust and equitable pay structure, ultimately contributing to a fairer workplace environment. This praxis aims to reduce the gap between current generative modeling techniques and practical outlier analysis, providing a more robust and accurate alternative for detecting pay outliers.

As part of an integrated and comprehensive framework, this method of unsupervised outlier detection can combine with other supervised methods to provide a strong and reliable source of information and support for a comprehensive outlier detection process.

### 1.2 Research Motivation

The need to measure activities and outcomes related to human resources can be traced back to the early 20th century (Kaufman, 2014), and the first book on how to measure human resources management was published in 1984 (Fitz-Enz & Davison, 2002). Since then the numbers of published papers on (Human Resources) HR Analytics has rapidly expanded (Marler & Boudreau, 2017) who also define HR Analytics as an HR practice enabled by information technology that uses descriptive, visual, and statistical analyses of data related to HR processes, human capital, organizational performance, and external economic benchmarks to establish business impact and enable data-driven decision-making. This definition emphasizes the linkage between HR metrics and impactful business outcomes.

The motivation for this research stems from a growing need for organizations to ensure a fair and equitable compensation structure. Large and rapidly evolving organizations struggle to quickly identify compensation outliers when data is unlabeled, and responses times are short. Yet the potential adverse outcome of a false negative is large, and the cost of false positive is often a significant burden for the operation of the organization.

Current methodologies struggle to account for the complex inter-relationship found in real data, where the data is mainly categorical and sparse, so outlier detection is nuanced.

CALDM introduces an improved alternative by incorporating LDMs and enhancing it with attention mechanisms. This approach addresses current limitations by:

1. **Improved Performance**: LDMs in general have shown success in dealing with complex data distributions. (Rombach et al., 2022).
2. **Contextual Understanding**: The addition of attention mechanisms incorporate into the machine learning process a level of understanding of the contextual relationship embedded in the data which ultimately determines what can be an outlier or not.
3. **Scalability and Flexibility**: CALDM is an unsupervised methodology, this translates to a process where data can go directly from source to CALDM without going through a complex data labeling process.

This research aims to contribute to both the field of Artificial Intelligence (AI)-driven HR analytics specifically and broad analytics in general. Concerning HR analytics, the findings of this praxis could influence policymaking, shape industry best practices, and empower organizations to make data-driven, fair, and equitable compensation decisions.

More broadly, the result of this praxis offers an efficient tool in CALDM for application in a wide range of fields where a reliable outlier detection tool is needed.

### 1.3 Problem Statement

*Outlier detection is critical problem in all industries. Large financial services firms specifically face costly legal fines due to undetected pay outlier issues, during their annual compensation review process, resulting in court fines and settlements in the U.S. of $91MM per year, based on adjusted United States (US) dollars to year 2025***.** (Healy & Ahamed, 2019) (Leutwiler & Kleiner, 2003)

Outlier detection for compensation, whether regular base pay or performance incentives, is a critical control and compliance task at all financial institutions. A significant portion of a financial institution's expenses is compensation, divided between base pay and a larger annual performance bonus. While compensation is important in many industries, the significance of year-end bonuses in financial services creates a compressed decision timeline, which places pressure on anomaly detection processes, leading to potential operational, reputational, and legal risks. Pay outliers come in two ways: they can be intentional as fraud or bias. Also, they can be unintentional, for example miskeyed entry fields in the pay system. Efficient and accurate detection enables audits and reviews to distinguish legitimate outliers from problematic discrepancies requiring further investigation.

Data in large multinational companies is complex. Operations in many geographies, lines of business and dealing with potential thousands of job codes and tens of thousands of job-level-country combinations, make anomaly detection systems susceptible to larger number of false positives (at an incremental cost to the company) or false negatives, with the potential of large reputational or regulatory risk.

## 1.4 Thesis Statement

*CALDM will improve the performance of current anomaly detection systems by the incorporation of attention mechanisms to contextualize complex interdependencies*. *CALDM's architecture will allow it to learn relationships between categorical variables such as geographic location, job codes, business lines, tenure, performance and others.*

This praxis will focus on CALDM as an outlier detection system for compensation cases, yet it will also test CALDM in the broader scope as a general-purpose outlier detection system.

Finally, the context-aware capabilities of CALDM will make it a more flexible and dynamically adaptable approach to changing patterns of data, making it less prone to false positives and negatives as data evolves in the normal course of business operations and activities.

## 1.5 Research Objectives

The primary objective is to develop CALDM to specifically address the challenges of detecting outliers in compensation in large companies with complex datasets. The secondary objective is to expand the scope of CALDM and assess it as a general-purpose outlier detection system for a wider range of business cases.

These objectives are broken down into the following specific research objectives:

1. To implement CALDM incorporating attention mechanism to address rare categorical cases and complex hierarchical interdependencies.

2. To compare CALDM against traditional unsupervised outlier detection systems, such as local outlier factor (LOF); (Cheng et al., 2019), K-near neighbors (KNN), empirical cumulative distribution-based outlier detector (ECOD). Performance

metrics to use will be: Average Precision (AP), F1 score, recall, and area under the received operating characteristic curve (AUROC).

3. To expand the scope of CALDM beyond compensation datasets, it will incorporate a wide variety of datasets from other fields outside compensation.

By achieving these objectives this research will contribute to the field of HR analytics specifically, and to wider field or risk management in general.

### 1.6 Research Questions and Hypotheses

Traditional outlier detection methods often struggle when dealing with high-dimensional and sparse categorical data.

To address these challenges, this praxis introduces CALDM. This approach aims to improve the accuracy and reliability of pay-anomaly detection by dynamically contextualizing rare categorical combinations. The following research questions and hypotheses have been formulated to guide the process and assess the effectiveness of the proposed methodology.

Research Questions:

- **RQ1:** Can a context-aware latent diffusion model outperform other unsupervised outlier detection algorithms in AUROC and AP metrics in the detection of outliers?
- **RQ2:** Can a context-aware latent diffusion model trained on a dataset that is greatly reduced in size outperform other unsupervised outlier detection algorithms in AUROC and AP metrics in the detection of outliers?

- **RQ3:** Can a context-aware latent diffusion model outperform other unsupervised outlier detection algorithms in AUROC and AP metrics when the target outliers are part of rare or newly introduced categories?

Hypotheses:

- **H1:** CALDMs enhanced with attention mechanisms improve Average Precision (AP) and Area Under the Receive Operating Characteristic (AUROC) in detecting outliers compared to traditional unsupervised methods, including LOF, ECOD, KNN, and others.
- **H2:** CALDMs with attention mechanisms require less training data to accurately detect outliers in underrepresented categories while maintaining or improving AP and AUROC compared to LOF, ECOD, KNN, and others.
- **H3**: CALDMs with attention mechanisms improve overall AP and AUROC in detecting outliers when the target outliers are part of rare or newly introduced categories.

### 1.7 Scope of Research

This research focuses primarily on the problem of compensation outlier detection. The proposed solution, CALDM utilizes compensation data, which includes categorical variables, such as state, metropolitan statistical area (MSA), job information, sector, and industry, alongside numerical data, such as salary and the individual's performance.

The research explores how the incorporation of attention mechanism improves performance by reducing both false positive and negatives vs other available unsupervised methodologies. The primary dataset was sourced from the BLS, and synthetic data were generated from it to simulate rare cases for model validation.

This praxis compares CALDM vs other unsupervised detection models, specifically: KNN, Principal Component Analysis (PCA), ECOD, and others. Performance metrics utilized are AP, AUROC, F1, and recall.

To expand the scope of this research outside compensation, the datasets contained in the ADBench (Han et al., 2022) benchmark project were used to test CALDM.

PyOD package (Zhao et al., 2019) a popular python package was used to test the various competing unsupervised methodologies.

The approach was unsupervised; therefore, the data had no labels available for training. This characteristic reduces the risk of biases, improves delivery times and reduces cost.

**1.8 Research Limitations**

This research encountered several limitations. Range of limitations starts from the availability of data to the inherent limitations encountered in hyper-parameter tuning of deep neural networks.

One key limitation of this praxis is the availability of detailed publicly available compensation data. The research ended up using aggregate data from the Bureau of Labor Statistics (BLS), therefore requiring a second step of synthetic data generation. This limitation means that the synthetically generated data, although sourced from real data from the BLS, may not completely reflect the complexities and nuances of every market where this system may be tested.

Another limitation in model interpretability. This is a challenge for any deep neural network, and more so for a system like CALDM which integrates three different Deep Neural Networks (DNN): A multi-head attention contextualizer, a variational

autoencoder and a diffusion model. Every effort will be made to improve the level of explainability of CALDM.

While this research has presented the advantages in general of unsupervised methodologies, it is important to mention that in general unsupervised methods underperform in detection accuracy vs supervised methods. Thus, methods like CALDM need to be deployed in conjunction with other supervised methodologies to offer a full range of benefits.

Computational requirements are a limitation, especially when dealing with large datasets and deep neural networks. Training and deploying models like CALDMs require increased processing and memory resources. CALDM was trained with Pytorch and a high-capacity Graphics Processing Unit (GPU), which can be a challenge for organizations with limited budgets. Optimization techniques should be explored to mitigate these concerns.

Finally regulatory concerns must always be considered, due to the sensitive nature of personal compensation data. Compliance with data protection regulations, such as General Data Protection Regulation (GDPR) (EU, 2016) and California Consumer Privacy Act (CCPA) (California, 2018) , may restrict the usage of data for training. It is important to review in detail data regulations for the markets where a system like CALDM will be used.

### 1.9 Organization of Praxis

This praxis is organized to present a general overview of the problem to address, the theoretical support on which it stands, the methodology followed, results and conclusions.

The first chapter **Introduction** introduces the general background of the problem to solve and the solution being proposed. The chapter also presents the research questions and hypotheses.

The second chapter, **Literature Review,** describes the theoretical background on which this praxis is supported. Starts with the definitions of outliers or anomalies and goes through current outlier detection methods. The emphasis will be on describing the observed gaps of each method, which this research is trying to mitigate with CALDM.

The third chapter, **Research Methodology**, goes through the design of CALDM. Starts with the data used for training and testing, describes the general architecture and algorithm of CALDM, and finalizes with the performance metrics used for evaluation.

The fourth chapter, **Results and Analysis**, presents the observed results from the experiments conducted using the proposed model. It compares CALDM to all other unsupervised methods tested ranking them by their performance on each individual dataset tested (31 datasets were tested) plus an overall ranking for all datasets.

The fifth chapter, **Discussion and Conclusion**, summarizes findings, addresses limitations found and suggests future areas of potential enhancement and research.

# Chapter 2—Literature Review

## 2.1 Introduction

The chapter begins with the definition of outliers and proceeds to describe the different current approaches for unsupervised outlier detection. Afterwards it proceeds into generative methods in general starting with variational autoencoders (VAEs); (Kingma & Welling, 2014), then LDMs (Rombach et al., 2022), and finalizing with transformers (Vaswani et al., 2017).

## 2.2 Understanding Outliers: Definitions And Types

To design mechanisms to detect them, outliers must be clearly defined. Outliers are broadly defined as data points that deviate significantly from the expected pattern or distribution. Outlier-related work can be traced back to 1949 in seminal papers from (Grubbs, 1950), who introduced one of the first statistical tests (Grubbs test) to identify a single outlier from within a normal population. More recently, the classical reference regarding outliers has become (Hawkins, 1980), whose outlier concepts are often cited today. Later, (Chandola et al., 2009) provided a taxonomy of outliers differentiating between three types: point outliers, contextual outliers, and collective outliers. In tabular data, especially compensation datasets, all three types may be present.

- Point outliers: Individual data points that deviate from a cohort's expected values, for example, salary records that differ from the cohort norm.
- Contextual outliers: Data points that are only anomalous within a specific context. For example, highly experienced employees have a job code–state combination but below-average salary.

- Collective outliers: A group of salaries that are anomalous collectively, despite individual values not appearing unusual in isolation. For example, a group of male employees receive the same performance bonus, which is higher than that received by all other females in the same job code–location.

Chandola (Chandola et al., 2009) highlighted the importance of data semantics and domain context in interpreting whether an observation is truly anomalous. For example, a Chief Executive Officer (CEO's) bonus may be statistically extreme but not necessarily anomalous within the appropriate peer group.

Aggarwal and team described the curse of dimensionality in outlier detection, where distance differences in vector space disappear as the number of dimensions grow (Aggarwal, 2017). This limitation emphasizes the need for utilization of latent space which offered a lower dimensional representation of high dimensional space.

### 2.2.1 Distinguishing Semantic Anomalies from Data Noise

In the practical application of anomaly detection systems, an important concept to understand is the difference between anomalies of interest and noise.

For example, a miskeyed bonus entry, such as $9,999,999, and a correctly entered but potentially unjustified $2,500,000 bonus may both be flagged as statistical outliers due to their large value vs. the median of payouts. However one is a data quality issue, the other may be a policy violation. Both need to be addressed, but they require different solutions.

This distinction between interesting outliers and noise has been addressed in more recent papers (Aggarwal, 2017; Micenková et al., 2015; Micenková et al., 2013). While Context-Aware Latent Diffusion Model (CALDM) effectively identifies statistical

outliers in a context-aware manner, differentiating between genuine anomalies and noisy data remains a broader challenge.

The value of CALDM is in reducing the number of cases that require human review.

### 2.3 Outlier Detection in Tabular Data

Traditional statistical approaches (Barnett & Lewis, 1994), such as z-scores, interquartile range IQR-based thresholds, and the Grubbs Test (Grubbs, 1950), assume underlying normality and homoscedasticity, which are rarely the case in compensation data. Unsupervised machine learning models, like isolation forests (Liu et al., 2008), local outlier factor (Breunig et al., 2000), and one-class support vector machines (Rätsch et al., 2002) are more flexible but can suffer from poor performance with imbalanced data.

Autoencoders have emerged as a deep learning-based alternative, with reconstruction error used as an anomaly score (An & Cho, 2015). However, standard autoencoders lack the ability to model contextual dependencies, which are important in real-world applications. Variational autoencoders (Kingma & Welling, 2014), generative adversarial networks (GAN); (Goodfellow et al., 2014), diffusion models (Sohl-Dickstein et al., 2015), and transformers  are some of the generative models being used to address outlier detection challenges.

#### 2.3.1 Taxonomy of Unsupervised Outlier Detection Methods

Outlier detection methods can be grouped based on their fundamental assumptions and how they define and score anomalies. The following categories summarize key approaches used in tabular anomaly detection:

(a) Proximity-Based Methods

Proximity-based models detect anomalies by analyzing distances or densities in the feature space.

- KNN detects outliers based on their average distance to the k-nearest data points. Larger distances indicate higher anomaly scores.
- Local Outlier Factor (LOF); (Breunig et al., 2000) measures the local density of a point relative to its neighbors. A lower local density implies a higher likelihood of being an outlier.
- Connectivity-based Outlier Factor (COF); (Tang et al., 2002) extends LOF by introducing the concept of chaining in the KNN graph, which, instead of measuring distance to neighbors directly, uses the distance through a connectivity chain, which is a connected path within the members of a cluster.

These methods are intuitive and interpretable but may suffer from high computational cost in large datasets and fail to perform well in high-dimensional spaces, as distance between all points becomes less meaningful when dimensions increase.

(b) Statistical and Probabilistic Models

Statistical models often assume some distributional properties of the data, while probabilistic approaches aim to estimate likelihoods from the observed data. Below is list of the most common statistical and probabilistic models found.

- Histogram-Based Outlier Score (HBOS); (Goldstein & Dengel, 2012) assumes feature independence and computes outlier scores by building

histograms for each feature and measuring the inverse frequency of each observation.

- Empirical Cumulative Distribution-based Outlier Detection (ECOD**)**; (Li et al., 2022) uses the empirical cumulative distribution function (CDF) for each feature to estimate how uncommon or extreme an observation is.
- Copula-Based Outlier Detection (COPOD); (Li et al., 2020) uses a concept called "copula" which is a way to model how variables move together. Copulas are used to calculate the joint outlier probability across features, which is then used to determine outliers.

(c) Projection and Subspace-Based Methods

These methods project high dimensional vector space in reduced or latent space. Common methods used:

- Principal Component Analysis (PCA); (Shyu et al., 2003) determines the principal vectors, which is a representation using orthogonal vectors that capture the maximum variance. Outliers are tagged based on the reconstruction error from this compressed space.
- Local Outlier Detection via Subspace Analysis (SOD); (Kriegel et al., 2009) detects anomalies by analyzing variance in local subspaces.
- Lightweight Online Detector of Anomalies (LODA**)**; (Pevný, 2016) combines projections with histogram analysis to produce anomaly scores.

(d) Ensemble-Based Methods

Ensemble methods use a collection of base detectors to improve robustness:

- Isolation Forest (iForest); (Liu et al., 2008) Utilizes random trees. Anomalies require fewer splits to isolate.
- LODA, mentioned earlier, also falls into this category due to its use of multiple random projections.

These methods are efficient and well-suited for unsupervised settings with complex data distributions.

(e) Cluster-Based Models

Cluster-based methods detect outliers by analyzing the clustering structure of the data. They include the following:

- Clustering-Based Local Outlier Factor (CBLOF); (He et al., 2003) assigns anomaly scores based on cluster size and the distance of a point from its assigned cluster centroid. Points in small or distant clusters are more likely to be anomalies.

These methods work well when the dataset has well-defined clusters, but they can be sensitive to the choice of clustering algorithm and hyperparameters.

(f) Kernel-Based and Neural Models

Kernel-Based and Neural models employ more advanced architectures to learn the structure of normal data:

- One-Class Support Vector Machines (OCSVM); (Schölkopf et al., 2001) learns a decision function that captures the high-density region of the data, treating all other observations as outliers.

- Deep Support Vector Data Description (DeepSVDD); (Ruff et al., 2018) uses a neural network to map data into a latent space and encloses normal points within the smallest-sized sphere in vector space.

These models perform well on complex, non-linear data but are computationally intensive and more difficult to interpret.

### 2.3.2 Context and Motivation

While each of the models described above has shown success in different settings, most lack the ability to model the feature dependencies and contextual relationships found in tabular compensation data. For example, the same salary may be considered normal in one job level and region but anomalous in another. These limitations motivate the development of context-aware generative approaches, such as the LDM presented in this praxis, which can learn complex joint distributions and provide meaningful reconstructions conditioned on structured context.

## 2.4 Generative Models for Tabular Data

Variational autoencoders (VAE) (Kingma & Welling, 2014) introduced an alternate way to learn latent representations with a probabilistic interpretation. For tabular data, conditional VAEs (Nazabal et al., 2020) have been proposed to better handle mixed-type features, for example, numerical and categorical.

Generative adversarial networks (GANs); (Goodfellow et al., 2014), such as Conditional Tabular Generative Adversarial Network CTGAN (Xu et al., 2019) and TableGAN (Park et al., 2018), are powerful for realistic data generation but are notoriously unstable and difficult to train for anomaly detection due to mode collapse.

Diffusion models (Sohl-Dickstein et al., 2015) (Ho et al., 2020) and score-based generative models (Song et al., 2021) offer a stable training procedure and strong sample quality. They rely on learning the score (gradient of the log-probability) of the data distribution through forward and reverse noising processes. For tabular data, (Kotelnikov et al., 2023) and (Shi et al., 2024) successfully applied a diffusion framework, showing improved efficacy in synthetic data generation.

Specific usage of generative models for outlier detection includes (An & Cho, 2015) who uses VAEs and reconstruction loss for the detection of outliers. (Correia et al., 2023) utilized a combination of VAE and multi-head attention to detect outliers in a time-series. Further, (Deecke et al., 2019) implemented a GAN to detect anomalies in images. (Zamberg et al., 2023), (Bhosale et al., 2024), (Graham et al., 2023), (He et al., 2024), (Sattarov et al., 2025) and (Li et al., 2025) were successful implementing diffusion for anomaly detection applications.

### 2.5 Latent Diffusion Models

LDMs (Rombach et al., 2022) extend diffusion models into compressed latent spaces learned by an encoder (typically a VAE). This architecture enables computational efficiency while retaining high data quality. In the latent space, noise is added and removed over successive timesteps, and the denoising network learns to restore latent representations that can be decoded back into fully noise-free data.

The LDM approach is especially suited for outlier detection: samples that are located far from the learned posterior distribution will result in poor reconstructions and produce high anomaly scores. Moreover, LDMs can be conditioned on context features,

which is critical in domains like compensation where outliers are defined relative to peer groups.

Reconstruction-based anomaly detection assumes that normal data can be effectively compressed and reconstructed, but anomalies, being dissimilar to the larger group, will have high reconstruction errors. This principle underlies the design of VAE both for outlier detection specifically (An & Cho, 2015) and in general (Kingma & Welling, 2014), and more recently, LDMs (Kotelnikov et al., 2023; Rombach et al., 2022).

The proposed CALDM in this praxis document follows this principle: it learns a compressed latent representation of the data and performs iterative noising–denoising steps to reconstruct the original input. Samples that deviate significantly from the learned distribution exhibit higher reconstruction loss and are flagged as anomalies.

### 2.6 Transformer-Based Architectures

Transformers (Vaswani et al., 2017) originally were developed for translation, yet they have revolutionized how models handle data. For tabular data, TabTransformer (Huang et al., 2020) and FT-Transformer (Gorishniy et al., 2021) have exhibited state-of-the-art performance by using self-attention to model dependencies between features.

Context-aware modeling is particularly critical in compensation, where the same salary value may be expected for one region and job level but anomalous in another. Integrating transformers into the VAE encoder and as a conditioning mechanism to the diffusion model enables the generative process to dynamically account for differences and variations in context during processes like synthetic data generation and anomaly detection.

### 2.7 Hybrid Models: VAE + Diffusion + Transformer

There is limited but growing literature on combining deep generative architectures. (Niu et al., 2020) explored VAE-GAN hybrids, and (Jiang et al., 2020) and proposed a transformer-VAE system for music representation learning. Combining a VAE for latent representation, a diffusion model for learning complex joint distributions, and a transformer with multi-head attention for context-awareness for outlier detection which is the combination that CALDM uses as its architecture, is superior and highly promising.

Such a design provides the following benefits:

- Robust probabilistic latent variable representations (via VAE)
- Expressive generative modeling (via diffusion)
- Context-sensitive conditioning (via transformer)

This multi-component architecture addresses key shortcomings in conventional outlier detection models, particularly in sparse, mainly categorical, and context-sensitive datasets like compensation data, where single categorical features, once transformed using one-hot encoding, often become hundreds or thousands of sparse features.

Research for this praxis found two papers where a latent diffusion model was enhanced with an attention mechanism:

- Kumar and team introduced a Vision Transformer (ViT)-based U-Net model as the core for diffusion models, treating noised images and time steps as tokens (Kumar et al., 2025).
- Hu and team introduced Anomaly Diffusion, a diffusion-based few-shot anomaly image generation model that utilizes the strong prior information of LDMs

learned from large-scale datasets. The model employs an adaptive attention re-weighting mechanism that dynamically guides the model to focus more on areas with less-noticeable generated anomalies (Hu et al., 2024).

### 2.8 Applications for Compensation Analytics and HR Data

Research on generative models for HR applications is an emerging field, with most literature focusing on predictive analytics or fairness audits using classical statistical methods. (Bučková & Przyłuska-Schmitt, 2025; Kumah et al., 2024; Saouabe et al., 2024) researched generative AI for HR analytics, but these scholars are not widely cited, so the conclusion for this Praxis is that GenAi for HR remains an area for further research and exploration.

Compensation fairness tools typically rely on regression and classification techniques or rule-based comparisons, which are vulnerable to overfitting, bias, and poor generalization for rare subgroups. Outlier detection in compensation data is important for the following reasons:

- Regulatory compliance
- Internal audit and bias detection
- Performance-based compensation adjustment

Generative models, particularly when context-aware, offer a data-driven, probabilistic framework for detecting anomalous pay patterns across diverse employee subgroups. This direction remains largely unexplored, making this work timely and impactful.

## 2.9 Summary and Conclusion

This chapter reviewed the foundational literature in outlier detection, beginning with defining outliers; presenting the main existing unsupervised outlier detection algorithms, emphasizing VAEs, diffusion models, and transformers; and finishing with hybrid models. The convergence of VAE, diffusion, and transformer-based techniques enables a powerful hybrid architecture better suited to the challenges of compensation data but applicable to all other use cases. Furthermore, this combination offers an improved way to address the challenges of detecting outliers when datasets are sparse and some categories in the data are rare (e.g., a new product, type of fraud, or new market), which are common in a scope of use cases wider than HR compensation.

While previous studies have laid the groundwork for each component, their integration in the domain of compensation analytics is emerging. CALDM is an improved generative approach for unsupervised outlier detection in tabular compensation data, specifically tailored to critical HR contexts like pay-equity audits. Unlike prior works, such as DTUNET and AnomalyDiffusion, which focus on computer vision and image-level anomalies, CALDM is designed from the ground up to operate on structured, categorical-rich, high-dimensional tabular data.

CALDM introduces a multi-head attention mechanism within a context encoder, which adds structured categorical contexts, for example, job code, location, and industry relationships as contextual embeddings into the latent space used by the diffusion model. In doing so, it differs from existing models where attention is used in the denoising step but not to embed contextual meaning from the first step in the VAE encoding process and again in the diffusion denoising step.

Research of existing methodologies didn't find other existing models which integrated latent diffusion + contextual attention + VAE-based decoding specifically for tabular HR data.

The next chapter details the architecture, training process, and evaluation strategy for the proposed CALDM for outlier detection.

# Chapter 3—Methodology

## 3.1 Introduction

This chapter details the model design, datasets, experiments, and evaluation metrics used to address the research questions defined for this praxis.

The primary goal of this praxis is to implement an unsupervised (no labels used for training) outlier detection model. Originally designed to detect pay outliers within complex high-dimensional tabular data and later expanded to address a wider range of anomaly detection problems and datasets.

The implemented model is an Latent Diffusion Model (LDM) (Rombach et al., 2022), or a Variational Autoencoder (VAE) (Kingma & Welling, 2014) that generates a reduced-dimension latent space of the data, which are used to train a Denoising Diffusion Probabilistic Model (DDPM**;** (Ho et al., 2020)**.**

A datapoint is encoded by the VAE into its latent space representation, and then it goes through the diffusion model, which converts the latent sample into noise, subsequently denoising it back to a reconstructed latent sample. The reconstructed latent sample is then fully reconstructed by the VAE. The reconstruction loss (difference between original sample and reconstructed sample) is used to determine which data samples may be considered outliers.

The novelty introduced in this praxis is the addition of transformers to the architecture to contextualize the data, which enables the model to be context-aware. This capability is a critical characteristic, as some outliers only exist within specific contexts. Some outlier definitions may be temporal; for example, a point may be an outlier if it occurs in June but not in May. Other outliers may exist due to their location within a

structure, for example, a geographical location, or within a specific job or industry. The complete integrated model is called Context Aware Latent Diffusion Model (CALDM).

### 3.2 Summary of Methodology

The primary purpose of this praxis is to detect outliers in pay from within a dataset that provides distinct levels of information about the location of the person and the type of work and industry where a job is performed and paid. The secondary purpose is to test CALDM as a general-purpose outlier detection system using a wide range of available datasets.

Due to the privacy issues and restrictions of obtaining public compensation datasets with enough information available by person to thoroughly test CALDM, a testing strategy was developed. The strategy was to test using two distinct groups of datasets:

The first group was a single dataset derived from aggregate pay distribution data across the labor force of the U.S., which is published annually by the BLS. From this general distribution dataset, synthetic samples were generated for outlier detection. The BLS dataset, despite being aggregated data, provides multiple levels of detail for each aggregated cohort, which made it both a challenging dataset to test an outlier detection algorithm and a feasible source to produce reliable synthetic data by person.

The second group of data was the complete inventory of outlier datasets provided by the ADBench project (Han et al., 2022). These datasets have a wide scope of sources, dimensionality, and levels of anomalies and provided CALDM with a thorough testing ground not only as a pay-specific outlier detection model but also as a general-purpose outlier detection model for all types of datasets with no labels available. A total of **30**

datasets from ADBench were utilized, each with distinct characteristics of source area, number of features, and the level of outliers embedded within the dataset.

The experiments will consistently test CALDM against other unsupervised outlier detection models available within the PyOD package (Zhao et al., 2019), which is a robust implementation of outlier detection models widely available and utilized for many purposes. Additionally, the TabAdm model (Zamberg et al., 2023), a diffusion outlier detection model, will be added to the list of comparison algorithms. The purpose is to complement those included within PyOD to evaluate the added value of the attention mechanism incorporated in CALDM and compare it to the overall performance in outlier detection found in standard LDMs.

Fourteen unsupervised outlier detection algorithms were used for testing and comparing results to CALDM, each one tested using the pool of 31 datasets (30 from ADBench plus the Bureau of Labor Statistics (BLS) dataset), ensuring a wide and thorough testing ground.

To test the robustness of CALDM using smaller datasets, a small portion (1% or less than 4,000 rows) of the BLS test dataset was also be used to test CALDM versus all other unsupervised algorithms.

A desired feature for outlier detection models, which was also tested using the full BLS dataset, is the capability of algorithms to detect outliers when they belong to rare categories of data. An example of this type of outlier is new products launched or new methods of fraud, for which few data exist, and from which the model must efficiently and accurately separate normal samples from anomalies.

Each comparison algorithm chosen to benchmark CALDM used the exact datasets and train-test split to ensure consistency. Accuracy, sensitivity, F1 score, AP, and AUROC were used to compare CALDM against all other unsupervised algorithms.

### 3.3 Datasets

Due to the lack of comprehensive publicly available compensation datasets that include enough detailed information by individual employees to create a challenging and realistic testing ground, a dual strategy was defined for this praxis to thoroughly test the outlier detection capabilities of CALDM, which utilized two large groups of datasets:

a) The Occupational Employment and Wage Statistics (OEWS) dataset. This is an aggregated dataset by MSA and Occupational Code with over 400,000 rows. This dataset was disaggregated into generating a comprehensive and detailed synthetic dataset with individuals' employment and compensation data.

b) Datasets from the ADBench project. This project has the specific goal of providing the most comprehensive inventory of datasets for benchmarking outlier detection algorithms.

This dual approach ensured the CALDM was thoroughly tested not only on compensation-specific data but also with a wider range of datasets to assess its capabilities as a general-purpose outlier detection model.

#### 3.3.1 Bureau of Labor Statistics (BLS) Dataset

The BLS originates from the United States Bureau of Labor, created in 1884, later transferred in 1913 as the BLS to the newly created Department of Labor.

It measures labor market activity, working conditions, price changes, and productivity in the U.S. economy to support public and private decision-making. The OEWS program within the BLS provides employment and occupation data across the U.S. and is published annually in May.

This dataset can be found at https://www.bls.gov/oes/special-requests/oesm23all.zip. For this praxis, the dataset from May 2023 was used, as it was the latest one published at the time development began.

#### 3.3.1.1 General Characteristics of OEWS dataset

The source file contained 413,327 rows and 32 columns. For a complete description of all columns, please see Appendix A.

Each row in the dataset contained aggregated statistics for each occupational code within each MSA code. Specifically, the BLS dataset provided the following distribution information for each MSA+OCC_Code combination:

- Total Employees: Total number of employees in that specific MSA + OCC code.
- Percentiles 10, 25, 50, 75, and 90: Annual salary at each of those percentiles.

| AREA | AREA_TITLE | AREA_TYPE | PRIM_STATE | NAICS | NAICS_TITLE | I_GROUP | OCC_CODE | OCC_TITLE | O_GROUP |
|---|---|---|---|---|---|---|---|---|---|
| 10180 | Abilene, TX | 4 | TX | 0 | Cross-industry | cross-industry | 11-1011 | Chief Executives | detailed |
| 10380 | Aguadilla-Isabela, PR | 4 | PR | 0 | Cross-industry | cross-industry | 11-1011 | Chief Executives | detailed |
| 10420 | Akron, OH | 4 | OH | 0 | Cross-industry | cross-industry | 11-1011 | Chief Executives | detailed |
| 10500 | Albany, GA | 4 | GA | 0 | Cross-industry | cross-industry | 11-1011 | Chief Executives | detailed |
| 10580 | Albany-Schenectady-Tr | 4 | NY | 0 | Cross-industry | cross-industry | 11-1011 | Chief Executives | detailed |
| 10740 | Albuquerque, NM | 4 | NM | 0 | Cross-industry | cross-industry | 11-1011 | Chief Executives | detailed |
| 10900 | Allentown-Bethlehem-E | 4 | PA | 0 | Cross-industry | cross-industry | 11-1011 | Chief Executives | detailed |
| 11020 | Altoona, PA | 4 | PA | 0 | Cross-industry | cross-industry | 11-1011 | Chief Executives | detailed |
| 11100 | Amarillo, TX | 4 | TX | 0 | Cross-industry | cross-industry | 11-1011 | Chief Executives | detailed |

| AREA | TOT_EMP | A_MEAN | A_PCT10 | A_PCT25 | A_MEDIAN | A_PCT75 | A_PCT90 | Major_OCC_Group |
|---|---|---|---|---|---|---|---|---|
| 10180 | 40 | 240,020 | 98,490 | 135,280 | 202,340 | 266,939 | 306,014 | 11-0000 |
| 10380 | 100 | 99,620 | 40,640 | 60,080 | 79,500 | 113,210 | 170,560 | 11-0000 |
| 10420 | 330 | 228,800 | 67,840 | 102,870 | 168,090 | 230,380 | 268,131 | 11-0000 |
| 10500 | 70 | 177,850 | 38,940 | 57,810 | 126,690 | 235,790 | 267,546 | 11-0000 |
| 10580 | 590 | 215,060 | 83,610 | 132,950 | 169,400 | 192,460 | 224,759 | 11-0000 |
| 10740 | 564 | 161,880 | 71,990 | 108,660 | 143,910 | 197,330 | 223,453 | 11-0000 |
| 10900 | 590 | 238,380 | 102,840 | 138,770 | 190,850 | 246,274 | 279,099 | 11-0000 |
| 11020 | 60 | 214,790 | 89,430 | 122,660 | 183,100 | 241,373 | 276,615 | 11-0000 |
| 11100 | 60 | 253,910 | 95,610 | 120,000 | 208,580 | 276,619 | 320,083 | 11-0000 |

**Figure 1- Example of raw data from BLS dataset**

From this information, a specific model was developed to select the best-fitting distribution from which individual synthetic samples by employee were generated.

#### 3.3.1.2 Generation of individual employees from aggregate data

For each MSA + OCC combination, the distribution information was fed into a model that tested the best-fitting distribution from the following distributions: Normal, Log-Normal, Gamma, Pareto, and Weibull. A Kolmogorov-Smirnov (KS); (Massey Jr., 1951) calculation was performed for each of the five candidate distributions. The distribution with the lowest KS score was chosen, and individual employees were synthetically generated as random draws from the chosen distribution.

The Kolmogorov–Smirnov (KS) test is a non-parametric method used to compare two distributions by measuring the greatest distance between their cumulative distribution functions. Because it makes no assumptions about the underlying shape of the data, the KS test is especially useful for detecting differences between distributions, even when those differences are not apparent.

$$KS = \max_{x} |F_n(x) - F(x)| \quad (1)$$

The process iterated through each MSA + OCC combination. Figure 2 shows an example of the completed distribution of synthetically generated employee salaries for the state of Wyoming.

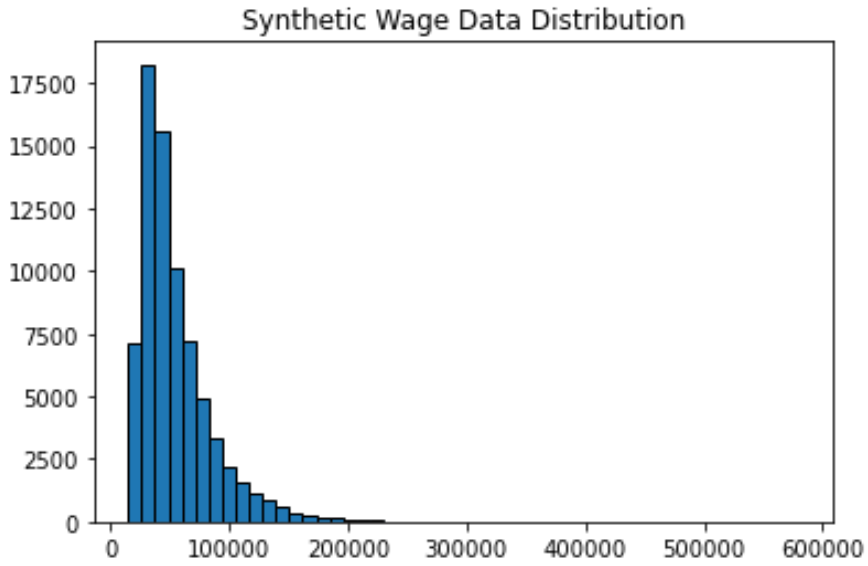



**Figure 2 - Example distribution of synthetically generated employees Wyoming**

### 3.3.1.3 Assignment of Performance Scores to each synthetically generated employee

In addition to geographical, industry, and occupational information provided by the BLS, an additional level of detail and complexity was added to the dataset to enable more comprehensive and challenging outlier detection testing.

A performance flag was added to each synthetically generated employee row. Performance in this project is a generic term that encompasses differences in output from employees driven by natural differentiators like motivation, skills, and experience. The performance ranged from 5 (lowest performance) to 1 (highest performance). The process to derive randomly assigned performance scores began with the annual salary of the employee. Each of the five performance score options had a specific distribution tied to the level of salary of the employee. Performance 1 had a higher probability of coming

from someone with a high salary, whereas Performance 5 had a higher probability of coming from someone with a lower salary.

The assignment of a performance score iterated through every employee, with every performance score (1–5) potentially possible but with different probabilities depending on the level of the employee's salary. Figure 3 shows how each of the performance score distributions were located and overlapped for a specific range of salaries.

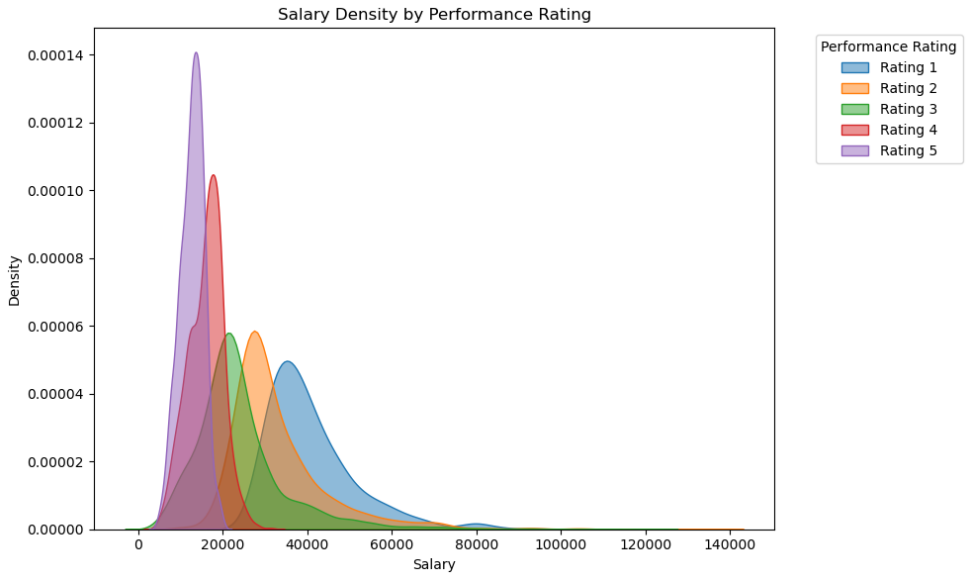



**Figure 3 - Theoretical performance distribution (1–5) for given salaries.**

### 3.3.1.4 Assignment of Outlier labels to BLS data

CALDM is an unsupervised algorithm; as such, it does not use labels for training. However, labels are required to evaluate performance and compare CALDM's performance with other competing algorithms. As the final step in generating the dataset for training, testing, and evaluation, different outlier labels were added to the data using the following rules.

**Table 1- BLS Dataset outlier types**

| Order of assignment | Type of Outlier | Outlier Rule |
|---|---|---|
| 1 | HRS | 97th Percentile of an MSA and Minor OCC group |
| 2 | LRS | 3rd Percentile of an MSA and Minor OCC group |
| 3 | INA | +/- 3 standard deviations from the mean of a Sub-Sector within a State |
| 4 | PVO | +/- 3 standard deviations from the mean of an MSA-OCC combination |
| 5 | PER | Performances 1 and 2, one standard deviation below the mean of MSA-OCC combination. Performances 4 and 5 and one standard deviation above the mean of an MSA–OCC combination |

#### 3.3.1.5 Synthetic MSAs for rare category testing

One of the research questions regards the capabilities of CALDM to detect outliers when a category is small and therefore has limited available training data.

For example, if a multinational firm with a large presence in most countries in the world opened a branch in a new country, it would only have a few employees in that country. The challenge for any model, and especially outlier detection models, is how to detect an outlier when the total category (in this case, the entire country) has few samples, even as few as just one. CALDM's capability to determine from a single sample whether it is "normal" or an "anomaly" was investigated.

To address this challenge, the usage of transformers was introduced with CALDM to contextualize information and provide an alternative to compare rare categories and identify outliers. For example, if the firm had only one employee in the country of Monaco, CALDM using multi-head attention, would learn to compare this single employee against similar employees in South France and Northwest Italy, and from that information determine whether the Monaco employee is an anomaly.

To test the capabilities of CALDM to detect outliers belonging to data defined by rare categories, four synthetic MSAs were created by randomly sampling 1%–5% of employees from four distinct MSAs in Alabama and Florida. Each employee's salary was slightly adjusted using normally distributed noise to ensure that the synthetic MSAs contained slightly different information from the original sources.

#### 3.3.1.6 Final characteristics of complete synthetic dataset

The final synthetic dataset by employee had the following characteristics:

- Total number of rows: 152,884,856
- Total number of columns: 13

**Table 2 - BLS Dataset columns**

| Column Name | Description |
|---|---|
| AREA | MSA code of the employee |
| PRIM_STATE | Two letter code for state |
| OCC_CODE | Occupational Code |
| Major_OCC_Group | Major group where the OCC_CODE belongs |
| Minor_OCC_Group | Minor group code where OCC_CODE belongs |
| Salary | Annual salary of the employee |
| Performance | 1–5 performance score from highest to lowest |
| Industry | Industry code |
| Sub-Sector | Sub-Sector code |
| Sector: | Sector code |
| Outlier: | 1 if outlier, 0 not outlier |
| Outlier_Type: | Type of outlier: HRS, LRS, INA, PER, PVO, NONE |

**Total number of outliers in the complete United States dataset: 8,995,411**

By using the rules defined in 3.3.1.4, the dataset included a challenging group of different anomalies, which helped to test CALDM and compare it to other unsupervised anomaly detection algorithms.

**Table 3- Composition of outliers in dataset**

| Outlier Type | Number of outliers in Dataset | Percent of Total Dataset |
|---|---|---|
| HRS | 1,666,633 | 1.09% |
| LRS | 2,685,856 | 1.76% |
| INA | 2,580,536 | 1.69% |
| PVO | 1,204,496 | 0.78% |
| PER | 857,890 | 0.56% |
| NONE | 143,889,445 | 94.11% |

The dataset included a complex, and therefore challenging structure for anomaly detection, which included a mostly categorical and multi-level hierarchical structure.

**Table 4- Description and count of categorical data in the salary dataset**

| Geographical | Number |
|---|---|
| States (included DC and PR) | 52 |
| MSAs | 396 |
| | |
| **Business / Industry** | |
| Sectors | 24 |
| Sub-Sectors | 85 |
| Industry Groups | 243 |
| Industries (NAICS) | 329 |
| | |
| **Job/Occupation** | |
| Major Occupation Code group | 22 |
| Minor Occupation Code group | 109 |
| Occupation Code (OCC) | 830 |

Due to the lack of comprehensive public compensation datasets and the contribution such a dataset could make not only to compensation professionals but also to the wider data science community, an application and website were developed to host the detailed synthetic BLS dataset. The website was called data.hrmatica.com/wagedata, and it was hosted in an Amazon Web Services (AWS) EC2 instance. The database used was DuckDB**,** and the interface was Python using the StreamLite API. This site is available

for anyone who wants to download specific datasets from within the complete U.S. labor force. See Appendix D.

### 3.3.1.7 Characteristics of Sub-Set used for Training and Testing

Due to the size of the complete dataset (152 MM rows), a smaller yet still material sample was selected. The randomly selected sub-sample was generated from the U.S. states of Georgia, Florida, and Alabama. These specific states were selected due to their geographical proximity and because they offered a diverse mix of larger metropolitan areas, including Miami, Jacksonville, and Atlanta, with smaller, poorer rural areas in central Alabama and Florida. See append E for details.

### 3.3.1.8 Exploratory Data Analysis (Synthetic data derived from BLS aggregate data)

Despite being a synthetic dataset by person, the detailed individual compensation data was created following very detailed BLS aggregate source data, which enabled a precise analysis of the compensation of the complete U.S. labor force.

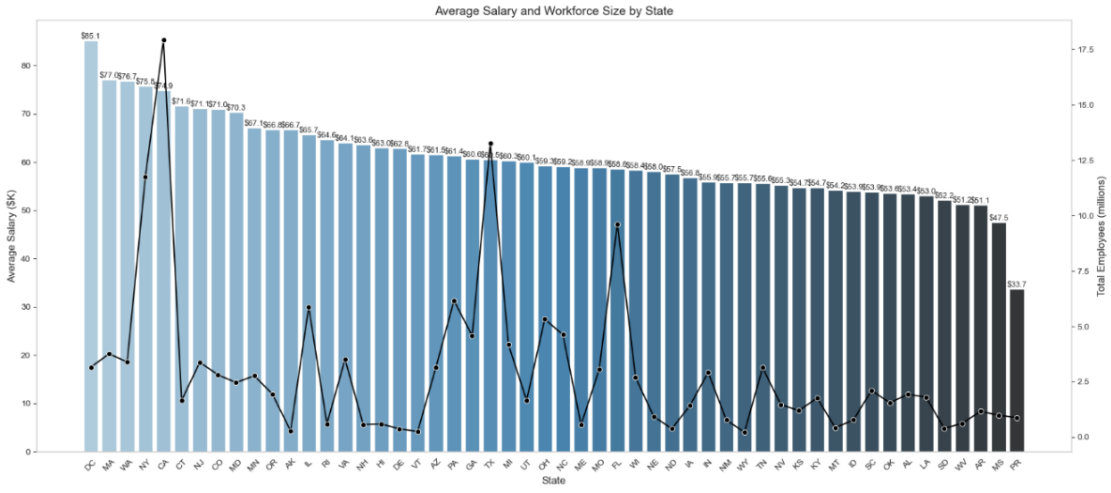



**Figure 4- U.S. States by Average Salary and Employed Population**

Figure 4 shows the wage differentiation between states, which must be accounted for by any outlier detection model. A few datapoints worth noting are the substantially

higher mean salary for DC, which is 10.3% higher than the second-highest state, Massachusetts. Moreover, California is not only the state with the highest employed population but also the fifth in terms of mean salary, with New York having similar stats as the third highest employed population and the fourth highest average salary.

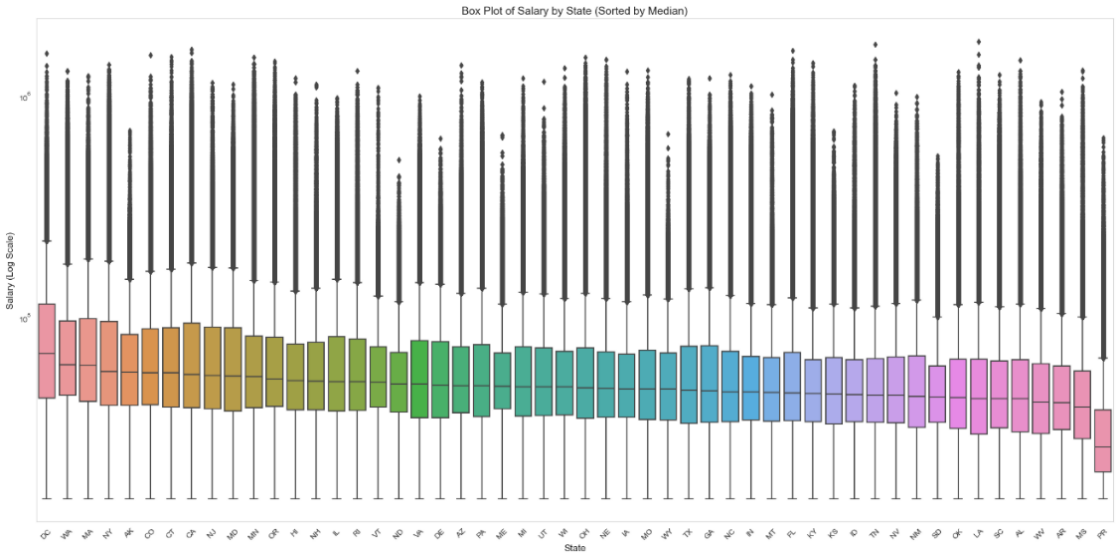



**Figure 5- Boxplot of States by MEDIAN Salary**

The boxplots provide a better perspective on salary distribution by state, showing that all states had a considerable number of outliers by the standard outlier definition found in most statistical packages which is 1.5 times the interquartile range (IQR).

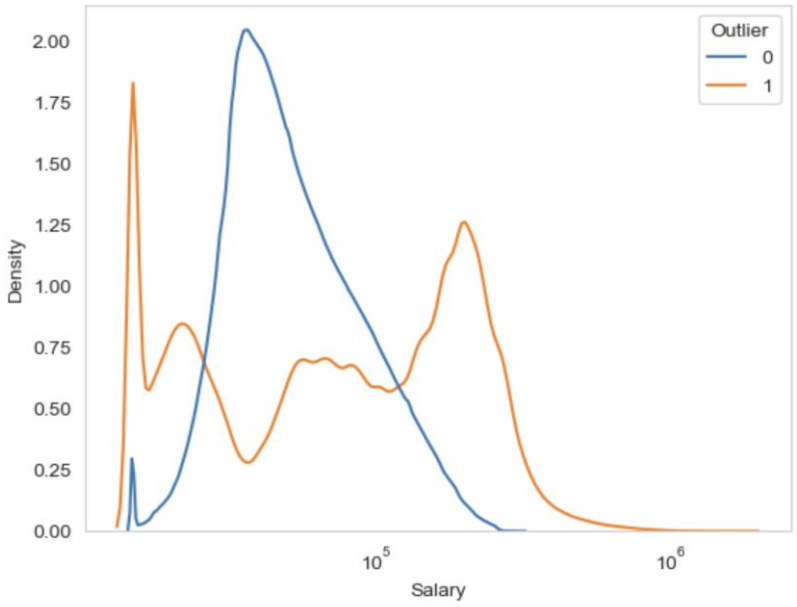



**Figure 6- Density Plots of outlier and non-outlier population**

Density plots of the population by outlier status show that the non-outlier population shows a more typical log-normal distribution, while outliers are present in peak densities at two extremes places in outlier probability density curve. , Outliers located in the lower salary region of the density plot, are people with high levels of performance but low salaries, while outliers located in the higher ranges of salary are people with higher salaries within their respective cohorts.

Appendix B provides state-by-state statistics that provide a deeper insight into the state differences within the data and the U.S. labor force. This dataset becomes challenging for challenging for any anomaly detection system, due to its more complex and hierarchical structure.

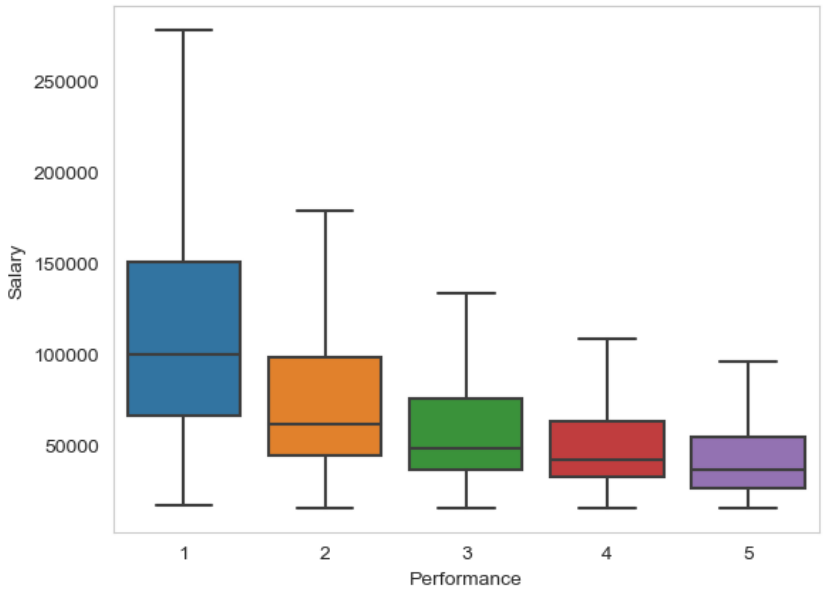



**Figure 7 - Average Salary by Performance Level**

A boxplot of salaries by performance level demonstrates that performance (output, experience, skills) impacts the level of expected salary. Thus, performance is another factor that should be considered by outlier detection models.

The final dataset used for training and testing contained 5.88% of outliers. Although CALDM is an unsupervised model (labels are not included in training), labels in the dataset were used to evaluate the performance of CALDM versus all other unsupervised outlier detection models.

A review of the top 20 occupations by average salary showed physicians in three of the top five positions, with pilots and CEOs completing the top five. Further, their flagged outlier rates were much higher than the rest of the job codes, with an extreme example in nurse anesthetists, who reached outlier rates of 79%. This was due to the logic used for flagging synthetic outliers, where some categories of outliers were chosen by comparing salaries within a sub-sector. Nurse anesthetists would receive, with a high

chance, an outlier label, as this occupational code interacts with other job codes with much lower average pay. See Appendix F for details.

### 3.3.2 ADBench Public Outlier Detection Datasets

To thoroughly evaluate the proposed model across diverse domains beyond the primary one in compensation, this praxis leverages datasets from ADBench (Han et al., 2022), a widely adopted benchmark suite for anomaly detection. ADBench contains 30 real-world (not synthetic data) and publicly available tabular datasets sourced from a variety of domains, including finance, healthcare, cybersecurity, and industrial systems. Table 5 lists all datasets and some characteristics, such as size, number of features, and number of outliers.

Each dataset in ADBench is labeled with a binary outcome indicating whether a sample is an inlier (normal) or an outlier (anomalous). However, consistent with the design of CALDM as an unsupervised model, these labels are used only for evaluation purposes; all models, including the proposed CALDM, are trained in a fully unsupervised manner. ADBench was chosen because of the following characteristics:

- Diversity & Relevance: The scope of fields where those data were sourced is wide (See Table 5) for more details). All the data are real (no synthetic data were used), offering a real testing ground for any algorithm.
- Ease of Comparability: All datasets are provided as NumPy datasets, which standardizes usability and testing by any algorithm that requires to utilize these datasets.

**Table 5 - ADBench datasets utilized**

| Dataset | # Samples | # Features | # Anomalies | % Anomaly | Category of Data |
|---|---|---|---|---|---|
| landsat | 6,435 | 36 | 1,333 | 20.71 | Astronautics |
| satellite | 6,435 | 36 | 2,036 | 31.64 | Astronautics |
| satimage-2 | 5,803 | 36 | 71 | 1.22 | Astronautics |
| shuttle | 49,097 | 9 | 3,511 | 7.15 | Astronautics |
| musk | 3,062 | 166 | 97 | 3.17 | Chemistry |
| PageBlocks | 5,393 | 10 | 510 | 9.46 | Document |
| SpamBase | 4,207 | 57 | 1,679 | 39.91 | Document |
| campaign | 41,188 | 62 | 4,640 | 11.27 | Finance |
| glass | 214 | 7 | 9 | 4.21 | Forensic |
| cardio | 1,831 | 21 | 176 | 9.61 | Healthcare |
| Cardiotocography | 2,114 | 21 | 466 | 22.04 | Healthcare |
| Hepatitis | 80 | 19 | 13 | 16.25 | Healthcare |
| mammography | 11,183 | 6 | 260 | 2.32 | Healthcare |
| thyroid | 3,772 | 6 | 93 | 2.47 | Healthcare |
| celeba | 202,599 | 39 | 4,547 | 2.24 | Image |
| InternetAds | 1,966 | 1,555 | 368 | 18.72 | Image |
| letter | 1,600 | 32 | 100 | 6.25 | Image |
| mnist | 7,603 | 100 | 700 | 9.21 | Image |
| optdigits | 5,216 | 64 | 150 | 2.88 | Image |
| pendigits | 6,870 | 16 | 156 | 2.27 | Image |
| speech | 3,686 | 400 | 61 | 1.65 | Linguistics |
| vowels | 1,456 | 12 | 50 | 3.43 | Linguistics |
| backdoor | 95,329 | 196 | 2,329 | 2.44 | Network |
| Amazon | 10,000 | 768 | 500 | 5.00 | NLP |
| Yelp | 10,000 | 768 | 500 | 5.00 | NLP |
| Ionosphere | 351 | 32 | 126 | 35.90 | Oryctognosy |
| Waveform | 3,443 | 21 | 100 | 2.90 | Physics |
| census | 299,285 | 500 | 18,568 | 6.20 | Sociology |
| donors | 619,326 | 10 | 36,710 | 5.93 | Sociology |
| smtp | 95,156 | 3 | 30 | 0.03 | Web |

### 3.4 Model Development

This section details the development of the CALDM, a novel architecture designed to detect pay outliers in high-dimensional or complex tabular data. The model integrates a VAE to capture a compact latent representation, a multi-head attention context encoder to condition the latent space on cohort-defining attributes, and a latent diffusion model trained to reconstruct clean samples from progressively noised data in

latent space. This section outlines the architectural components, training procedures, and implementation choices made in constructing CALDM.

### 3.4.1 Baseline Models for Comparison

To benchmark CALDM as a general outlier detection model, a comprehensive list of competing unsupervised outlier detection models was assembled.

**Table 6 - Unsupervised algorithms utilized for comparison and benchmarking**

| Method | Type | Detection Principle |
|---|---|---|
| **CBLOF** | Cluster-based | Groups data into clusters and assigns an outlier score based on distance and cluster size (He et al., 2003). Suitable for datasets where clustering assumptions hold. |
| **HBOS** | Histogram-based (Statistical) | Assumes feature independence; builds histograms per feature and computes score by inverse frequency (Goldstein & Dengel, 2012). |
| **iForest** | Ensemble-based (Isolation) | Uses random tree structures to isolate anomalies. Efficient and scalable (Liu et al., 2008). |
| **KNN** | Proximity-based | Outliers are data points far from their k-nearest neighbors. Simple and interpretable. |
| **LOF** | Proximity-based (Density) | Measures local density deviation from neighbors (Breunig et al., 2000). |
| **PCA** | Projection-based | Projects data to principal components; outliers deviate in reconstruction error. |
| **ECOD** | Probabilistic (Empirical CDF) | Uses empirical cumulative distribution functions to compute outlier scores (Li et al., 2022). |
| **COPOD** | Probabilistic (Outlier Prob.) | Uses copulas and empirical probability functions to detect outliers. |
| **COF** | Proximity-based (Connectivity) | Extends LOF by considering connectivity and chaining of neighborhoods. |
| **LODA** | Ensemble (Projection Histograms) | Uses random projections + histogram density estimation (Pevný, 2016). |
| **SOD** | Subspace-based | Finds sparse outliers by computing deviation in local subspaces. |
| **DeepSVDD** | Deep Learning | Trains a neural network to map normal data into a hypersphere; anomalies lie outside. |
| **OCSVM** | Kernel-based | Learns a decision boundary around normal instances in feature space (Schölkopf et al., 2001). |
| **TabAdm** | Diffusion (Reconstruction based) | Trains a diffusion and uses the reconstruction loss from the diffusion process to determine outliers (Zamberg et al., 2023) |

### 3.4.2 CALDM – Architecture And Other Design Considerations

CALDM provides an improved approach for unsupervised outlier detection and was designed specifically to detect pay outliers within a high-dimensional and multi-level occupation and compensation dataset. Furthermore, it was tested as a general-purpose outlier detection system.

The model integrates three core components: a VAE for compressing input data into a lower-dimensional latent representation, an LDM trained on that compressed space to reconstruct this latent space after converting it to pure noise, and a transformer-based context encoder that uses transformers for self-attention, which adds a contextual understanding to both the VAE and the LDM.

This architecture enables the model not only to learn a compact representation of each data point but also to condition the encoding to deeper internal relationships captured by attention mechanisms.

Training goes through two stages: In the first stage, the VAE is trained using a combination of mean squared error (MSE) for reconstruction loss and Kullback-Leibler metric for regularization.

In the next stage, the diffusion model is trained to predict the noise added to latent vectors during a forward diffusion process.

The input data are normalized using standard z-score normalization, then split into 80% training and 20% testing data, with an additional 20% split from the training data used for validation.

Training was conducted using a batch size of 256 and a learning rate of 0.001, with the VAE trained over 20–50 epochs and the diffusion model trained over 200 epochs.

The implementation was built entirely in PyTorch using a combination of built-in neural network modules and custom-defined classes for the VAE, transformer-based context encoder, and diffusion components. Attention was incorporated through PyTorch's nn.TransformerEncoderLayer using a 2-layer Transformer with 4 attention heads and a hidden dimension of 64, followed by a linear projection to a 10–20-dimensional context vector. This context vector is concatenated to the input features before VAE encoding, to the latent vector during decoding, and to the diffusion process, allowing context to influence both compression and reconstruction. For a detailed list of all parameters and architectural configurations used for the contextual encoder, VAE, and diffusion, please refer to Appendix C**.**

The use of a transformer-based context mechanism allows the model to move beyond the more commonly used distance-based anomaly detection mechanism and instead adds a contextual review of a data point, which increases the model's reliability to detect outliers in the data. Figure 8 shows the overall structure and data flow of CALDM.

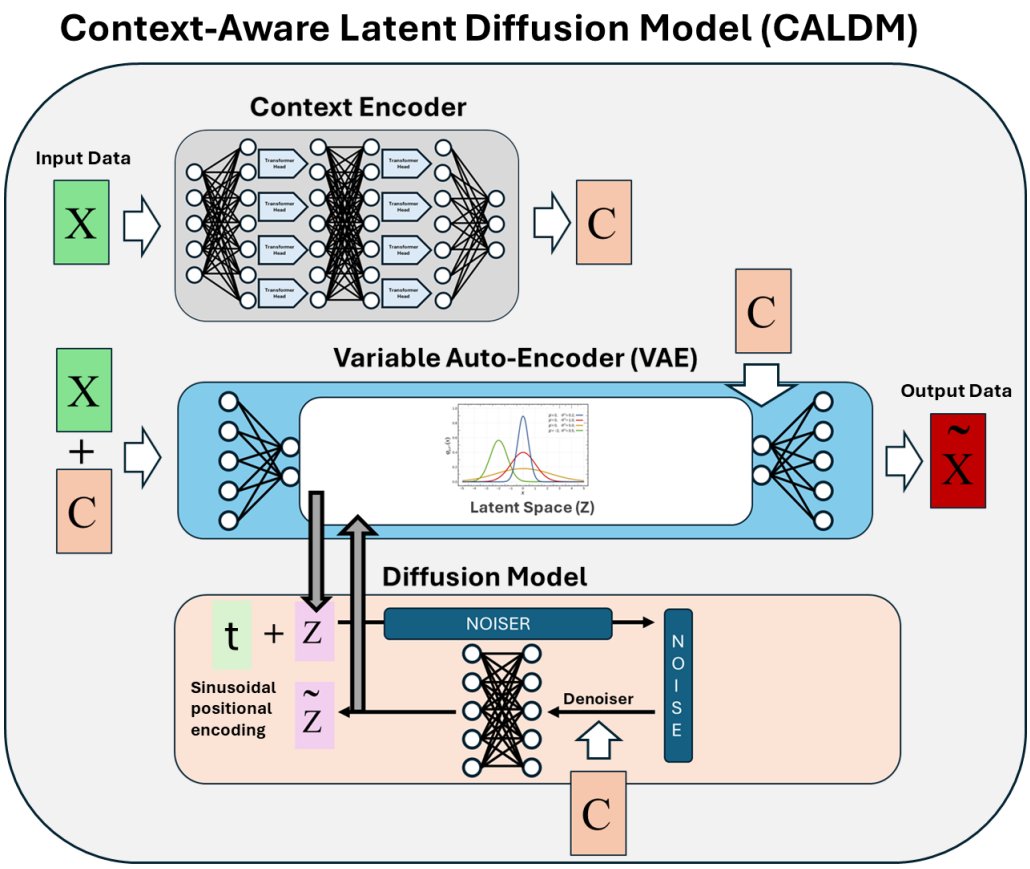



**Figure 8 - Overall architecture and data flow of CALDM.**

### 3.4.2.1 Variable Auto-Encoder

A VAE (Kingma & Welling, 2014) is an improvement of standard autoencoders that, instead of producing a deterministic compressed latent space, learns to compress data into a probabilistic distribution, typically a Gaussian distribution.

The main objective of a VAE is to maximize the marginal likelihood of the data $p(x)$. However, $p(x)$ calculation is intractable due to high dimensionality and a lack of close-ended solutions.

$$logp(x) = log \int p(x \mid z)\, p(z)\, dz \tag{2}$$

To overcome the intractability problem, variational inference was used to turn the original equation into an optimization problem by maximizing the lower bound of *log p(x)*.

$$\log p\,(x) \geq E_{q(z|x)}[\log p\,(x \mid z)] - D_{\mathrm{KL}}\big(q(z \mid x) \parallel p(z)\big) \tag{3}$$

The right side of this inequality is called the Evidence Lower Bound (ELBO).

$$\mathrm{ELBO} = E_{q(z|x)}[\log p\,(x|z)] - D_{KL}\big(q(z|x) \parallel p(z)\big) \tag{4}$$

Maximizing the ELBO maximizes the *log p(x)*, which is the main objective to achieve. It is achieved by maximizing the log likelihood of *p(x|z)* and minimizing the Kullback-Leibler (KL**)** divergence between *q(z|x*) and *p(z)*. The KL divergence is a measure of the distance between two probability distributions. Through a series of transformation steps, it provides an objective loss function that VAEs can minimize, as follows:

$$\mathcal{L}_{\mathcal{VAE}}(x, \hat{x}, \mu, \log\sigma^2) = \mathrm{MSE}(x,\hat{x}) + \beta \cdot D_{\mathrm{KL}}\left(\mathcal{N}(\mu, \sigma^2)\middle|\middle|\mathcal{N}(0, I)\right)\Big) \tag{5}$$

Specifically,

- **x**: original input
- $\hat{\mathbf{x}}$: reconstruction from the decoder
- μ, log σ²: mean and log-variance vectors output by the encoder
- $MSE\big(x, \hat{x}\big) = (1/n)\sum(x_i - \hat{x_i})^2$
- $D_{KL} = -½\sum(1 + log\sigma_j^2 - \mu_j^2 - \sigma_j^2)$

β is the weight controlling the strength of the KL regularization.

For CALDM, a VAE serves the purpose of generating a latent space, which will later be fed into a diffusion model. The reconstruction error of the VAE after its latent

space has been noised and denoised by the diffusion model, it is what is used by CALDM to determine outliers.

#### 3.4.2.2 Diffusion Model

Diffusion models (Sohl-Dickstein et al., 2015) (Ho et al., 2020) are a class of generative models that learn to generate new data by learning how to reverse the effect of adding noise to an original data point, for example, an image. Noise is added in *t* steps to produce an increasingly noisy version of the original data until the version is almost pure noise. Diffusion learns to reverse this process by predicting the incremental noise, which was added at a specific step, therefore learning to return to the original data point. The following equations describe the math behind diffusion models.

$$q(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\overline{\alpha_t}}x_0, (1-\overline{\alpha_t})I\right) \tag{6}$$

This equation represents how noise is added progressively to a datapoint until there is pure noise.

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t)\right) \tag{7}$$

This equation shows how the model learns to predict noise.

$$\log p_\theta(x_0) \geq E_q\left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T} \mid x_0)}\right] \tag{8}$$

$$\log p_\theta(x_0) \geq D_{\mathrm{KL}}\left(q(x_T \mid x_0) \parallel p(x_T)\right) \tag{9}$$

$$+ \sum_{t=2}^{T} E_{q(x_t|x_0)}\left[D_{\mathrm{KL}}\left(q(x_{t-1} \mid x_t, x_0)\right.\right.$$

$$\left.\left. \parallel p_\theta(x_{t-1} \mid x_t)\right)\right] - E_{q(x_1|x_0)}[\log p_\theta(x_0 \mid x_1)]$$

The ELBO technique, similar to what was used in the VAE, offers a tractable objective to learn the reverse denoising process.

For CALDM, the technique utilized is Latent Diffusion (Rombach et al., 2022) in which, rather than learning to denoise the original data, the algorithm learns to denoise the latent space of data points derived from a VAE data compression process.

Additionally, in CALDM, the time variable is added in the backward denoising process. For this purpose, CALDM uses a sinusoidal position encoding, which was introduced in Vaswani's (2017) article "Attention is all you need":

$$\mathrm{PE}_{(\mathrm{pos},2\mathrm{i})} = \sin\left(\frac{\mathrm{pos}}{10000^{\frac{2\mathrm{i}}{\mathrm{d}_{\mathrm{model}}}}}\right) \tag{10}$$

In the reverse process, the noisy data sample from the forward process at time ***t*** is concatenated to the context ***C***, which is generated by the attention process and the positional encoding at time *t*.

During training, the diffusion model minimizes the mean squared error between predicted and true noise.

$$\mathcal{L}_{\mathrm{diff}} = E_{z_{\mathbb{0}},t,\epsilon}[|\epsilon - \epsilon_{\theta}(z_t, t, c)|^2] \tag{11}$$

The interactive denoising process aims to modify pure noise back to the original input. For CALDM, this original input is the latent representation of the sample produced by the VAE.

#### 3.4.2.3 Context Encoder

The context encoder enhances the LDM with attention (Vaswani et al., 2017) by implementing a transformer-style multi-head attention network to encode each input's

contextual relationships into a dynamic vector $\boldsymbol{C}$, which conditions every stage of the model (VAE encoder, VAE decoder, and diffusion denoiser).

Attention is an important development in AI research. It allowed models to add contextual importance information in their training process and inference.

$$\text{Attention}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_k}}\right) V_i \tag{12}$$

Where *Q, K,* and *V* are learned vectors that are multiplied by the input *X*. The produced attention vector can be used in feed-forward networks for a variety of tasks.

The context encoder in CALDM is designed to capture deeper interrelationships within the different columns or features in the data, as outliers in many instances exist only in specific contexts. For CALDM, contextual encoding was implemented using two layers of feed-forward neural networks and two layers of transformers with four attention heads. The output of this two-layer self-attention process is a context vector (*C*), which is concatenated to the input of a VAE and to the input of a diffusion model in their reconstruction processes.

The attention mechanism models interactions among input features and contextual relationships within each sample (e.g., job code, MSA, industry, performance tier), producing a learned context embedding $\boldsymbol{C}$ that reflects each individual's cohort.

This renders CALDM different from a conditional VAE by producing a generalized, dynamic, high-capacity version of conditional generation. By injecting context into both the VAE encoder and decoder and conditioning the diffusion model on it, CALDM aligns the learned latent space with cohort-informed expectations. VAE's KL and reconstruction losses now reflect whether a sample conforms to its contextual norm, not just the global distribution.

This process enabled the model to be sensitive to contextual outliers, rare category interactions (combinations not seen during training), and anomalous reconstruction behavior, even if the raw input seems reasonable.

In summary, the novel addition of a contextual encoder in both the VAE and Diffusion Model in CALDM was intended to add a desired capability in any outlier detector: the capacity to incorporate context within a data point to determine its incorporation or exclusion as an outlier.

### 3.4.2.4 CALDM Algorithm

**Algorithm 1** Context-Aware Latent Diffusion for Outlier Detection (CALDM)

**Input:**
Dataset $X \in R^{N \times D}$, Context $C \in R^{N \times D_c}$
Number of diffusion steps $T$, noise schedule $\{\beta_t\}_{t=1}^{T}$
VAE Encoder $E_\phi$, Decoder $D_\theta$
Context Encoder $f_{\text{ctx}}$, Denoising Network $\epsilon_\theta$

**Output:** Outlier scores $\{L_i\}_{i=1}^{N}$

**for** $i = 1$ to $N$ **do**
  $\mathbf{h}_i \leftarrow f_{\text{ctx}}(C_i)$ ▷ Encode context using multi-head attention
  $(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) \leftarrow E_\phi([\mathbf{X}_i, \mathbf{h}_i])$
  $\mathbf{z}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \text{diag}(\boldsymbol{\sigma}_i^2))$ ▷ Latent sample from VAE
**end for**

▷ Forward Diffusion Process (Training)
**for** each training step **do**
  Sample $t \sim \mathcal{U}(\{1, \ldots, T\})$
  Sample noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$
  $\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_i + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}$
  $\hat{\boldsymbol{\epsilon}}_t \leftarrow \epsilon_\theta([\mathbf{z}_t, \mathbf{h}_i, t])$
  Compute loss: $\mathcal{L}_{\text{diff}} = \|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_t\|^2$
**end for**

▷ Reverse Process (Sampling)
Initialize $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$
**for** $t = T$ to 1 **do**
  $\hat{\boldsymbol{\epsilon}}_t \leftarrow \epsilon_\theta([\mathbf{z}_t, \mathbf{h}_i, t])$
  Estimate $\mathbf{z}_{t-1}$ via reverse diffusion step
**end for**

▷ Reconstruction and Scoring
$\hat{\mathbf{X}}_i \leftarrow D_\theta([\mathbf{z}_0, \mathbf{h}_i])$
$L_i = \|\mathbf{X}_i - \hat{\mathbf{X}}_i\|$ ▷ Reconstruction loss
Identify top-$k$% of samples with highest $L_i$ as outliers

**Figure 9 - CALDM Algorithm**

The CALDM algorithm in Figure 9 outlines the full pipeline for the CALDM, which integrates contextual information into a latent diffusion process for the purpose of detecting anomalies. The model combines three components: (1) a transformer-based context encoder that generates context embeddings, (2) a VAE that compresses the input and context into a latent space, and (3) a denoising diffusion model trained to reconstruct the latent representations from noisy versions.

During training the diffusion model takes samples from the latent space and proceeds to put them through the diffusion process (noising and denoising). The denoising model is then trained to predict the added noise, conditioned on both the noisy latent and contextual embedding, minimizing a reconstruction loss between true and predicted noise.

During inference, the reverse diffusion process generates clean latent representations from standard Gaussian noise, and these representations are then decoded by the VAE to reconstruct the original input. Anomaly scores are computed as the reconstruction error, and instances with the top-k% highest scores are flagged as potential outliers.

The final determination of an outlier from a reconstruction follows these steps: First, reconstruction loss between the original and reconstructed data point is determined:

$$\mathcal{L}_{\text{recon}}(\mathrm{x}, \hat{\mathrm{x}}) = \text{loss}(\mathrm{x}, \hat{\mathrm{x}}) \tag{13}$$

Reconstruction loss is defined for CALDM as the mean squared difference (MSE) between the original and reconstructed data point. For this praxis MSE was used but other losses like Huber Loss (Huber, 1992) could be tested to check whether this alternative loss calculation improves CALDM's performance

$$\mathcal{L}_{\text{MSE}}(\mathrm{x}, \hat{\mathrm{x}}) = \frac{1}{\mathrm{d}} \sum_{\mathrm{i}=1}^{\mathrm{d}} (\mathrm{x}_{\mathrm{i}} - \widehat{\mathrm{x}_{\mathrm{i}}})^2 \tag{14}$$

The final determination of a point as an outlier is whether the reconstruction loss is in the top N percentile of losses above a preset threshold. For example, if model aims to identify 1% of data points as outliers, it will tag as outliers all the points with a reconstruction loss in percentile 99.

### 3.4.3 Advantages over a VAE model for outlier detection

CALDM combines three large components: a context encoder that uses multi-head attention, a VAE, and a diffusion model. This architecture was chosen to exploit advantages each component can offer in the outlier detection process. However, VAEs and diffusion models can and have been used for outlier detection on their own. For example, (An & Cho, 2015), (Chen et al., 2018), and (Correia et al., 2023) used VAEs for anomaly detection. Nevertheless, CALDM offers advantages over VAEs as follows.

VAEs are widely used for anomaly detection due to their ability to learn a latent space representation and reconstruct input data. The reconstruction loss, typically based on MSE or a similar metric, is commonly used to identify outliers. However, using only a VAE for compensation outlier detection presents several limitations:

#### 3.4.3.1 Limitations of VAEs for Outlier Detection

Despite their advantages and wide use, VAE's have some specific disadvantages applicable to outlier detection.

- Overly Smooth Latent Representations: VAEs tend to generate overly smooth latent spaces due to the imposed KL divergence constraint, which can lead to poor reconstruction of complex data distributions.
- Blurry Reconstructions: VAEs often struggle with fine details, as they prioritize learning a structured latent space over high-fidelity reconstructions. In VAE generated images, this can be observed as blurry images which lack detail. This is problematic in tabular data as well, for example when detecting subtle compensation anomalies, for example two nurses with same

performance in same MSA, but different salaries, yet one is a regular nurse, the other a nurse practitioner.

- Mode Collapse in Latent Space: VAEs may fail due to mode collapse, which is the result in which the model always compresses the sample into the same latent space.
- Difficulty Handling Complex Distributions: Compensation data involves complect distribution, involving potentially different job codes, countries, and hierarchical levels.

#### 3.4.3.2 The Benefits of Latent Diffusion Models

LDMs provide several advantages over VAEs when applied to compensation outlier detection:

- Simplified representation: LDMs operate in a simple latent space, enabling them to preserve information while reducing dimensionality.
- Better generalization: Unlike VAEs, LDMs learn high-quality latent representations reducing the blurriness found in VAE reconstructions.
- More varied sampling: The diffusion process ensures that LDMs explore a broader range of data variations, so they are better equipped to detect outlier where the variances are subtle.

### 3.4.4 Advantages of Latent Diffusion Models over pure Diffusion models

DDPMs have demonstrated impressive success in generating new data, especially images, but they have some challenges when dealing with tabular data.

#### 3.4.4.1 Limitations of DDPMs

- High computational cost: DDPMs require many diffusion steps. Models using 1,000 more diffuse steps are common. Because of this, they are computationally expensive for high-dimensional tabular data.
- Difficulty in Handling Sparse Data: Tabular data are commonly sparse due to methods of encoding of categorical variables. The lack of latent representation forces DDPM's to work in this sparse space creating learning difficulties during training.

#### 3.4.4.2 Latent Diffusion Models' Compensation for the Limitations of Diffusion Models

LDMs improve over traditional diffusion models by:

- -Efficient computation: LDMs operate in latent space, which means they operate in few dimensions. For example, an image may easily a 1 million dimensions, while its compressed space may be 32 or lower dimensions.
- Better scalability: Due to more efficient computation described previously, LDMs offer advantages where there is a need to scale the size of the data inputs in both speed and hardware requirements.

### 3.4.5 Conclusion of the advantages of a Latent Diffusion Model

While both VAE's and DDPM's were great advancements in generative ai, they suffer some disadvantages for outlier detection of sparse categorical data.

VAE's suffered from "blurred edges" which not a great issue when generating new images, but it is an issue when those edges represent potential outliers.

DDPM's are computationally expensive, which can be a problem with large datasets both in terms of training time and cost and size of resources required.

LDMs are a convenient middle point which reduces computational needs by using a reduced latent space and eliminates the issue of blurry edges by using the diffusion reconstruction process in determining outliers.

### 3.5 Experimental Setup

First, the input dataset was normalized using standard z-score normalization across all numerical features. The data were then split to 80% training and 20% testing, with an additional 20% of the training set aside as a validation set.

Data augmentation was applied to the training features by adding small Gaussian noise, improving generalization during the VAE training stage. Assigned outlier labels were removed during training to maintain an unsupervised evaluation protocol.

The VAE component of CALDM was trained for 15–50 epochs using batch sizes of 64–256. The optimizer was Adam with a learning rate of 0.01–0.0005. A warm-up strategy was employed for the KL divergence coefficient $\beta$, beginning near zero and gradually increasing over 10 epochs to its maximum value of 1. However, other values could be used, for example, 0.5 or 2, to adjust the weight of the KL value in the overall loss assessment of the reconstruction error.

The diffusion model was trained separately for 200 epochs on the latent representations generated by the VAE, using MSE loss to predict the added noise at each timestep.

Performance was evaluated on the test set by computing the reconstruction loss from the sampled latent vectors, with the top percentage of loss samples flagged as outliers.

Precision, recall, F1-score, and AUROC were calculated using the held-out true labels to benchmark detection performance. All models were implemented in PyTorch using custom modules for the VAE, context encoder, and diffusion process.

### 3.6 Evaluation Metrics

To evaluate the performance of CALDM, several standard metrics for outlier detection and binary classification were employed. The primary evaluation metrics for the publicly available benchmark datasets from ADBench were area under the precision-recall curve (AP) and AUROC.

These metrics are well-suited for imbalanced anomaly detection settings where the proportion of outliers is typically much smaller than that of normal data. AUROC evaluates the model's ability to rank anomalies higher than normal points across all possible thresholds, while AP focuses on performance at the high-precision, high-recall end of the curve, which is critical when false positives must be minimized. These metrics are threshold independent and robust to class imbalance and are thus ideal for assessing unsupervised models on benchmark datasets with known outlier labels.

#### 3.6.1 Average Precision (AP)

AP is a common evaluation metric for ranking-based anomaly detection tasks, particularly when true outliers are rare and high precision at the top ranks is critical. In CALDM, where reconstruction loss was used to rank samples by their likelihood as outliers, AP summarizes the precision-recall curve as a single number.

AP is calculated by ranking all samples according to their anomaly scores (for CALDM, the anomaly score was its reconstruction loss), then computing precision at each position where a true outlier appears. AP is then the average of these precision values over all true outliers.

$$AP = \sum_{k=1}^{n} P(k)\ \Delta R(k) \tag{15}$$

Precision measures how many of the items predicted as positive are actually correct. It reflects the model's ability to avoid false positives.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{16}$$

### 3.6.2 Area Under the Receiver Operating Characteristic Curve

AUROC is a widely used metric to evaluate the performance of binary classifiers and anomaly detection models like CALDM. AUROC measures the model's ability to distinguish between true outliers and normal samples based on their anomaly scores (for CALDM the anomaly score is its reconstruction error).

The AUROC is the area under the curve plotting true positive and false positive rates at different thresholds. It ranges from 0 to 1, with 1.0 indicating perfect separation and 0.5 indicating random performance.

AUROC was one of the chosen metrics for performance evaluation because it is threshold independent. So, it measures how good the model is at all threshold levels.

$$AUROC = \int_{0}^{1} TPR(FPR)\ d(FPR) \tag{17}$$

### 3.6.3 Other Metrics Used

For the BLS data only other metrics were added in addition to AP and AUROC. F1-score and recall were also computed. These are threshold-dependent metrics. Recall quantifies the model's ability to detect all actual outliers, which is important for regulatory and fairness audits, while the F1-score offers a more balanced measure.

- Recall (True Positive Rate) measures how many of the actual positive cases the model successfully identifies. High recall indicates the model is missing few true positives.

$$Recall\ (TPR) = \frac{TP}{TP + FN} \tag{18}$$

- The F1 score is the harmonic mean of precision and recall, balancing both error types in one number. It is especially useful when dealing with imbalanced datasets.

$$F_1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{19}$$

- False Positive Rate measures how often negative cases are incorrectly classified as positive. Low FPR indicates the model avoids generating unnecessary false alarms.

$$FPR = \frac{FP}{FP + TN} \tag{20}$$

## 3.7 Research Questions Operationalization

The experimental design of this research was explicitly structured to align with and operationalize the three core research questions posed in the praxis document. To address the first research question regarding whether CALDM outperforms traditional

unsupervised outlier detection methods, the model was trained and evaluated on full or complete datasets. This included both public datasets from ADBench and the custom synthetic dataset derived from BLS data.

The second research question, concerning the model's performance under data scarcity, was tested through experiments on a reduced-size dataset. A stratified random sample representing just 1% of the original BLS dataset was used.

The third research question focuses on the model's ability to generalize to and detect outliers within rare or underrepresented categorical segments. To simulate this ability, a set of "cloned" MSAs was synthetically generated by duplicating small subsets of data from four selected MSAs in Florida and Alabama. These clones were slightly perturbed using Gaussian noise to introduce diversity while maintaining realism.

### 3.8 Ethical Considerations

This praxis uses synthetic data originating from public aggregate data and therefore does not have any of the ethical or legal privacy issues that practitioners will face if running a paid outlier detection process with real employee data (Allen, 2012; McCoy et al., 2023).

Extreme care is required to ensure that embedded biases in the compensation structure are not perpetuated, for example, that people belonging to a protected minority class are not identified as outliers because, as a group, they are overrepresented in the lower tiers of compensation.

Statistical tests for pay equity should be performed on the outliers identified. For example, chi-squared could be utilized on the outliers to determine whether the proportion of people from a protected class is statistically different from the proportion

within the rest of the population. A higher statistically significant meaningful difference may signal systemic biases that negatively impact such a population. A lower statistically meaningful difference may signal biases on the outlier detection model itself, which needs to be documented and passed on to developers for potential calibration hyper-parameter adjustment.

### 3.9 Limitations of Research

The first limitation of this research originates from the use of synthetic data derived from basic distribution assumptions and interpolations. This methodology may not capture the complexities observed in real compensation data.

The second limitation was the hardware constraints faced. This praxis comprises readily available GPUs offered by Google, such as the Nvidia A100. This is a powerful GPU offered by Google with 40 GB of memory in its virtual (VM) environment. Additionally, a smaller RTX 3070 was used for several runs. These specifications may be adequate for most academic projects or small use case scenarios but may prove insufficient with real compensation data, which may have millions of rows and thousands of columns. For example, if the complete BLS synthetics dataset were used for training, the dimensions would have been over 152 million rows and over 2,000 columns due to the process of one-hot encoding all available categories in the data. This praxis dealt with this limitation by selecting a smaller random sample from only three states (FL, AL, GA). However, this approach may not have captured or tested complexities potentially found in other U.S. states. For a complete test of the complete U.S. labor force specifically, or other high-dimensional datasets in general, a larger cluster of GPUs and memory banks may be required at a considerably higher training cost.

CALDM utilized two independent deep neural networks (DNN). Neural networks contain many hyperparameters and architecture design decisions that substantially impact the performance of the DNN. Due to the resource requirements and time required to perform a single complete training run on the BLS dataset (approximately 10 minutes), a more robust and thorough testing of different hyperparameters and design options was not possible. There is an opportunity to improve performance in the future by expanding the hyperparameter tuning steps within the ML pipeline.

### 3.10 Summary

The methodology for this research aimed to evaluate CALDM as both a pay outlier detection model and a general-purpose outlier detection model.

A significant part of this praxis was spent creating a detailed occupation and compensation synthetic dataset of the complete U.S. labor force, using as basic aggregate statistical data provided by the Bureau of Labor Statistics.

To test CALDM as a general-purpose outlier detection model, 30 datasets from the ADBench projects were used and run through CALDM. The tests include comparison using full datasets, as well as smaller subsets to test the performance of CALDM under a reduced training dataset.

CALDM was compared with 14 or 13 (depending on test) other unsupervised outlier detection models and compared using classification metrics such as accuracy, AP, AUROC, F1, and recall.

These metrics were also calculated for outliers belonging to specifically tagged, rare categories, which were created to test CALDM for its capability to identify outliers when categories are rare, such as new products, markets, or even new data dynamics.

Testing CALDM using 31 diverse datasets against up to14 unsupervised algorithms ensured the model was thoroughly tested as both a specific-pay and general-purpose outlier detection model under the most challenging and diverse scenarios found.

# Chapter 4—Results

## 4.1 Introduction

This chapter presents the empirical results obtained from the experiments conducted on Context-Aware Latent Diffusion Model (CALDM) to determine its outlier detection. A clear and objective presentation of the model's performance across multiple datasets and experimental setups is given, without providing in-depth interpretation or theoretical analysis. Detailed discussion, contextualization, and implications of these findings will be addressed in Chapter 5.

The results of the TabAdm project (Zamberg et al., 2023), which ran outlier detection experiments on the same ADBench datasets, were used to compare CALDM's performance.

## 4.2 Datasets Overview

Testing was divided into groups depending on the type of data used. The purpose was to test CALDM as both compensation-specific and as a general-purpose outlier detection system.

The first group contained the synthetically generated dataset of the complete U.S. labor force derived from detailed aggregate data published by the Bureau of Labor Statistics. The second group was a collection of datasets from the ADBench project.

### 4.2.1 Synthetic BLS Compensation Dataset

The Bureau of Labor Statistics (BLS) publishes annually detailed salary information by occupational code for every Metropolitan Statistical Area (MSA) in the United States. The published dataset, although presented in aggregated form, provided

enough statistical information to generate detailed synthetic data, which was used for testing CALDM.

The dataset comprised approximately 152 million rows. The features include occupational, geographical, industry, and compensation data, which was used by CALDM for outlier detection.

For this praxis, only a subset of the complete dataset was used for training and inference, specifically a subset of the population in Florida, Alabama, and Georgia. The final dataset had a total of 596,856 rows and 202 features (due to one-hot encoding of categorical variables), and 7.3% of the data was tagged as outliers.

### 4.2.2 Public Outlier Detection Datasets

To test CALDM as a general-purpose outlier, 30 datasets from the ADBench project were used. For a complete description of each, please refer to Table 5.

## 4.3 Experimental Results

Three experiments were conducted to test CALDM's capabilities to address the three hypotheses, as follows:

- Capability of CALDM in outlier detection versus the most common unsupervised outlier detection models available. To test CALDM as a general-purpose outlier detection tool, the datasets used were a collection of 30 datasets from the ADBench project (Han et al., 2022). A synthetic dataset derived from occupational and wage information from the BLS was then used to test CALDM as a compensation-specific outlier detection system. CALDM was compared to all other models using the appropriate metrics (Area Under

the Receiver Operating Characteristic (AUROC), Average Precision (AP), recall, and F1-score)

- Change in performance of CALDM versus all other models in outlier detection using a (1%; 4,000 rows) BLS dataset for training that was materially reduced in size.
- Capability of CALDM in detecting outliers within rare or infrequent categories within the BLS dataset.

### 4.3.1 Hypothesis 1

The main goal of this research was to develop an unsupervised anomaly detection system which would outperform other unsupervised alternatives. Research question 1 (RQ1) is closely linked to this main goal:

**RQ1:** Can a context-aware latent diffusion model outperform other unsupervised outlier detection algorithms in AUROC and AP metrics in the detection of outliers?

Hypothesis 1 (H1) addressed this wider-scope research question and objective, by focusing on the full dataset, and evaluating overall performance metrics.

**H1:** CALDMs enhanced with attention mechanisms improve Average Precision (AP) and Area Under the Receive Operating Characteristic (AUROC) in detecting outliers compared to traditional unsupervised methods, including LOF, ECOD, KNN, and others.

For this wider scope experiment, CALDM ranked **number one** as the best-performing outlier detection system for both the BLS and the ADBench datasets experiments.

#### 4.3.1.1 Outlier Detection Datasets

With ADBench. CALDM was ranked number one compared to the other 12 unsupervised outlier detection models. Overall performance was determined as the average ranking in AP and AUROC performance measures for all 30 ADBench datasets tested. The average ranking of CALDM was 4.03 for AUROC and 4.57 for AP.

CALDM was the highest performer eight times for AUROC and seven times for AP in the 30 datasets tested. It performed in the top three models in AUROC for 43% of the 30 datasets, and 50% of the time in AP for all 30 datasets tested.

**Table 7- Overall ranking of CALDM by dataset tested**

| Dataset | CALDM_AUROC | CALDM_AP |
|---|---|---|
| Smtp | 1 | 7 |
| Mammography | 4 | 3 |
| Thyroid | 1 | 1 |
| Glass | 9 | 12 |
| Shuttle | 3 | 5 |
| Donors | 2 | 3 |
| PageBlocks | 1 | 1 |
| Vowels | 5 | 3 |
| Pendigits | 2 | 4 |
| Hepatitis | 5 | 3 |
| Cardio | 1 | 1 |
| Cardiotocography | 5 | 5 |
| Waveform | 10 | 7 |
| Letter | 4 | 6 |
| Ionosphere | 7 | 6 |
| Landsat | 8 | 8 |
| Satellite | 3 | 2 |
| Satimage-2 | 5 | 8 |
| Celeba | 8 | 8 |
| SpamBase | 5 | 4 |
| Campaign | 5 | 8 |
| Optdigits | 6 | 9 |
| MNIST | 1 | 1 |
| Musk | 1 | 1 |
| Backdoor | 5 | 3 |
| Speech | 6 | 3 |
| Census | 1 | 1 |
| Amazon | 1 | 1 |
| Yelp | 4 | 3 |
| InternetAds | 2 | 10 |
| **Average Ranking** | **4.03** | **4.57** |
| **Final overall ranking** | **1** | **1** |

| | | |
|---|---|---|
| **Times #1 for AUROC** | **8** | **27%** |
| **Times #1 for AP** | 7 | 23% |
| | | |
| **Times in Top 3 for AUROC** | 13 | 43% |
| **Times in Top 3 for AP** | 15 | 50% |

AUROC Results

Credit for the experimental results of unsupervised outlier detection algorithms in this table is given to the TabADM authors (Zamberg et al., 2023).

**Table 8 - Detailed AUROC results of all systems**

| Dataset | PCA | OCSVM | LOF | CBLOF | HBOS | KNN | COPOD | IF | ECOD | GOAD | NeuTraL | TabADM | CALDM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Smtp | 81.01 | 74.94 | 93.07 | 83.82 | 81.48 | 90.61 | 90.95 | 89.19 | 87.30 | 87.61 | 79.15 | 86.21 | **94.36** |
| Mammogra | 87.05 | 83.72 | 73.15 | 80.15 | 83.91 | 82.69 | **89.00** | 84.56 | 88.96 | 57.27 | 52.89 | 82.79 | 86.31 |
| Thyroid | 96.04 | 87.58 | 86.37 | 94.10 | 95.91 | 95.49 | 94.07 | 97.73 | 97.84 | 73.19 | 92.20 | 93.84 | **98.00** |
| Glass | 74.72 | 55.27 | 66.02 | 85.59 | 83.23 | 88.39 | 78.06 | 80.65 | 76.02 | 64.41 | 49.25 | **88.71** | 73.06 |
| Shuttle | 98.95 | 98.15 | 53.59 | 75.59 | 98.68 | 63.40 | 99.41 | **99.72** | 99.27 | 98.56 | 95.11 | 98.97 | 99.37 |
| Donors | 82.85 | 73.11 | 61.65 | 65.36 | 71.88 | 62.68 | 81.41 | 77.76 | **88.80** | 30.90 | 50.26 | 72.48 | 85.58 |
| PageBlocks | 90.59 | 89.30 | 72.67 | 87.37 | 81.14 | 81.42 | 87.74 | 89.26 | 91.55 | 80.24 | 83.98 | 90.83 | **91.81** |
| Vowels | 62.54 | 58.31 | 94.76 | 90.69 | 68.74 | 98.23 | 52.45 | 79.17 | 61.43 | 93.46 | 45.94 | **96.80** | 93.33 |
| Pendigits | 92.61 | 92.62 | 51.44 | 90.26 | 92.20 | 72.97 | 88.99 | **94.77** | 91.60 | 63.17 | 86.04 | 86.20 | 94.06 |
| Hepatitis | 79.05 | 67.86 | 79.52 | 73.81 | 79.76 | 71.90 | **82.38** | 74.29 | 75.48 | 69.05 | 55.75 | 71.43 | 76.47 |
| Cardio | 94.27 | 91.96 | 64.08 | 80.27 | 83.34 | 72.17 | 91.16 | 90.91 | 92.90 | 59.40 | 26.30 | 81.70 | **98.14** |
| Cardiotocogr | 75.82 | **79.19** | 58.43 | 65.97 | 60.51 | 54.63 | 66.80 | 68.85 | 79.13 | 39.79 | 77.29 | 60.97 | 68.85 |
| Waveform | 62.08 | 51.86 | 70.51 | 71.39 | 67.84 | **72.21** | 71.94 | 69.17 | 59.38 | 70.80 | 66.16 | 71.40 | 65.70 |
| Letter | 53.31 | 51.33 | 87.43 | 76.28 | 61.43 | 87.77 | 55.03 | 63.19 | 56.82 | 80.55 | 23.06 | **91.04** | 81.47 |
| Ionosphere | 79.03 | 74.42 | 85.93 | 89.52 | 65.65 | 92.43 | 79.54 | 84.43 | 73.66 | 87.49 | 83.41 | **92.67** | 84.14 |
| Landsat | 36.63 | 36.62 | 55.94 | 62.05 | 57.01 | 59.00 | 42.50 | 49.49 | 37.09 | 59.11 | **64.16** | 58.61 | 55.09 |
| Satellite | 60.34 | 59.97 | 55.17 | 72.97 | 75.27 | 65.53 | 63.47 | 71.38 | 58.51 | 64.90 | **77.25** | 72.53 | 74.77 |
| Satimage-2 | 97.41 | 97.13 | 49.86 | 99.87 | 97.66 | 92.92 | 97.18 | 99.21 | 96.14 | 97.47 | **99.94** | 99.31 | 98.51 |
| Celeba | **78.58** | 69.70 | 42.49 | 60.19 | 75.63 | 57.46 | 75.17 | 69.91 | 75.81 | 26.69 | 70.79 | 69.05 | 69.15 |
| SpamBase | 56.36 | 54.96 | 45.14 | 57.78 | 65.69 | 54.82 | **69.11** | 61.37 | 66.00 | 38.26 | 40.02 | 58.93 | 59.36 |
| Campaign | 73.80 | 66.77 | 56.57 | 65.85 | **79.61** | 72.40 | 78.74 | 71.12 | 77.51 | 41.93 | 74.77 | 72.36 | 74.41 |
| Optdigits | 49.97 | 52.84 | 53.05 | 73.28 | **80.34** | 38.92 | 66.71 | 69.23 | 58.93 | 69.49 | 58.32 | 59.23 | 62.03 |
| MNIST | 85.35 | 82.69 | 68.30 | 79.67 | 61.94 | 81.42 | 77.92 | 80.47 | 75.10 | 83.58 | 88.49 | 86.13 | **93.36** |
| Musk | **100.00** | 81.19 | 37.80 | **100.00** | **100.00** | 70.93 | 94.71 | 99.97 | 95.58 | **100.00** | 76.15 | **100.00** | **100.00** |
| Backdoor | 88.68 | 84.77 | 71.57 | 83.10 | 75.66 | 68.20 | 78.91 | 73.22 | 84.56 | 90.48 | 89.99 | **91.83** | 88.14 |
| Speech | 51.39 | 50.99 | 53.99 | 51.17 | 51.92 | 52.26 | 53.13 | 53.50 | 51.46 | 46.74 | 40.19 | **54.54** | 52.05 |
| Census | 65.58 | 52.77 | 48.05 | 58.85 | 62.75 | 64.12 | 66.62 | 60.77 | 65.41 | 53.62 | 50.03 | 64.99 | **68.92** |
| Amazon | 52.61 | 52.87 | 55.39 | 53.33 | 53.14 | 56.32 | 52.31 | 52.92 | 52.32 | 50.36 | 54.57 | 54.71 | **60.65** |
| Yelp | 56.09 | 56.30 | 61.94 | 60.12 | 56.44 | **62.17** | 55.33 | 56.93 | 55.60 | 50.36 | 59.58 | 58.77 | 59.92 |
| InternetAds | 60.24 | 68.75 | 64.77 | 68.41 | 68.28 | 70.10 | 67.51 | 67.90 | 67.57 | 41.82 | 46.10 | **76.48** | 75.24 |

The ADBench group of datasets offered a comprehensive and diverse baseline to anomaly testing. AUROC's for all datasets range approximately between 50% and 100% for all systems tested. CALDM's AUROC range went as low as 52% for the Speech dataset, and as high as 10 datasets with an AUROC of 90%+.

**Table 9 - Detail AUROC rankings of all systems**

| Dataset | PCA | OCSVM | LOF | CBLOF | HBOS | KNN | COPOD | IF | ECOD | GOAD | NeuTraL | TabADM | CALDM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Smtp | 11 | 13 | 2 | 9 | 10 | 4 | 3 | 5 | 7 | 6 | 12 | 8 | 1 |
| Mammogra | 3 | 7 | 11 | 10 | 6 | 9 | 1 | 5 | 2 | 12 | 13 | 8 | 4 |
| Thyroid | 4 | 11 | 12 | 7 | 5 | 6 | 8 | 3 | 2 | 13 | 10 | 9 | 1 |
| Glass | 8 | 12 | 10 | 3 | 4 | 2 | 6 | 5 | 7 | 11 | 13 | 1 | 9 |
| Shuttle | 6 | 9 | 13 | 11 | 7 | 12 | 2 | 1 | 4 | 8 | 10 | 5 | 3 |
| Donors | 3 | 6 | 11 | 9 | 8 | 10 | 4 | 5 | 1 | 13 | 12 | 7 | 2 |
| PageBlocks | 4 | 5 | 13 | 8 | 11 | 10 | 7 | 6 | 2 | 12 | 9 | 3 | 1 |
| Vowels | 9 | 11 | 3 | 6 | 8 | 1 | 12 | 7 | 10 | 4 | 13 | 2 | 5 |
| Pendigits | 4 | 3 | 13 | 7 | 5 | 11 | 8 | 1 | 6 | 12 | 10 | 9 | 2 |
| Hepatitis | 4 | 12 | 3 | 8 | 2 | 9 | 1 | 7 | 6 | 11 | 13 | 10 | 5 |
| Cardio | 2 | 4 | 11 | 9 | 7 | 10 | 5 | 6 | 3 | 12 | 13 | 8 | 1 |
| Cardiotocogr | 4 | 1 | 11 | 8 | 10 | 12 | 7 | 5 | 2 | 13 | 3 | 9 | 5 |
| Waveform | 11 | 13 | 6 | 4 | 8 | 1 | 2 | 7 | 12 | 5 | 9 | 3 | 10 |
| Letter | 11 | 12 | 3 | 6 | 8 | 2 | 10 | 7 | 9 | 5 | 13 | 1 | 4 |
| Ionosphere | 10 | 11 | 5 | 3 | 13 | 2 | 9 | 6 | 12 | 4 | 8 | 1 | 7 |
| Landsat | 12 | 13 | 7 | 2 | 6 | 4 | 10 | 9 | 11 | 3 | 1 | 5 | 8 |
| Satellite | 10 | 11 | 13 | 4 | 2 | 7 | 9 | 6 | 12 | 8 | 1 | 5 | 3 |
| Satimage-2 | 8 | 10 | 13 | 2 | 6 | 12 | 9 | 4 | 11 | 7 | 1 | 3 | 5 |
| Celeba | 1 | 7 | 12 | 10 | 3 | 11 | 4 | 6 | 2 | 13 | 5 | 9 | 8 |
| SpamBase | 8 | 9 | 11 | 7 | 3 | 10 | 1 | 4 | 2 | 13 | 12 | 6 | 5 |
| Campaign | 6 | 10 | 12 | 11 | 1 | 7 | 2 | 9 | 3 | 13 | 4 | 8 | 5 |
| Optdigits | 12 | 11 | 10 | 2 | 1 | 13 | 5 | 4 | 8 | 3 | 9 | 7 | 6 |
| MNIST | 4 | 6 | 12 | 9 | 13 | 7 | 10 | 8 | 11 | 5 | 2 | 3 | 1 |
| Musk | 1 | 10 | 13 | 1 | 1 | 12 | 9 | 7 | 8 | 1 | 11 | 1 | 1 |
| Backdoor | 4 | 6 | 12 | 8 | 10 | 13 | 9 | 11 | 7 | 2 | 3 | 1 | 5 |
| Speech | 9 | 11 | 2 | 10 | 7 | 5 | 4 | 3 | 8 | 12 | 13 | 1 | 6 |
| Census | 3 | 11 | 13 | 9 | 7 | 6 | 2 | 8 | 4 | 10 | 12 | 5 | 1 |
| Amazon | 10 | 9 | 3 | 6 | 7 | 2 | 12 | 8 | 11 | 13 | 5 | 4 | 1 |
| Yelp | 10 | 9 | 2 | 3 | 8 | 1 | 12 | 7 | 11 | 13 | 5 | 6 | 4 |
| InternetAds | 11 | 4 | 10 | 5 | 6 | 3 | 9 | 7 | 8 | 13 | 12 | 1 | 2 |
| **Average Ranking** | **6.77** | **8.90** | **9.07** | **6.57** | **6.43** | **7.13** | **6.40** | **5.90** | **6.73** | **9.00** | **8.57** | **4.97** | **4.03** |
| **Final Ranking** | **8** | **11** | **13** | **6** | **5** | **9** | **4** | **3** | **7** | **12** | **10** | **2** | **1** |

Results showed how CALDM was the number one performing system in about a quarter of the tests, top three in about half of the tests. Once all results are scored, CALDM average ranking is about 4 which makes it the best performing algorithm tested.

Average Precision Results

Credit for the experimental results of unsupervised outlier detection algorithms in this table is given to the TabADM authors (Han et al., 2022).

## Table 10 - AP results of all systems

| Dataset | PCA | OCSVM | LOF | CBLOF | HBOS | KNN | COPOD | IF | ECOD | GOAD | NeuTraL | TabADM | CALDM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Smtp | 44.75 | 5.35 | 1.43 | 44.42 | 8.70 | 48.38 | 0.46 | 0.44 | **62.09** | 41.92 | 31.36 | 46.10 | 33.54 |
| Mammogra | 19.93 | 11.84 | 10.33 | 7.96 | 19.35 | 16.10 | 40.23 | 19.50 | **40.67** | 5.84 | 3.04 | 15.68 | 32.46 |
| Thyroid | 42.26 | 19.49 | 16.87 | 26.16 | 52.55 | 29.88 | 19.41 | 60.16 | 49.60 | 35.65 | 34.44 | 31.19 | **71.68** |
| Glass | 20.65 | 16.81 | 22.53 | 35.28 | 25.42 | 26.82 | 18.97 | 21.59 | 25.68 | 16.67 | 7.80 | **39.06** | 14.86 |
| Shuttle | 90.35 | 94.69 | 9.68 | 40.51 | 95.07 | 16.42 | 95.56 | **97.92** | 89.81 | 85.60 | 52.89 | 92.52 | 93.23 |
| Donors | 16.68 | 10.42 | 10.50 | 8.25 | 12.24 | 10.64 | 20.76 | 13.07 | **26.33** | 5.13 | 5.58 | 11.59 | 18.25 |
| PageBlocks | 52.82 | 51.88 | 35.77 | 56.64 | 35.34 | 45.44 | 37.42 | 45.78 | 52.34 | 46.15 | 29.95 | 57.27 | **59.01** |
| Vowels | 8.54 | 9.19 | 43.12 | 23.99 | 11.64 | **67.26** | 4.05 | 15.79 | 10.46 | 39.49 | 5.28 | 64.12 | 56.33 |
| Pendigits | 19.86 | 19.51 | 3.88 | 16.37 | 24.16 | 6.23 | 16.54 | **28.71** | 25.19 | 5.92 | 14.95 | 9.83 | 22.85 |
| Hepatitis | **50.94** | 28.67 | 37.08 | 31.76 | 43.31 | 29.73 | 50.84 | 34.82 | 36.97 | 33.23 | 22.54 | 32.30 | 45.07 |
| Cardio | 57.18 | 53.33 | 18.25 | 36.56 | 42.68 | 31.25 | 52.56 | 48.48 | 53.55 | 29.43 | 7.95 | 31.77 | **83.49** |
| Cardiotocogr | 48.48 | 54.43 | 30.42 | 44.03 | 39.37 | 33.88 | 40.78 | 43.03 | 51.65 | 24.40 | **61.81** | 37.07 | 46.36 |
| Waveform | 4.80 | 3.94 | 9.22 | 15.11 | 5.08 | 10.77 | 5.67 | 6.21 | 4.47 | **44.14** | 5.32 | 6.26 | 6.01 |
| Letter | 10.16 | 8.62 | 43.76 | 19.55 | 10.32 | 34.73 | 7.36 | 9.91 | 8.63 | 30.47 | 4.25 | **49.66** | 15.61 |
| Ionosphere | 73.13 | 72.76 | 81.22 | 87.74 | 46.00 | 92.36 | 68.86 | 79.49 | 65.74 | 85.59 | 75.54 | **92.37** | 79.52 |
| Landsat | 16.43 | 16.33 | 26.49 | **29.36** | 23.26 | 25.95 | 17.89 | 20.12 | 16.61 | 24.89 | 26.72 | 24.95 | 22.15 |
| Satellite | 60.80 | 58.98 | 38.45 | 62.19 | 67.70 | 50.55 | 57.18 | 66.20 | 52.65 | 43.94 | **72.87** | 56.15 | 71.32 |
| Satimage-2 | 87.19 | 86.03 | 3.37 | **97.32** | 80.89 | 34.83 | 79.94 | 91.31 | 68.14 | 43.72 | 96.35 | 62.75 | 75.65 |
| Celeba | **11.28** | 7.41 | 1.78 | 3.30 | 9.77 | 2.69 | 9.63 | 7.08 | 9.86 | 1.39 | 4.62 | 4.37 | 4.52 |
| SpamBase | 42.21 | 41.59 | 35.91 | 43.27 | 5.36 | 42.70 | **54.74** | 50.14 | 52.21 | 34.74 | 35.71 | 43.69 | 44.58 |
| Campaign | 28.97 | 28.58 | 12.12 | 21.12 | **37.45** | 26.43 | 37.01 | 30.23 | 35.57 | 10.46 | 31.10 | 29.52 | 28.81 |
| Optdigits | 2.67 | 2.90 | 5.00 | 4.99 | **11.01** | 2.54 | 4.21 | 5.19 | 3.40 | 5.01 | 3.70 | 4.22 | 3.49 |
| MNIST | 39.63 | 32.97 | 22.73 | 30.91 | 13.41 | 38.31 | 22.37 | 28.43 | 18.27 | 38.35 | 41.96 | 44.17 | **53.73** |
| Musk | **100.00** | 10.47 | 3.57 | **100.00** | **100.00** | 10.78 | 35.61 | 99.80 | 48.73 | **100.00** | 6.98 | **100.00** | **100.00** |
| Backdoor | 52.13 | 8.62 | 20.69 | 7.19 | 5.47 | 30.55 | 6.92 | 3.91 | 9.25 | **55.48** | 36.38 | 41.29 | 50.93 |
| Speech | 2.98 | 3.07 | 3.90 | 3.08 | 3.19 | 3.37 | 3.03 | 2.84 | 3.28 | 2.38 | 1.48 | **3.99** | 3.54 |
| Census | 8.76 | 6.27 | 5.51 | 6.94 | 7.88 | 8.16 | 8.98 | 7.33 | 8.58 | 6.54 | 5.78 | 8.51 | **9.35** |
| Amazon | 5.43 | 5.44 | 5.69 | 5.48 | 5.48 | 5.79 | 5.40 | 5.49 | 5.35 | 5.26 | 5.63 | 5.60 | **7.37** |
| Yelp | 6.01 | 6.11 | 7.56 | 6.50 | 6.15 | **7.57** | 6.04 | 6.17 | 5.84 | 5.01 | 6.85 | 6.55 | 6.89 |
| InternetAds | 27.54 | **53.29** | 37.81 | 53.15 | 52.33 | 41.63 | 50.35 | 50.04 | 50.46 | 53.05 | 17.89 | 53.05 | 38.71 |

Average Precision, the second core metric tested provides another threshold independent way to measure the performance of anomaly detection systems. Once again, the ADBench proves that it offers a challenging testing ground for anomaly detection, with AP numbers going down to the single digits. CALDM obtained the highest AP seven times out of the 30 ADBench datasets tested.

**Table 11- Detailed AP Rankings of all systems**

| Dataset | PCA | OCSVM | LOF | CBLOF | HBOS | KNN | COPOD | IF | ECOD | GOAD | NeuTraL | TabADM | CALDM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Smtp | 4 | 10 | 11 | 5 | 9 | 2 | 12 | 13 | 1 | 6 | 8 | 3 | 7 |
| Mammogra | 4 | 9 | 10 | 11 | 6 | 7 | 2 | 5 | 1 | 12 | 13 | 8 | 3 |
| Thyroid | 5 | 11 | 13 | 10 | 3 | 9 | 12 | 2 | 4 | 6 | 7 | 8 | 1 |
| Glass | 8 | 10 | 6 | 2 | 5 | 3 | 9 | 7 | 4 | 11 | 13 | 1 | 12 |
| Shuttle | 7 | 4 | 13 | 11 | 3 | 12 | 2 | 1 | 8 | 9 | 10 | 6 | 5 |
| Donors | 4 | 10 | 9 | 11 | 6 | 8 | 2 | 5 | 1 | 13 | 12 | 7 | 3 |
| PageBlocks | 4 | 6 | 11 | 3 | 12 | 9 | 10 | 8 | 5 | 7 | 13 | 2 | 1 |
| Vowels | 11 | 10 | 4 | 6 | 8 | 1 | 13 | 7 | 9 | 5 | 12 | 2 | 3 |
| Pendigits | 5 | 6 | 13 | 8 | 3 | 11 | 7 | 1 | 2 | 12 | 9 | 10 | 4 |
| Hepatitis | 1 | 12 | 5 | 10 | 4 | 11 | 2 | 7 | 6 | 8 | 13 | 9 | 3 |
| Cardio | 2 | 4 | 12 | 8 | 7 | 10 | 5 | 6 | 3 | 11 | 13 | 9 | 1 |
| Cardiotocogr | 4 | 2 | 12 | 6 | 9 | 11 | 8 | 7 | 3 | 13 | 1 | 10 | 5 |
| Waveform | 11 | 13 | 4 | 2 | 10 | 3 | 8 | 6 | 12 | 1 | 9 | 5 | 7 |
| Letter | 8 | 11 | 2 | 5 | 7 | 3 | 12 | 9 | 10 | 4 | 13 | 1 | 6 |
| Ionosphere | 9 | 10 | 5 | 3 | 13 | 2 | 11 | 7 | 12 | 4 | 8 | 1 | 6 |
| Landsat | 12 | 13 | 3 | 1 | 7 | 4 | 10 | 9 | 11 | 6 | 2 | 5 | 8 |
| Satellite | 6 | 7 | 13 | 5 | 3 | 11 | 8 | 4 | 10 | 12 | 1 | 9 | 2 |
| Satimage-2 | 4 | 5 | 13 | 1 | 6 | 12 | 7 | 3 | 9 | 11 | 2 | 10 | 8 |
| Celeba | 1 | 5 | 12 | 10 | 3 | 11 | 4 | 6 | 2 | 13 | 7 | 9 | 8 |
| SpamBase | 8 | 9 | 10 | 6 | 13 | 7 | 1 | 3 | 2 | 12 | 11 | 5 | 4 |
| Campaign | 7 | 9 | 12 | 11 | 1 | 10 | 2 | 5 | 3 | 13 | 4 | 6 | 8 |
| Optdigits | 12 | 11 | 4 | 5 | 1 | 13 | 7 | 2 | 10 | 3 | 8 | 6 | 9 |
| MNIST | 4 | 7 | 10 | 8 | 13 | 6 | 11 | 9 | 12 | 5 | 3 | 2 | 1 |
| Musk | 1 | 11 | 13 | 1 | 1 | 10 | 9 | 7 | 8 | 1 | 12 | 1 | 1 |
| Backdoor | 2 | 9 | 7 | 10 | 12 | 6 | 11 | 13 | 8 | 1 | 5 | 4 | 3 |
| Speech | 10 | 8 | 2 | 7 | 6 | 4 | 9 | 11 | 5 | 12 | 13 | 1 | 3 |
| Census | 3 | 11 | 13 | 9 | 7 | 6 | 2 | 8 | 4 | 10 | 12 | 5 | 1 |
| Amazon | 10 | 9 | 3 | 7 | 7 | 2 | 11 | 6 | 12 | 13 | 4 | 5 | 1 |
| Yelp | 11 | 9 | 2 | 6 | 8 | 1 | 10 | 7 | 12 | 13 | 4 | 5 | 3 |
| InternetAds | 12 | 1 | 11 | 2 | 5 | 9 | 7 | 8 | 6 | 3 | 13 | 3 | 10 |
| **Average Ranking** | **6.33** | **8.40** | **8.60** | **6.33** | **6.60** | **7.13** | **7.47** | **6.40** | **6.50** | **8.33** | **8.50** | **5.27** | **4.57** |
| **Final Ranking** | **3** | **11** | **13** | **3** | **7** | **8** | **9** | **5** | **6** | **10** | **12** | **2** | **1** |

With similar results to what was observed with AUROC, the results comparing models on AP, once again showed CALDM to be the best performing system, with an average ranking of 4.6 place among 13 systems tested, with the second place, and third place system averaging 5.3 and 6.3 place respectively.

#### 4.3.1.2 Synthetic BLS Dataset

For the complete BLS compensation dataset, CALDM was number one in all four metrics tested (AUROC, AP, Recall, F1), as it outperformed the other 13 models tested.

**Table 12- Performance results for BLS dataset of all systems**

| Algorithm | AUROC | AP | Recall | F1 | RANK AUROC | RANK AP | RANK Recal | Rank F1 | Total Average Ra | Final Overall Ra |
|---|---|---|---|---|---|---|---|---|---|---|
| CBLOF | 69.46 | 26.72 | 26.91 | 25.91 | 4 | 3 | 4 | 4 | **3.8** | **4** |
| HBOS | 63.12 | 11.12 | 14.94 | 14.43 | 8 | 7 | 8 | 8 | **7.8** | **7** |
| iForest | 59.06 | 9.94 | 11.92 | 11.76 | 9 | 8 | 9 | 9 | **8.8** | **9** |
| KNN | 76.69 | 20.65 | 28.34 | 27.62 | 2 | 4 | 3 | 3 | **3.0** | **3** |
| LOF | 64.35 | 9.55 | 5.59 | 5.46 | 7 | 10 | 13 | 13 | **10.8** | **10** |
| PCA | 57.22 | 8.62 | 8.49 | 8.24 | 11 | 12 | 10 | 10 | **10.8** | **10** |
| ECOD | 68.94 | 14.22 | 20.70 | 19.93 | 5 | 5 | 6 | 5 | **5.3** | **5** |
| COPOD | 67.16 | 13.18 | 18.59 | 18.04 | 6 | 6 | 7 | 6 | **6.3** | **6** |
| COF | 58.40 | 8.67 | 3.23 | 4.40 | 10 | 11 | 14 | 14 | **12.3** | **13** |
| LODA | 50.01 | 7.80 | 7.07 | 7.17 | 14 | 14 | 12 | 12 | **13.0** | **14** |
| SOD | 56.56 | 9.86 | 26.80 | 17.46 | 13 | 9 | 5 | 7 | **8.5** | **8** |
| DeepSVDI | 57.00 | 8.55 | 8.32 | 8.04 | 12 | 13 | 11 | 11 | **11.8** | **12** |
| OCSVM | 73.66 | 33.27 | 30.67 | 30.29 | 3 | 2 | 2 | 2 | **2.3** | **2** |
| CALDM | 79.29 | 37.05 | 36.72 | 36.58 | 1 | 1 | 1 | 1 | **1.0** | **1** |

A confusion matrix (Figure 10) provides an alternative to show a model's performance and summarizes the number of true positives, false positives, true negatives, and false negatives. In outlier detection, a confusion matrix is particularly useful when ground-truth labels are available. Ground-truth in this case is the true assessment of whether a data sample is an outlier. ADBench datasets come with ground-truth outlier labels. The synthetically created United States (US) labor force compensation dataset derived from the BLS dataset also has outlier labels added which serve as ground truth,

Ground-truth labels enable the usage of confusion matrices which allow detail insight into the types of errors made, especially the critical balance between detecting true outliers (true positives) and avoiding false alarms (false positives).

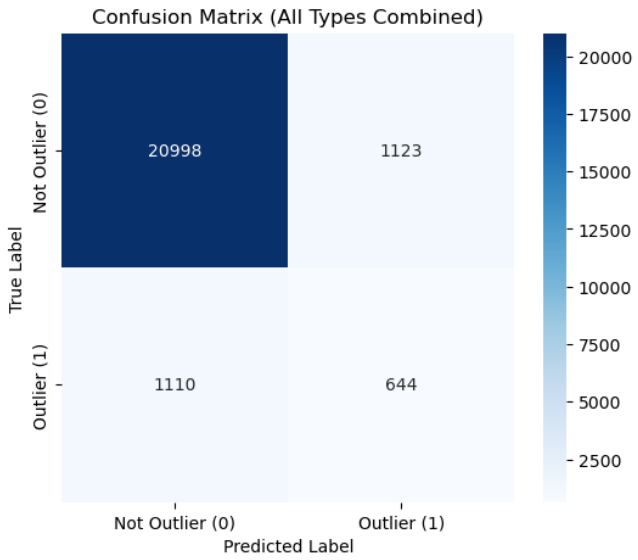



**Figure 10 - BLS dataset test - confusion matrix**

AUROC is an appropriate performance metric for unsupervised outlier detection because it evaluates a model's ability to discriminate between normal and anomalous instances across all possible classification thresholds. Since unsupervised methods do not rely on labeled data during training, AUROC was computed afterward by incorporating the withheld ground-truth labels to assess how well the detection scores separate normal data from outliers.

AUROC's threshold-independent nature renders it especially valuable in anomaly detection, where score distributions can vary. A high AUROC indicates that the model assigns consistently higher anomaly scores to true outliers than to normal observations, and it is therefore a strong indicator of ranking quality even in imbalanced datasets, which are, by definition, the nature of outliers.

Figure 11 shows that CALDM had a smooth curve with a value of **0.7929** for the BLS dataset test.

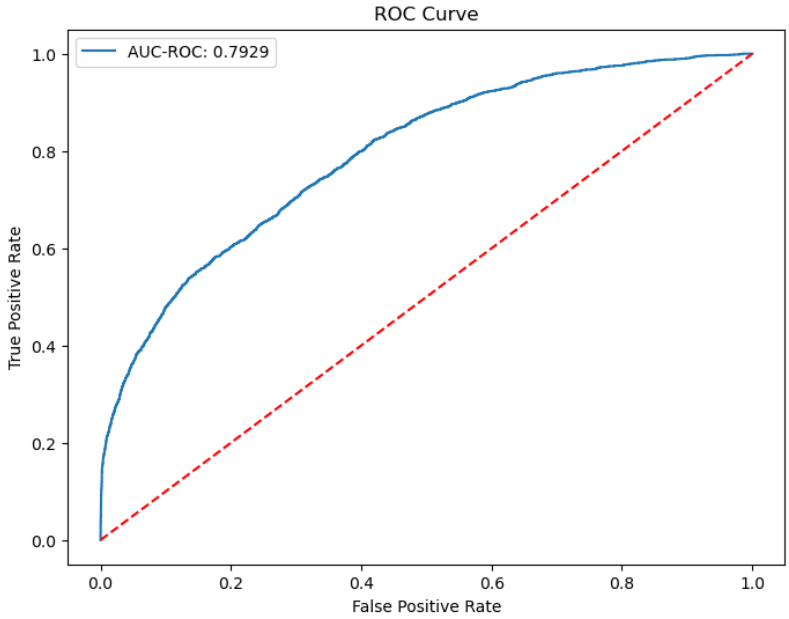



**Figure 11 - AUROC for BLS dataset test**

CALDM performs unsupervised detection of outliers by comparing reconstruction loss, which is the difference between the original data point $X$ and the same data point after being contextualized, converted into its latent representation by the VAE encoder, noised and denoised by the diffusion model, and finally reconstructed back by the VAE decoder. Therefore, reconstruction loss is defined as follows:

$$\mathcal{L}_{rec} = X - \tilde{X} \tag{21}$$

The points with the highest reconstruction loss are flagged as outliers. Figure 12 shows a histogram of all reconstruction losses from the BLS test set. The highly skewed plot demonstrates the capacity of CALDM to identify points that are vastly distant from the generalization of "normal points."

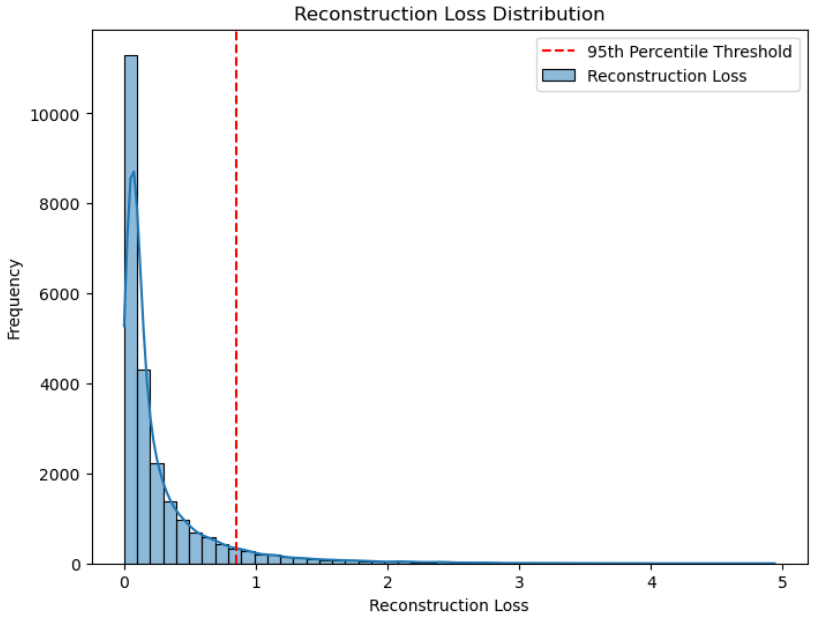



**Figure 12– Reconstruction Loss Histogram**

Plotting the reconstruction loss of normal and outlier plots separately gives insight into how the model perceives them and highlights the challenges in identifying some points that show the same level of reconstruction loss as normal points.

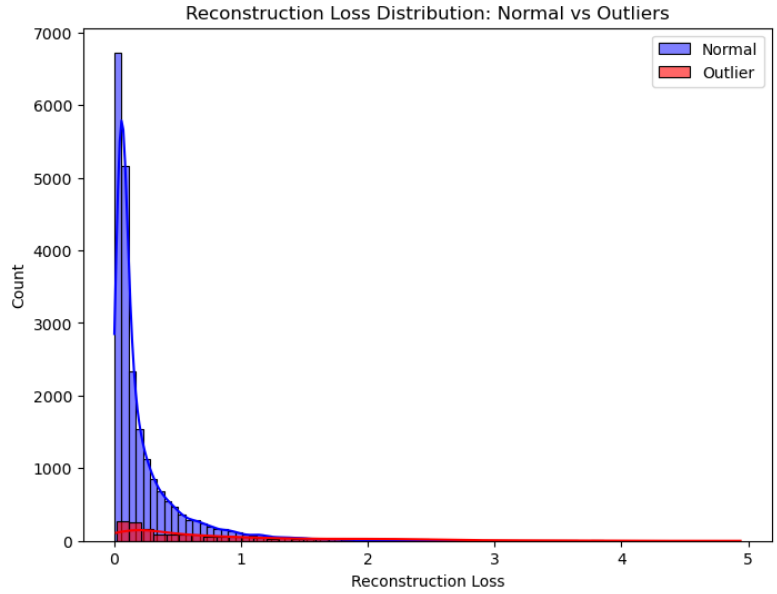



**Figure 13 - Histogram Reconstruction Loss Normal vs Outliers.**

Principal component analysis (PCA) is often used to reduce high-dimensional data to two dimensions for visualization. PCA does it by projecting the data onto the two principal components that capture the most variance. This two-dimensional representation helps to provide some insights into the data and their patterns. In unsupervised tasks like anomaly detection, PCA plots are especially helpful to visually validate whether detected outliers are located far from the main data distribution in the compressed space.

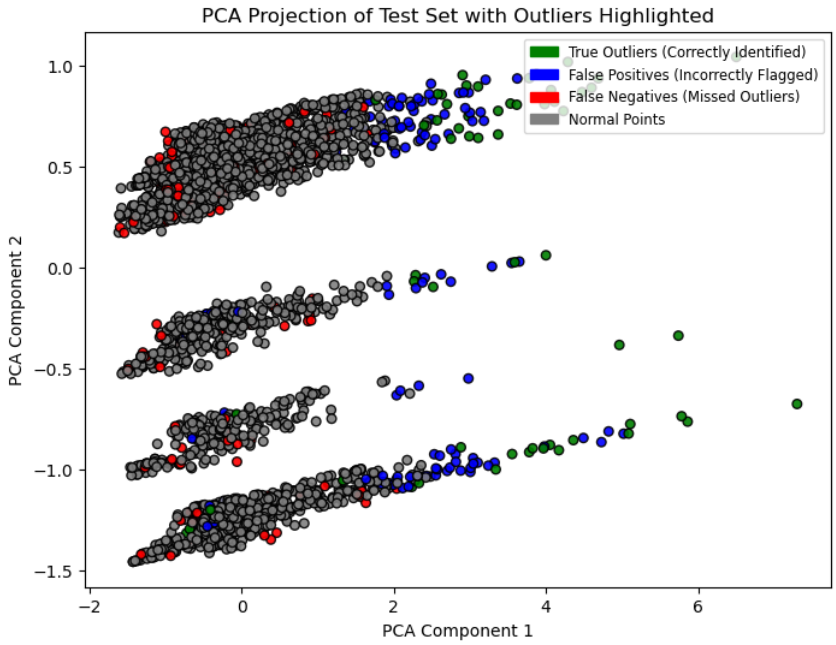



**Figure 14- PCA dimensional reduction of test points**

### 4.3.2 Hypothesis 2

Research question 2 (RQ2) was intended to compare anomaly detection models under conditions where training data may be limited in size.

**RQ2:** Can a context-aware latent diffusion model trained on a dataset that is greatly reduced in size outperform other unsupervised outlier detection algorithms in AUROC and AP metrics in the detection of outliers?

Hypothesis 2 (H2) addressed the second research question. The approach taken was to repeat all experiments with a much smaller dataset to train and test each system.

**H2:** CALDMs with attention mechanisms require less training data to accurately detect outliers in underrepresented categories while maintaining or improving AP and AUROC compared to LOF, ECOD, KNN, and others.

This specific test was only performed on the BLS dataset because not only it is the main dataset used for this praxis, but it was also the largest dataset use for training, thus re-training and testing using only this dataset was a decision made early when the set of tests were designed. Additionally, the set of comparison metrics was expanded to include F1 and Recall.

**Table 13- Performance results BLS reduced-size training dataset**

| Algorithm | AUROC | AP | Recall | F1 | RANK AUROC | RANK AP | RANK Recal | Rank F1 | Total Average Ra | Final Overall Ra |
|---|---|---|---|---|---|---|---|---|---|---|
| CBLOF | 70.64 | 25.89 | 25.29 | 27.67 | 5 | 3 | 3 | 2 | 3.3 | 2 |
| HBOS | 65.02 | 13.98 | 17.24 | 18.29 | 8 | 8 | 9 | 8 | 8.3 | 9 |
| iForest | 67.04 | 16.50 | 20.69 | 21.56 | 7 | 6 | 6 | 6 | 6.3 | 6 |
| KNN | 73.57 | 21.11 | 21.84 | 22.89 | 3 | 4 | 5 | 5 | 4.3 | 5 |
| LOF | 61.90 | 8.90 | 3.45 | 3.64 | 10 | 13 | 13 | 13 | 12.3 | 13 |
| PCA | 59.58 | 9.44 | 13.79 | 12.63 | 12 | 12 | 10 | 11 | 11.3 | 11 |
| ECOD | 72.10 | 18.11 | 27.59 | 26.23 | 4 | 5 | 2 | 3 | 3.5 | 3 |
| COPOD | 69.26 | 16.00 | 18.39 | 19.88 | 6 | 7 | 8 | 7 | 7.0 | 7 |
| COF | 63.98 | 10.67 | - | - | 9 | 10 | 14 | 14 | 11.8 | 12 |
| LODA | 49.67 | 7.96 | 10.34 | 10.84 | 14 | 14 | 12 | 12 | 13.0 | 14 |
| SOD | 61.46 | 11.04 | 25.29 | 17.81 | 11 | 9 | 3 | 9 | 8.0 | 8 |
| DeepSVDD | 58.43 | 9.47 | 13.79 | 12.97 | 13 | 11 | 10 | 10 | 11.0 | 10 |
| OCSVM | 76.02 | 32.69 | 20.69 | 24.16 | 2 | 2 | 6 | 4 | 3.5 | 3 |
| CALDM | 78.68 | 43.00 | 52.87 | 52.27 | 1 | 1 | 1 | 1 | 1.0 | 1 |

For the reduced-in-size (1%) BLS compensation dataset, CALDM performed the best in all four metrics tested (AUROC, AP, Recall, F1) outperforming all 13 other models tested in every single metric tested.

### 4.3.3 Hypothesis 3

The final part of testing was related to research question three (RQ3). It was intended to measure and compare different anomaly detection systems when dealing with

data which contained rare categorical variables. The challenge for anomaly detection is that despite being “anomalies” with respect to the categories, they may not be anomalies once reviewed within its context on the variable of interest (for example pay).

**RQ3:** Can a context-aware latent diffusion model outperform other unsupervised outlier detection algorithms in AUROC and AP metrics when the target outliers are part of rare or newly introduced categories?

Hypothesis 3 (H3) addresses this research question and compares the different systems when facing data which is rare within the dataset.

**H3**: CALDMs with attention mechanisms improve overall AP and AUROC in detecting outliers when the target outliers are part of rare or newly introduced categories.

#### 4.3.3.1 Summary of Rare Category Problem

Rare categories are a special challenge in anomaly detection. Most anomaly detection systems rely on training with sufficient samples, which enables them to learn what “normal” data points look like. Thus, these systems are able to flag anomalies as points with large enough differences to warrant a review.

The problem or rare categories is not the same as a class imbalance problem (Ali et al., 2013), in which the target is a minority. For these samples the variable of interest (for example salary) may be perfectly within norm, but it may belong to a category with very samples in training. Dealing with datapoints which belong to rare categories remains a challenge of any outlier detection system, since outliers are a minority of points that differ from normality.

For class imbalance, there are several techniques like SMOTE (Chawla et al., 2002) or other over- and under-sampling techniques.

Rare category problems encompass cases where the predictors belong to categories that are underrepresented but may have a similar number of outliers as the target. One example could be Monaco and France. Monaco may have within its own population a similar proportion of pay outliers as France, but "Monaco" as a category is small in comparison, thus challenging any outlier system to learn what an outlier within Monaco looks like. Another example is new products launched where only a few normal data points may even exist. New types of fraud may be hard to detect when they originate from new or evolving business dynamics or customer behavior among many others.

Techniques to address the rare category problem have been researched, including active learning (Pelleg & Moore, 2004) and KNN-based techniques (He & Carbonell, 2007), but these either depend on human intervention or have shown limited success.

CALDM's approach to addressing the rare category problem is by the incorporation of contextualization conditioning. CALDM incorporates multi-head attention into its process which is used to generate contextual data to the reconstruction process. The principle is that rare categories may share similar characteristics with other categories, enabling CALDM to use what was learned for those more common categories as a reference to identify outliers. In the France-Monaco example, the principle would be that the attention mechanism contextualizes the Monaco rare category, learns that it is similar to South France or North Italy, and is able to correctly flag Monaco outliers if present.

#### 4.3.3.2 Results of Rare Category Test

For this praxis, four synthetic MSAs were specifically created to test CALDM's capabilities in addressing rare categories. These four MSAs comprised a small, randomly

sampled selection from existing MSAs in Florida and Alabama, with some normal noise added to ensure some differentiation. These new four MSA's were clones of four specific real MSA's. These MSA's were the top two and bottom two MSAs on average salary in Florida and Alabama.

The MSA's chose for testing were very with only a range of 9 to 122 total employees each, and 3 out of 4 with no tagged outliers.

**Table 14- Rare category test - Sample characteristics**

| MSA | Total Employees | Average Salary | Number of Outliers |
|---|---|---|---|
| **99995** | 9 | $ 38,721 | 0 |
| **99996** | 26 | $ 49,134 | 0 |
| **99997** | 1 | $ 33,313 | 0 |
| **99998** | 122 | $ 56,988 | 17 |

CALDM's performance measured only these specific MSA's was compared to the performance on the complete test dataset with all MSA's included.

**Table 15- Rare category test - Performance results**

| Metric | Result only on rare MSA's | Overall Results all MSA's | Overall Ranking |
|---|---|---|---|
| **AUROC** | 72.11 | 79.29 | 1 |
| **Average Precision (AP)** | 37.07 | 37.05 | 1 |
| **F1 Score** | 55.17 | 36.58 | 1 |
| **Recall** | 47.06 | 36.72 | 1 |

Overall CALDM again outperformed other systems when dealing with rare categories. Interestingly, AP was slightly better, and F1 and Recall much better than the results obtained with all MSA's included.

The confusion matrix illustrates in more detail how CALDM classified these samples from rare MSA's.

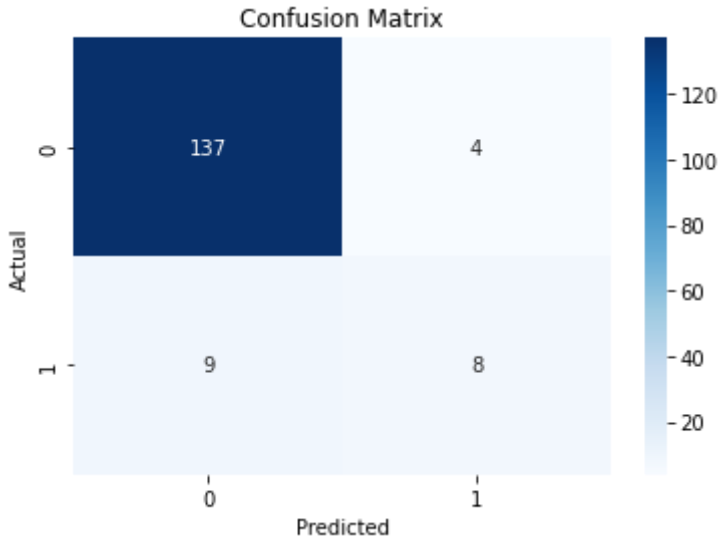



**Figure 15- Rare category test confusion matrix**

The table below shows in detail who CALDM's score each MSA's data and outliers, Showing a false positive only once.

**Table 16 - Rare category test, detailed results**

| MSA | Total Employees | Number of Outliers | Flagged as Outlier |
|---|---|---|---|
| **99995** | 9 | 0 | 0 |
| **99996** | 26 | 0 | 1 |
| **99997** | 1 | 0 | 0 |
| **99998** | 122 | 17 | 11 |

The rare category test showed comparable results to the full BLS dataset test, where a significant number (for an unsupervised method) of outliers were detected.

### 4.4 Characteristics of the Latent Space generated by CALDM

As a check of the VAE's theoretical capability to approximate the latent space using normal distributions. Code was implemented that plotted the shape of the latent space post-training.

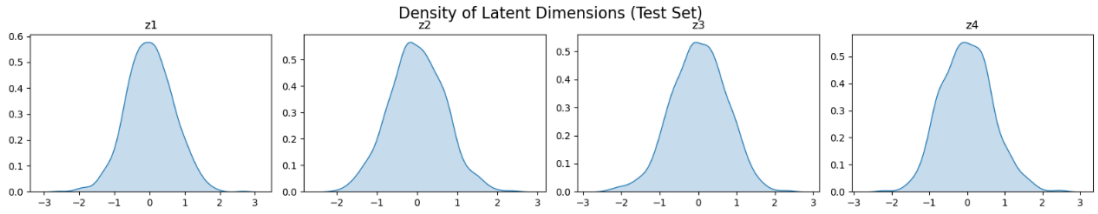



**Figure 16 - Example of  CALDM's VAE's latent space**

Figure 16 shows how CALDM, which includes in its internal structure a VAE, is able to model the original feature space into a series of gaussian distributions. This plot was used to verify the correct operation of CALDM in creating its latent space.

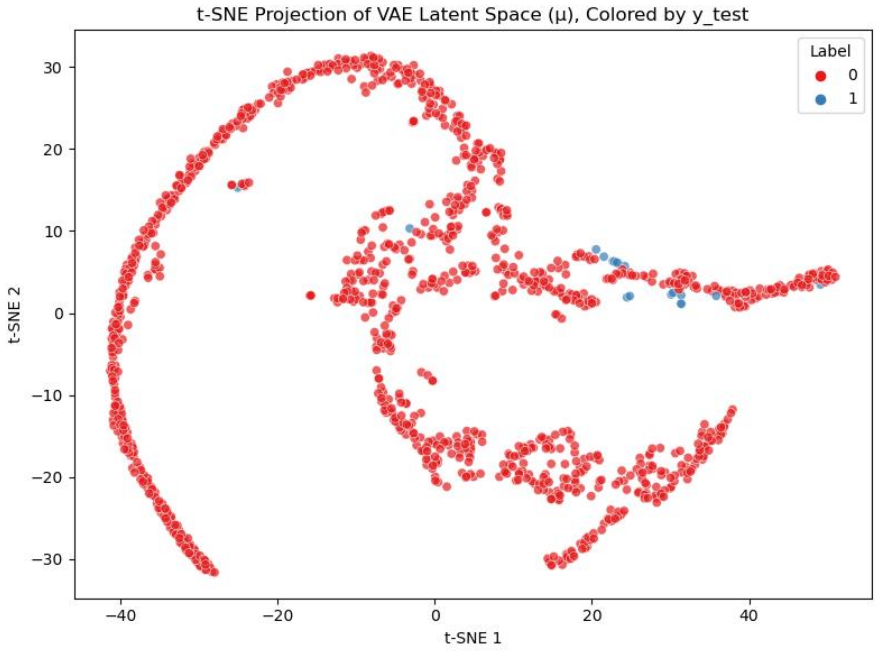



**Figure 17– PCA approximation of Latent Space**

Figure 17 shows a t-SNE (reduced dimensionality) plot of the same latent space, adding color differentiation to denote the normal and anomalous (outlier) points. In this case, the detection of outliers would be challenging using traditional methods as anomalous data points are not located far away from the rest of normal data points.

### 4.5 Summary of Key Results

Three sets of experiments were conducted to evaluate the performance of the CALDM in addressing the core research questions of this praxis document.

In the first experiment, CALDM demonstrated strong generalizability across 30 benchmark datasets from the ADBench project, achieving the highest overall ranking based on AUROC and AP, including eight first-place AUROC scores and seven first-place AP scores.

In domain-specific testing using a synthetic dataset derived from the BLS, CALDM again ranked first across all four evaluation metrics: AUROC, AP, Recall, and F1, outperforming 13 other unsupervised outlier detection models.

In the second experiment, which tested CALDM's robustness under data scarcity by using only 1% of the BLS dataset for training, the model retained its top position across all metrics.

The third experiment, the rare category test, also showed a strong performance from CALDM with an AUROC of 72.11% versus the 79.29% observed from the full dataset test. Despite being somewhat lower in AUROC, the AP was 37.07%, similar to the 37.05% observed in the full test. Finally, recall was 47.06%, also better than the 36.58% observed in the full test.

# Chapter 5—Discussion and Conclusions

## 5.1 Discussion

This praxis aimed to develop Context-Aware Latent Diffusion Model (CALDM), which merged a Variational Auto Encoder (VAE), a diffusion model, and transformer-based attention mechanisms. CALDM was originally intended to detect outliers in structured tabular compensation data but was later expanded and found effective in other low- and high-dimensionality datasets of different types and sources.

Specifically, CALDM aimed to address the challenge of outlier detection in employee compensation, where other anomaly detection techniques may underperform because the data are often largely categorical and sparse, the salary distributions are skewed, and the dependencies between these categorical features are complex.

This final chapter discusses the findings of the research, connecting them to the original research questions; interprets the results while referencing current related literature; and explores the theoretical and practical implications and applications of CALDM not only in compensation and people analytics but also in other realms of science and practice where outlier detection remains a challenge.

### 5.1.1 Alignment with Research Questions

The primary research questions aimed to determine whether a Latent Diffusion Model (LDM) enhanced with attention mechanisms could outperform traditional unsupervised outlier detection algorithms in detecting both overall outliers and outliers belonging to rare categories.

Compared to benchmark models in unsupervised outlier detection like LOF, KNN, PCA, Isolation Forest, DeepSVDD, GOAD, CBLOF, HBOS, OCSVM, ECOD,

and pure diffusion methods, CALDM showed superior performance in the wider-scope ADBench datasets and in the specific Bureau of Labor Statistics (BLS) dataset, in both overall and rare-category testing.

The second research question was related to capabilities of outlier detection models when training data were limited in size. CALDM also showed superior performance to all other algorithms selected for benchmarking when trained with a materially smaller size (1% of original) dataset, or about 4,000 rows.

The third research question was related to the capability of CALDM to detect outliers when the defining features of a sample were rare compared to other categories in the same feature. An example of rare categories would be new products, new geographies, or new types of fraud. In these cases, model training could lack an adequate number of samples to learn the defining characteristics of a normal versus an outlier sample. Algorithms that are impacted by rare categories could either classify all samples in rare categories as outliers or miss them completely, classifying them as non-outliers.

### 5.1.2 Interpretation of Key Findings

A key factor of CALDM's observed superior performance in Average Precision (AP) and Area Under the Receiver Operating Characteristic (AUROC) over the traditional unsupervised outlier detection models was the integration of a contextualizer, which utilizes an attention mechanism. The attention mechanism conditions both the VAE and the diffusion model with contextual embeddings related to the interactions between categorical features. For example, when using the BLS dataset, these contextual categories are industry, Metropolitan Statistical Area (MSA), or family of occupational codes.

Not all datasets have a deeply embedded contextual structure. The BLS dataset with several hierarchical levels of information structure provided a rich and challenging ground for models to be tested. Nevertheless, CALDM showed superior performance even in simple datasets with fewer features.

CALDM outperformed other models in the comprehensive ADBench dataset test, where simpler datasets (some of them with as little as three features) were used. This superiority can only be explained by two reasons: 1) There is always some level of contextualization potential in all datasets, which attention mechanisms are able to utilize. 2) Even when there is not much to contextualize in the dataset, attention mechanism and its produced contextual embeddings do not harm performance, and the pure LDM capabilities for outlier detection are still powerful enough to outperform other models.

The latent diffusion process introduces a robust mechanism for reconstruction-based anomaly detection. Samples showing high reconstruction loss are flagged as potential outliers. The experimental results indicated that this reconstruction-based metric provides improved sensitivity and specificity compared to distance-based or density-based methods.

Regarding the trade-off between model complexity and interpretability, while CALDM offered improved detection accuracy, its architecture is complex, and integrating three types of generative AI methods (attention, VAE, and Denoising Diffusion Probabilistic Model (DDPM) introduces a degree of model opacity. This can be an issue since model outputs often go through review from internal and external stakeholders who may not have the technical background to understand the technicalities of more complex generative models.

The main finding is that adding an attention mechanism to the LDM provides additional performance in outlier detection, particularly in settings characterized by high-dimensional, largely categorical, and sparse data.

### 5.1.3 Contrast with Traditional Methods

The core objective of this praxis was to evaluate the performance of CALDM against a broad and diverse set of existing, unsupervised outlier detection models. This evaluation was critical to determine whether the introduction of architectural choices, such as the integration of an attention mechanism into a latent diffusion process, would translate into meaningful performance improvements.

The following competing outlier detection models are the most popular for unsupervised tasks and include the major families of models based on the approach taken to identify potential outliers:

- Distance and Density-Based Methods: LOF, KNN
- Dimensionality Reduction Approaches: PCA
- Ensemble and Tree-Based Models: iForest
- Deep Learning-Based Techniques: Deep Support Vector Data Description (DeepSVDD), Generative Outlier Detection (GOAD)
- Clustering and Histogram-Based Approaches: Clustering-Based Local Outlier Factor (CBLOF), Histogram-Based Outlier Score (HBOS)
- Statistical and One-Class Methods: OCSVM, Empirical Cumulative Outlier Detection (ECOD)
- Standard Diffusion-Based Methods: Diffusion models applied to raw data without context-aware enhancements.

Overall results showed strong evidence that CALDM performs better than other unsupervised models, especially when dealing with high-dimensional sparse categorical data.

### 5.1.4 Practical Implications

The first practical implication is related to improving pay fairness. While current methods provide adequate performance detecting pay outliers when data have a simple structure, current methods struggle when data are high-dimensional, sparse, largely categorical, and have complex internal interactions that define an anomaly or potential outlier. CALDM is a more robust outlier detector capable of addressing a wider range of use cases, from simple datasets to the more complete structure found in compensation datasets, where pay is dependent on factors like geographic location, job-specific data, tenure, performance, and industry—all of which impact what the expected pay for a given performance or experience should be.

CALDM also addresses operational issues related to how some pay outliers are handled nowadays. A common practice used in outlier detection and in general pay equity reviews is to group people subjectively into cohorts and use traditional methods like standard deviations from the mean of a cohort. While the general principle is valid, in which the main interest is in identifying people within a homogeneous group who deviate from the norm of the group, this principle is highly dependent on the subjective way in which a group is defined.

If the cohort definition is too broad and the cohort is large, the cohort may not capture the subtleties embedded within such a large group that are important for anomaly

detection (e.g., create cohorts of people by state or country as opposed to smaller and more homogeneous cohorts using MSA or zip code).

If the cohort definition is too detailed, the cohorts become small in elements per cohort, and the total number of resulting cohorts would be large. Besides the operational concerns of trying to review many diverse cohorts, since the number of people is small, standard statistical measures of distance, like standard deviation, become more volatile or unreliable.

CALDM offers a more streamlined approach that works even in cases where the categorical information of a sample is unique or rare, which could make the sample of interest the only one of its kind. CALDM, through its contextualization capabilities, is able to flag the sample as normal or abnormal.

CALDM was originally developed to address the specific task of identifying compensation outliers, specifically in datasets with high dimensionality, a product of the potentially large number of categorical variables normally found in real-world datasets. After benchmarking CALDM with a much wider range of datasets and anomaly detection use cases, the strength of CALDM as a general-purpose anomaly detection algorithm for unsupervised cases became evident.

By using an unsupervised methodology, CALDM can be utilized without the complex, costly, and still-subjective labeling process required for supervised methodologies. Rather, CALDM can be used to quickly flag outliers as a complement to supervised methodologies by having a complete and more robust supervised and unsupervised anomaly detection framework.

This research showed that the enhancements built into CALDM, like the addition of an attention mechanism to the LDM, also show superior performance in all other types of datasets for outlier detection.

### 5.1.5 Limitations and Considerations

While CALDM offers significant improvements in outlier detection, there are limitations which need to be recognized and addressed by any organization reviewing CALDM as part of their outlier detection pipeline.

The first consideration is model complexity and interpretability. CALDM is an amalgamation of three generative AI architectures: an attention mechanism used for contextualization, a VAE used for latent space creation, and a DDPM used to enhance the reconstruction loss data used for outlier flagging.

These three individual component models are complex, with a deeper mathematical principle guiding them. In many other use cases, stakeholders may not be interested in internal workings, for example, when an image is generated from a user prompt. However, if there is a specific desire for the outlier detection system to be fully transparent and simple to explain, CALDM will require purpose-specific methodologies.

Computational cost and scalability. CALDM consists of three deep neural network architectures, which were trained using PyTorch and require a GPU for training and inference. For this praxis, several types of GPUs were utilized at different times, mostly driven by availability and convenience. For cloud processing, Google's powerful Colab A100 GPU was used. Locally, an RTX 3070 GPU in a personal computer with 32 GB of RAM was used. The datasets used for experiments were not larger than 1 MM rows and 200 columns. For larger datasets in rows and especially columns, the "curse of

dimensionality" (Bellman & Kalaba, 1957) may kick in. A single GPU and current memory would have quickly proven to be insufficient. Groups considering adopting CALDM should evaluate the potential size of their datasets and evaluate the internal GPU and memory capabilities, as well as the incremental cost of scaling up versus the benefits of greater accuracy in anomaly detection.

### 5.2 Conclusions

This doctoral praxis presented the design, implementation, and evaluation of CALDM for finding outliers in complex tabular datasets in an unsupervised manner. It was originally intended to work with compensation data from large organizations but later expanded and tested with a broader range of datasets and use cases. Existing unsupervised anomaly detection methods struggle with the messy reality of HR data: rare categorical combinations, contextual differences, and the complex distributions found in real compensation systems.

The performance improvements observed with CALDM with the broader outlier datasets from the ADBench benchmarking project revealed through internal patterns and interactions within other datasets that CALDM was able to learn with better results than other available unsupervised methods.

CALDM consistently beat a wide range of baseline unsupervised methods, specifically LOF, KNN, iForest, PCA, DeepSVDD, GOAD, HBOS, CBLOF, ECOD, OCSVM, and standard diffusion models like TabAdm.

The key breakthrough with CALDM is the addition of an attention mechanism to learn contextual relationships between variables. This is particularly important in outlier detection, as many outliers are so dependent on the context in which they live.  CALDM

can harvest this and add an additional level of performance with those difficult to detect and more nuanced outliers.

CALDM offers a scalable, context-aware solution for outlier detection with improved performance over traditional methods. This research also demonstrates the potential of generative methods mostly associated with image and text generation to other use cases**.**

### 5.3 Contributions to Body of Knowledge

This doctoral research contributes to the body of knowledge in machine learning, generative modeling, and HR analytics in several distinct and original ways:

1. Integration of Context into Latent Diffusion for Tabular data

This work presents the first known application of an LDM that is explicitly conditioned on learned context embeddings derived from categorical metadata (MSA, industry, occupation code, performance) in a tabular setting. While diffusion models have been explored in image and language domains, this praxis advances the field by adapting these models for structured, mixed-type business data and showing that contextual conditioning significantly improves anomaly detection performance.

2. Unified Architecture Combining VAE, Context Encoding, and Diffusion

The research proposes an improved three-stage architecture that integrates a VAE for latent compression, a transformer-based encoder for contextual embedding, and a DDPM for reconstruction-based anomaly calculation. This layered but integrated design is a new contribution to generative modeling approaches for anomaly detection.

3. Advancing Unsupervised Anomaly Detection in Sparse Categorical Cohorts

Traditional unsupervised anomaly detection methods often underperform in rare categorical combinations, such as small MSA-industry-occupational code segments in HR data. CALDM introduces a learned approach to context-sensitive outlier detection that does not require manually defined groupings or pre-specified subjective rules.

4. Demonstrating the Practical Utility of Latent Diffusion Beyond Image/Text Domains

Normally LDMs have been associated with image generation from natural language with impressive results. This praxis demonstrated a non-traditional application of an LDM for outlier detection, emphasizing the incredible potential of generative AI to solve a wider range of problems.

5. Comprehensive Benchmark Against Existing Anomaly Detectors

This research offers a comprehensive benchmark of a wide range of outlier detection models under a large list of datasets. CALDM was compared to over 10 existing detection methods (KNN, LOF, PCA, iForest, DeepSVDD, ECOD, CBLOF, and others), using 31 different datasets. This test can be used as a baseline for future improvements in outlier detection when future technologies emerge, or new theories are tested.

**5.4** Recommendations For Future Research

There are multiple places to refine and extend this work. The following recommendations are offered for researchers interested in improving upon or adapting the CALDM framework:

1. Explore Alternative Methods to include Attention Mechanisms within the Latent Diffusion Model

The current implementation of CALDM uses a transformer-based encoder to generate context embeddings from categorical features. These embeddings are concatenated with the data input features and passed through the VAE and diffusion pipeline, effectively acting as a form of contextual conditioning. However, the integration of attention can be further expanded in several directions, one of which was considered for this praxis: Self-Attention on Input Features. Instead of conditioning context externally, models like TabTransformer (Huang et al., 2020) use self-attention directly on tabular columns to capture inter-feature dependencies. This could allow the model to learn feature-level relationships that are context specific.

2. Investigate Alternative Neural Network Architectures for the VAE and Diffusion Components

The current CALDM implementation relies on feed-forward, fully connected neural networks for both the encoder/decoder (in the VAE) and the denoising model. While CALDM was very effective when tested for this praxis, several other neural architectures should be explored, as they may yield better results:

- Residual Networks (ResNets; (He et al., 2016): This is a common architecture used when the LDM's main purpose is image generation. Due to the tabular nature of the data used to benchmark CALDM, it was decided early in this Praxis to utilize a simpler architecture in lieu of a ResNet.

- Graph Neural Networks (GNNs); (Kipf & Welling, 2017), (Scarselli et al., 2009): When specific relationships exist, like job hierarchies GNNs can capture dependencies.
- Transformer-based Denoising Networks: As seen in diffusion transformers (Chen et al., 2024; Hatamizadeh et al., 2024; Peebles & Xie, 2023), replacing MLP's with attention-based denoisers may improve results
- Normalizing Flows (Rezende & Mohamed, 2015) : Replacing the VAE decoder with flow models can improve the quality of the output.

3. Improved Interpretability of Model Outputs

As CALDM is a more complex model than more traditional models, it is crucial to improve model explainability. Things to explore include:

- Feature attribution methods like SHAP
- Attention layers visualizations.
- Surrogate models to approximate CALDM outputs using decision trees.

4. Semi-Supervised and Human-in-the-Loop Extensions

While CALDM is fully unsupervised, there is potential for further research by incorporating semi-supervised methods. This could include the following:

- Integrating contrastive learning or weak supervision to better separate normal and anomalous clusters in latent space
- Using active learning to prioritize uncertain cases for review and label acquisition
- Incorporating feedback from compensation experts to refine decision thresholds or update context embeddings

- Such extensions would make CALDM more adaptive, auditable, and aligned with enterprise needs where human oversight is mandatory.

The development of this Praxis was challenging yet incredibly rewarding. What started as a specific problem-solving process for compensation expanded into a more relevant, general-purpose anomaly detection model.

Generative AI has the potential for a wider range of applications beyond what is currently available. As researchers and practitioners expand the applications of GenAI to existing problems, new and exciting avenues of research and development will open.

## References


Aggarwal, C. C. (2017). *Outlier Analysis, 2nd Edition.* Springer. https://doi.org/10.1007/978-3-319-47578-3

Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2013). Classification with class imbalance problem. *International Journal of Advances in Soft Computing and Its Applications,, 5*(3), 176-204. https://www.researchgate.net/publication/288228469_Classification_with_class_imbalance_problem_A_review

Allen, A. L. (2012). An ethical duty to protect one's own information privacy? *Alabama Law Review, 64*, 845-866. https://scholarship.law.upenn.edu/faculty_scholarship/451/

An, J., & Cho, S. (2015). *Variational autoencoder based anomaly detection using reconstruction probability* Special lecture on IE, http://dm.snu.ac.kr/static/docs/TR/SNUDM-TR-2015-03.pdf

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data (3rd ed.)*.

Bellman, R., & Kalaba, R. (1957). Dynamic programming and statistical communication theory. *Proceedings of the National Academy of Sciences, 43*(8), 749-751. https://doi.org/10.1073/pnas.43.8.749

Bhosale, A., Mukherjee, S., Banerjee, B., & Cuzzolin, F. (2024). Anomaly detection using Diffusion-based methods. *arXiv preprint arXiv:2412.07539*. https://arxiv.org/pdf/2412.07539

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). *LOF: identifying density-based local outliers* Proceedings of the 2000 ACM SIGMOD international conference on Management of data, https://doi.org/10.1145/342009.335388

Bučková, J., & Przyłuska-Schmitt, J. (2025). The Use of Generative Artificial Intelligence in Human Resource Management: Transformation of HR Processes and Challenges for the Future. In *Data-Centric Business and Applications: Advancing Success Through Operational Excellence, Financial Innovation, Digital Transformation, and Data-Driven Human Resource Management* (pp. 361-372). Springer. https://doi.org/10.1007/978-3-031-89718-4_29

California. (2018). *California Consumer Privacy Act of 2018 (CCPA)*. State of California. https://oag.ca.gov/privacy/ccpa

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR), 41*(3). https://doi.org/10.1145/1541880.1541882

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321-357. https://doi.org/10.1613/jair.953

Chen, S., Xu, M., Ren, J., Cong, Y., He, S., Xie, Y., Sinha, A., Luo, P., Xiang, T., & Perez-Rua, J.-M. (2024). GenTron: Diffusion transformers for image and video generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, https://doi.org/10.1109/CVPR52733.2024.00616

Chen, Z., Yeo, C. K., Lee, B. S., & Lau, C. T. (2018). Autoencoder-based network anomaly detection. 2018 Wireless telecommunications symposium (WTS), https://doi.org/10.1109/WTS.2018.8363930

Cheng, Z., Zou, C., & Dong, J. (2019). Outlier detection using isolation forest and local outlier factor. Proceedings of the conference on research in adaptive and convergent systems, https://doi.org/10.1145/3338840.3355641

Correia, L., Goos, J.-C., Klein, P., Bäck, T., & Kononova, A. V. (2023). *MA-VAE: Multi-head Attention-based Variational Autoencoder Approach for Anomaly Detection in Multivariate Time-series Applied to Automotive Endurance Powertrain Testing* International Conference on Neural Computation Theory and Applications, https://doi.org/10.48550/arXiv.2309.02253

Deecke, L., Vandermeulen, R., Ruff, L., Mandt, S., & Kloft, M. (2019). Image anomaly detection with generative adversarial networks. In F. B. M. Berlingerio, T. Gärtner, N. Hurley, & G. Ifrim, Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18, https://doi.org/10.1007/978-3-030-10925-7_1

EU. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation)*. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2016/679/oj

EU. (2023). *Directive (EU) 2023/970 on strengthening the application of the principle of equal pay through pay transparency and enforcement mechanisms*. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32023L0970

Fitz-Enz, J., & Davison, B. (2002). *How to measure human resources management*. McGraw-Hill.

Foorthuis, R. (2021). On the nature and types of anomalies: a review of deviations in data. *International journal of data science and analytics, 12*(4), 297-331. https://doi.org/10.1007/s41060-021-00265-1

Goldstein, M., & Dengel, A. (2012). *Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm* 35th German Conference on Artificial Intelligence, https://www.researchgate.net/publication/231614824_Histogram-based_Outlier_Score_HBOS_A_fast_Unsupervised_Anomaly_Detection_Algorithm

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative Adversarial Networks* https://doi.org/10.48550/arxiv.1406.2661

Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. Advances in Neural Information Processing Systems, https://proceedings.neurips.cc/paper/2021/hash/9d86d83f925f2149e9edb0ac3b49229c-Abstract.html

Graham, M. S., Pinaya, W. H., Tudosiu, P.-D., Nachev, P., Ourselin, S., & Cardoso, J. (2023). Denoising diffusion models for out-of-distribution detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, https://doi.org/10.1109/CVPRW59228.2023.00297

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *Annals of Mathematical Statistics, 21*(1), 27-58. https://doi.org/10.1214/aoms/1177729885

Han, S., Hu, X., Huang, H., Jiang, M., & Zhao, Y. (2022). Adbench: Anomaly detection benchmark. *Advances in Neural Information Processing Systems, 35*, 32142-32159. https://proceedings.neurips.cc/paper_files/paper/2022/hash/cf93972b116ca5268827d575f2cc226b-Abstract-Datasets_and_Benchmarks.html

Hatamizadeh, A., Song, J., Liu, G., Kautz, J., & Vahdat, A. (2024). Diffit: Diffusion vision transformers for image generation. European Conference on Computer Vision, https://doi.org/10.1007/978-3-031-73242-3_3

Hawkins, D. (1980). *Identification of outliers*. Chapman and Hall.

He, H., Zhang, J., Chen, H., Chen, X., Li, Z., Chen, X., Wang, Y., Wang, C., & Xie, L. (2024). A diffusion-based framework for multi-class anomaly detection. Proceedings of the AAAI Conference on Artificial Intelligence, https://doi.org/10.1609/aaai.v38i8.28690

He, J., & Carbonell, J. (2007). Nearest-neighbor-based active learning for rare category detection. Advances in Neural Information Processing Systems, https://proceedings.neurips.cc/paper/2007/hash/2838023a778dfaecdc212708f721b788-Abstract.html

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, https://doi.org/10.1109/CVPR.2016.90

He, Z., Deng, S., & Xu, X. (2005). An optimization model for outlier detection in categorical data. International conference on intelligent computing, https://doi.org/10.1007/11538059_42

He, Z., Xu, X., & Deng, S. (2003). Discovering cluster-based local outliers. *Pattern recognition letters, 24*(9-10), 1641-1650. https://doi.org/10.1016/S0167-8655(03)00003-5

Healy, G., & Ahamed, M. M. (2019). Gender pay gap, voluntary interventions and recession: the case of the British financial services sector. *British Journal of Industrial Relations, 57*(2), 302-327. https://doi.org/10.1111/bjir.12448

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

Hu, T., Zhang, J., Yi, R., Du, Y., Chen, X., Liu, L., Wang, Y., & Wang, C. (2024). Anomalydiffusion: Few-shot anomaly image generation with diffusion model.

Proceedings of the AAAI conference on artificial intelligence, https://doi.org/10.1609/aaai.v38i8.28696

Huang, X., Khetan, A., Cvitkovic, M., & Karnin, Z. (2020). Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*. https://arxiv.org/pdf/2012.06678

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution* (pp. 492-518). Springer. https://doi.org/10.1007/978-1-4612-4380-9_35

Jiang, J., Xia, G. G., Carlton, D. B., Anderson, C. N., & Miyakawa, R. H. (2020). Transformer vae: A hierarchical model for structure-aware and interpretable music representation learning. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), https://doi.org/10.1109/ICASSP40776.2020.9054554

Kaufman, B. E. (2014). The historical development of American HRM broadly viewed. *Human Resource Management Review, 24*(3), 196-218. https://doi.org/10.1016/j.hrmr.2014.03.003

Kingma, D. P., & Welling, M. (2014). *Auto-Encoding Variational Bayes* International Conference on Learning Representations, https://doi.org/10.48550/arxiv.1312.6114

Kipf, T. N., & Welling, M. (2017). *Semi-Supervised Classification with Graph Convolutional Networks* International Conference on Learning Representations, https://openreview.net/pdf?id=SJU4ayYgl

Kotelnikov, A., Baranchuk, D., Rubachev, I., & Babenko, A. (2023). *TabDDPM: Modelling Tabular Data with Diffusion Models* Proceedings of the 40th International Conference on Machine Learning (ICML 2023), http://proceedings.mlr.press/v202/kotelnikov23a.html

Kriegel, H.-P., Kröger, P., Schubert, E., & Zimek, A. (2009). Outlier detection in axis-parallel subspaces of high dimensional data. Advances in Knowledge Discovery and Data Mining: 13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 Proceedings 13, https://doi.org/10.1007/978-3-642-01307-2_86

Kumah, P., Nketia, I. K., Yaokumah, W., & Asante-Offei, K. O. (2024). Exploring the Application of Generative AI in Human Resource Management. *Generative AI for Transformational Management*, 51-82.

Kumar, K., Chakraborty, S., Mahapatra, D., Bozorgtabar, B., & Roy, S. (2025). Self-Supervised Anomaly Segmentation via Diffusion Models with Dynamic Transformer UNet. 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), https://openaccess.thecvf.com/content/WACV2025/papers/Kumar_Self-Supervised_Anomaly_Segmentation_via_Diffusion_Models_with_Dynamic_Transformer_UNet_WACV_2025_paper.pdf

Lavin, A., & Ahmad, S. (2015). Evaluating real-time anomaly detection algorithms--the Numenta anomaly benchmark. 2015 IEEE 14th international conference on machine learning and applications (ICMLA), https://doi.org/10.1109/ICMLA.2015.141

Leutwiler, J., & Kleiner, B. H. (2003). Statistical analysis for determining pay discrimination. *Equal Opportunities International, 22*(6/7), 1-8. https://doi.org/10.1108/02610150310787559

Li, J., Gao, Y., Lu, J., Fang, J., Wen, C., Lin, H., & Wang, X. (2025). *DiffGAD: A Diffusion-based Unsupervised Graph Anomaly Detector* Proceedings of the 13th International Conference on Learning Representations (ICLR 2025), https://openreview.net/forum?id=PxTIG12RRHS

Li, Z., Zhao, Y., Botta, N., Ionescu, C., & Hu, X. (2020). Copod: copula-based outlier detection. 2020 IEEE international conference on data mining (ICDM), https://doi.org/10.1109/ICDM50108.2020.00135

Li, Z., Zhao, Y., Hu, X., Botta, N., Ionescu, C., & Chen, G. H. (2022). Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE transactions on knowledge and data engineering, 35*(12), 12181-12193. https://doi.org/10.1109/TKDE.2022.3159580

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. 2008 eighth ieee international conference on data mining, https://doi.org/10.1109/ICDM.2008.17

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2019). Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering, 31*(12), 2346-2363. https://doi.org/10.1109/TKDE.2018.2876857

Marler, J. H., & Boudreau, J. W. (2017). An evidence-based review of HR Analytics. *The International Journal of Human Resource Management, 28*(1), 3-26. https://doi.org/10.1080/09585192.2016.1244699

Massey Jr., F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association, 46*(253), 68-78. https://doi.org/10.1080/01621459.1951.10500769

McCoy, M. S., Allen, A. L., Kopp, K., Mello, M. M., Patil, D., Ossorio, P., Joffe, S., & Emanuel, E. J. (2023). Ethical responsibilities for companies that process personal data. *The American Journal of Bioethics, 23*(11), 11-23. https://doi.org/10.1080/15265161.2023.2209535

Micenková, B., McWilliams, B., & Assent, I. (2015). Learning Representations for Outlier Detection on a Budget. https://arxiv.org/abs/1507.08104

Micenková, B., Ng, R. T., Dang, X.-H., & Assent, I. (2013). Explaining outliers by subspace separability. 2013 IEEE 13th international conference on data mining, https://doi.org/10.1109/ICDM.2013.132

Nazabal, A., Olmos, P. M., Ghahramani, Z., & Valera, I. (2020). Handling incomplete heterogeneous data using VAEs. *Pattern Recognition, 107*, Article 107501. https://doi.org/10.1016/j.patcog.2020.107501

Niu, Z., Yu, K., & Wu, X. (2020). LSTM-based VAE-GAN for time-series anomaly detection. *Sensors, 20*(13), Article 3738. https://doi.org/10.3390/s20133738

Pang, G., Cao, L., & Chen, L. (2021). Homophily outlier detection in non-IID categorical data. *Data Mining and Knowledge Discovery, 35*(4), 1163-1224. https://doi.org/10.1007/s10618-021-00750-y

Pang, G., Shen, C., Cao, L., & Hengel, A. V. D. (2021). Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR), 54*(2), 1-38. https://doi.org/10.1145/3439950

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., & Kim, Y. (2018). Data synthesis based on generative adversarial networks. *Proc. VLDB Endow., 11*(10), 1071–1083. https://doi.org/10.14778/3231751.3231757

Peebles, W., & Xie, S. (2023). Scalable diffusion models with transformers. Proceedings of the IEEE/CVF international conference on computer vision, https://doi.org/10.1109/ICCV51070.2023.00387

Pelleg, D., & Moore, A. (2004). *Active learning for anomaly and rare-category detection* Advances in Neural Information Processing Systems, https://proceedings.neurips.cc/paper/2004/hash/8c59fd6fbe0e9793ec2b27971221cace-Abstract.html

Pevný, T. (2016). Loda: Lightweight on-line detector of anomalies. *Machine Learning, 102*(2), 275-304. https://doi.org/10.1007/s10994-015-5521-0

Rätsch, G., Mika, S., Schölkopf., B., & Muller, K.-R. (2002). Constructing boosting algorithms from SVMs: An application to one-class classification. *IEEE*

*Transactions on Pattern Analysis and Machine Intelligence, 24*(9), 1184-1199. https://doi.org/10.1109/TPAMI.2002.1033211

Rezende, D., & Mohamed, S. (2015). Variational inference with normalizing flows. Proceedings of the 32nd International Conference on Machine Learning, http://proceedings.mlr.press/v37/rezende15.html

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models.*Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, June 19, 2022 - June 24, 2022, New Orleans, LA, United states. https://doi.org/10.1109/CVPR52688.2022.01042

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., & Kloft, M. (2018). Deep one-class classification. *Proceedings of the 35th* International conference on machine learning, Stockholm, Sweden. http://proceedings.mlr.press/v80/ruff18a.html

Saouabe, A., Oualla, H., Mazar, M., & Messadia, M. (2024). Evolution of Human Resources Management Methods: From Traditional Approaches to Generative Artificial Intelligence. 2024 International Conference on Decision Aid Sciences and Applications (DASA), https://doi.org/10.1109/DASA63656.2024.10836655

Sattarov, T., Schreyer, M., & Borth, D. (2025). Diffusion-Scheduled Denoising Autoencoders for Anomaly Detection in Tabular Data. Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, https://doi.org/10.1145/3711896.3736910

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). The graph neural network model. *IEEE transactions on neural networks, 20*(1), 61-80. https://doi.org/10.1109/TNN.2008.2005605

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation, 13*(7), 1443-1471. https://doi.org/10.1162/089976601750264965

Shi, J., Xu, M., Hua, H., Zhang, H., Ermon, S., & Leskovec, J. (2024). Tabdiff: a multi-modal diffusion model for tabular data generation. *arXiv e-prints*, arXiv: 2410.20626. https://arxiv.org/pdf/2410.20626

Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., & Chang, L. (2003). *A novel anomaly detection scheme based on principal component classifier* Proceedings of the IEEE foundations and new directions of data mining workshop, https://www.researchgate.net/publication/228709094_A_Novel_Anomaly_Detection_Scheme_Based_on_Principal_Component_Classifier

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. Proceedings of the 32nd International conference on machine learning, Lille, France. http://proceedings.mlr.press/v37/sohl-dickstein15.html

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). *Score-Based Generative Modeling through Stochastic Differential Equations* Proceedings of the 9th International Conference on Learning Representations (ICLR 2021), https://openreview.net/forum?id=PxTIG12RRHS

Talagala, P. D., Hyndman, R. J., & Smith-Miles, K. (2021). Anomaly detection in high-dimensional data. *Journal of Computational and Graphical Statistics, 30*(2), 360-374. https://doi.org/10.1080/10618600.2020.1807997

Tang, J., Chen, Z., Fu, A. W.-C., & Cheung, D. W. (2002). Enhancing effectiveness of outlier detections for low density patterns. Advances in knowledge discovery and data mining: 6th Pacific-Asia conference, PAKDD 2002 Taipei, Taiwan, May 6–8, 2002 proceedings 6, https://doi.org/10.1007/3-540-47887-6_53

UK. (2017). *Equality Act 2010 (Gender Pay Gap Information) Regulations 2017, S.I. 2017/172*. UK Government. https://www.legislation.gov.uk/uksi/2017/172/contents

USA. (1963). *Equal Pay Act of 1963, 29 U.S.C. § 206(d)*. U.S. Congress. https://www.eeoc.gov/statutes/equal-pay-act-1963

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need* Advances in Neural Information Processing Systems, https://papers.nips.cc/paper/7181-attention-is-all-you-need

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). *Modeling Tabular data using Conditional GAN* Advances in Neural Information Processing Systems 32, https://dl.acm.org/doi/10.5555/3454287.3454946

Zamberg, G., Salhov, M., Lindenbaum, O., & Averbuch, A. (2023). TabADM: Unsupervised Tabular Anomaly Detection with Diffusion Models. *arXiv preprint arXiv:2307.12336*. https://arxiv.org/pdf/2307.12336

Zhao, Y., Nasrullah, Z., & Li, Z. (2019). PyOD: A Python toolbox for scalable outlier detection. *Journal of machine learning research, 20*(96), 1-7. http://jmlr.org/papers/v20/19-011.html

# Appendix A

## BLS RAW Dataset – Column Descriptions

| Field | Field Description |
|---|---|
| area | U.S. (99), state FIPS code, Metropolitan Statistical Area (MSA) or New England City and Town Area (NECTA) code, or OEWS-specific nonmetropolitan area code |
| area_title | Area name |
| area_type | Area type: 1= U.S.; 2 = State; 3 = U.S. Territory; 4 = Metropolitan Statistical Area (MSA) or New England City and Town Area (NECTA); 6 = Nonmetropolitan Area |
| prim_state | The primary state for the given area. “U.S.” is used for the national estimates. |
| naics | North American Industry Classification System (NAICS) code for the given industry |
| naics_title | North American Industry Classification System (NAICS) title for the given industry |
| i_group | Industry level. Indicates cross-industry or NAICS sector, 3-digit, 4-digit, 5-digit, or 6-digit industry. For industries that OEWS no longer publishes at the 4-digit NAICS level, the “4-digit” designation indicates the most detailed industry breakdown available: either a standard NAICS 3-digit industry or an OEWS-specific combination of 4-digit industries. Industries that OEWS has aggregated to the 3-digit NAICS level (for example, NAICS 327000) will appear twice, once with the “3-digit” and once with the “4-digit” designation. |
| own_code | Ownership type: 1= Federal Government; 2 = State Government; 3 = Local Government; 123 = Federal, State, and Local Government; 235 =Private, State, and Local Government; 35 = Private and Local Government; 5 = Private; 57 = Private, Local Government Gambling Establishments (Sector 71), and Local Government Casino Hotels (Sector 72); 58 = Private plus State and Local Government Hospitals; 59 = Private and Postal Service; 1235 = Federal, State, and Local Government and Private Sector |
| occ_code | The 6-digit Standard Occupational Classification (SOC) code or OEWS-specific code for the occupation |
| occ_title | SOC title or OEWS-specific title for the occupation |
| o_group | SOC occupation level. For most occupations, this field indicates the standard SOC major, minor, broad, and detailed levels, in addition to all-occupations totals. For occupations that OEWS no longer publishes at the SOC detailed level, the “detailed” designation indicates the most detailed data available: either a standard SOC broad occupation or an OEWS-specific combination of detailed occupations. Occupations that OEWS has aggregated to the SOC broad occupation level will appear in the file twice, once with the “broad” and once with the “detailed” designation. |
| tot_emp | Estimated total employment rounded to the nearest 10 (excludes self-employed). |

| | |
|---|---|
| emp_prse | Percent relative standard error (PRSE) for the employment estimate. PRSE is a measure of sampling error, expressed as a percentage of the corresponding estimate. Sampling error occurs when values for a population are estimated from a sample survey of the population, rather than calculated from data for all members of the population. Estimates with lower PRSEs are typically more precise in the presence of sampling error. |
| jobs_1000 | The number of jobs (employment) in the given occupation per 1,000 jobs in the given area. Only available for the state and MSA estimates; otherwise, this column is blank. |
| loc quotient | The location quotient represents the ratio of an occupation's share of employment in a given area to that occupation's share of employment in the U.S. as a whole. For example, an occupation that makes up 10% of employment in a specific metropolitan area compared with 2% of U.S. employment would have a location quotient of 5 for the area in question. Only available for the state, metropolitan area, and nonmetropolitan area estimates; otherwise, this column is blank. |
| pct_total | Percent of industry employment in the given occupation. Percents may not sum to 100 because the totals may include data for occupations that could not be published separately. Only available for the national industry estimates; otherwise, this column is blank. |
| pct_rpt | Percent of establishments reporting the given occupation for the cell. Only available for the national industry estimates; otherwise, this column is blank. |
| h_mean | Mean hourly wage |
| a_mean | Mean annual wage |
| mean_prse | Percent relative standard error (PRSE) for the mean wage estimate. PRSE is a measure of sampling error, expressed as a percentage of the corresponding estimate. Sampling error occurs when values for a population are estimated from a sample survey of the population, rather than calculated from data for all members of the population. Estimates with lower PRSEs are typically more precise in the presence of sampling error. |
| h_pct10 | Hourly 10th percentile wage |
| h_pct25 | Hourly 25th percentile wage |
| h_median | Hourly median wage (or the 50th percentile) |
| h_pct75 | Hourly 75th percentile wage |
| h_pct90 | Hourly 90th percentile wage |
| a_pct10 | Annual 10th percentile wage |
| a_pct25 | Annual 25th percentile wage |
| a_median | Annual median wage (or the 50th percentile) |
| a_pct75 | Annual 75th percentile wage |
| a_pct90 | Annual 90th percentile wage |

| | |
|---|---|
| annual | Contains “TRUE” if only annual wages are released. The OEWS program releases only annual wages for some occupations that typically work fewer than 2,080 hours per year, but are paid on an annual basis, such as teachers, pilots, and athletes. |
| hourly | Contains “TRUE” if only hourly wages are released. The OEWS program releases only hourly wages for some occupations that typically work fewer than 2,080 hours per year and are paid on an hourly basis, such as actors, dancers, and musicians and singers. |

## Notes:

* = indicates that a wage estimate is not available

** = indicates that an employment estimate is not available

# = indicates a wage equal to or greater than $115.00 per hour or $239,200 per year

~ = indicates that the percentage of establishments reporting the occupation is less than 0.5%

# Appendix B

## General Statistics by State

| State | Employees | Mean Salary | Median Salary | Std Deviation of Salary | Outlier Rate |
|---|---|---|---|---|---|
| DC | 3,157,800 | 85,140 | 68,406 | 54,751 | 7.73 |
| MA | 3,760,648 | 77,001 | 60,210 | 49,241 | 6.90 |
| WA | 3,392,746 | 76,747 | 60,715 | 48,086 | 6.67 |
| NY | 11,735,501 | 75,785 | 56,492 | 52,870 | 7.25 |
| CA | 17,937,802 | 74,905 | 54,918 | 52,749 | 7.10 |
| CT | 1,662,872 | 71,594 | 56,030 | 48,773 | 6.52 |
| NJ | 3,380,199 | 71,095 | 54,466 | 46,496 | 6.41 |
| CO | 2,814,920 | 71,001 | 56,070 | 44,616 | 6.11 |
| MD | 2,462,028 | 70,289 | 54,134 | 47,136 | 6.33 |
| MN | 2,771,803 | 67,074 | 53,604 | 44,866 | 6.09 |
| OR | 1,942,631 | 66,761 | 52,446 | 43,503 | 5.86 |
| AK | 284,950 | 66,731 | 56,279 | 38,073 | 5.53 |
| IL | 5,882,990 | 65,662 | 50,962 | 42,376 | 5.78 |
| RI | 598,377 | 64,649 | 50,869 | 42,104 | 5.55 |
| VA | 3,512,617 | 64,073 | 49,662 | 43,491 | 5.76 |
| NH | 573,071 | 63,589 | 51,376 | 40,690 | 5.57 |
| HI | 593,378 | 62,995 | 51,497 | 42,037 | 5.32 |
| DE | 376,159 | 62,808 | 49,140 | 41,208 | 5.68 |
| VT | 251,968 | 61,685 | 50,744 | 35,989 | 5.58 |
| AZ | 3,134,575 | 61,550 | 48,873 | 37,930 | 5.51 |
| PA | 6,160,171 | 61,361 | 48,780 | 39,739 | 5.41 |
| GA | 4,583,542 | 60,642 | 46,404 | 42,579 | 5.47 |
| TX | 13,277,299 | 60,488 | 46,726 | 42,232 | 5.52 |
| MI | 4,172,473 | 60,343 | 48,329 | 38,022 | 5.44 |
| UT | 1,669,248 | 60,090 | 48,304 | 37,503 | 5.15 |
| OH | 5,329,218 | 59,316 | 47,813 | 39,816 | 5.27 |
| NC | 4,623,382 | 59,177 | 45,893 | 41,625 | 5.36 |
| ME | 554,663 | 58,917 | 48,665 | 32,101 | 5.30 |
| MO | 3,060,541 | 58,854 | 47,190 | 38,326 | 5.31 |
| FL | 9,624,993 | 58,569 | 45,405 | 41,433 | 5.32 |
| WI | 2,694,709 | 58,403 | 48,270 | 36,073 | 5.25 |
| NE | 935,221 | 58,049 | 47,441 | 38,408 | 5.26 |
| ND | 379,881 | 57,523 | 49,678 | 30,179 | 5.10 |
| IA | 1,431,354 | 56,760 | 47,285 | 33,701 | 5.03 |
| IN | 2,923,622 | 55,861 | 45,753 | 35,860 | 5.03 |
| NM | 772,827 | 55,740 | 43,802 | 36,658 | 5.49 |

| | | | | | |
|---|---|---|---|---|---|
| WY | 225,191 | 55,727 | 47,094 | 31,846 | 5.01 |
| TN | 3,135,585 | 55,621 | 44,355 | 38,708 | 5.15 |
| NV | 1,473,211 | 55,322 | 44,134 | 36,831 | 5.03 |
| KS | 1,195,145 | 54,698 | 44,813 | 33,776 | 4.85 |
| KY | 1,770,847 | 54,689 | 45,084 | 37,786 | 5.02 |
| MT | 427,707 | 54,222 | 45,531 | 32,283 | 4.94 |
| ID | 772,125 | 53,938 | 44,404 | 34,240 | 4.70 |
| SC | 2,114,098 | 53,857 | 42,779 | 38,726 | 5.20 |
| OK | 1,551,190 | 53,598 | 43,118 | 36,173 | 4.88 |
| AL | 1,934,583 | 53,405 | 42,620 | 36,218 | 5.02 |
| LA | 1,807,894 | 53,012 | 42,830 | 37,044 | 4.80 |
| SD | 392,935 | 52,205 | 43,522 | 30,914 | 5.46 |
| WV | 626,468 | 51,220 | 41,200 | 35,959 | 5.04 |
| AR | 1,166,334 | 51,140 | 40,990 | 32,639 | 5.09 |
| MS | 988,687 | 47,481 | 39,159 | 31,898 | 4.44 |
| PR | 882,677 | 33,735 | 25,885 | 23,783 | 3.56 |

# Appendix C

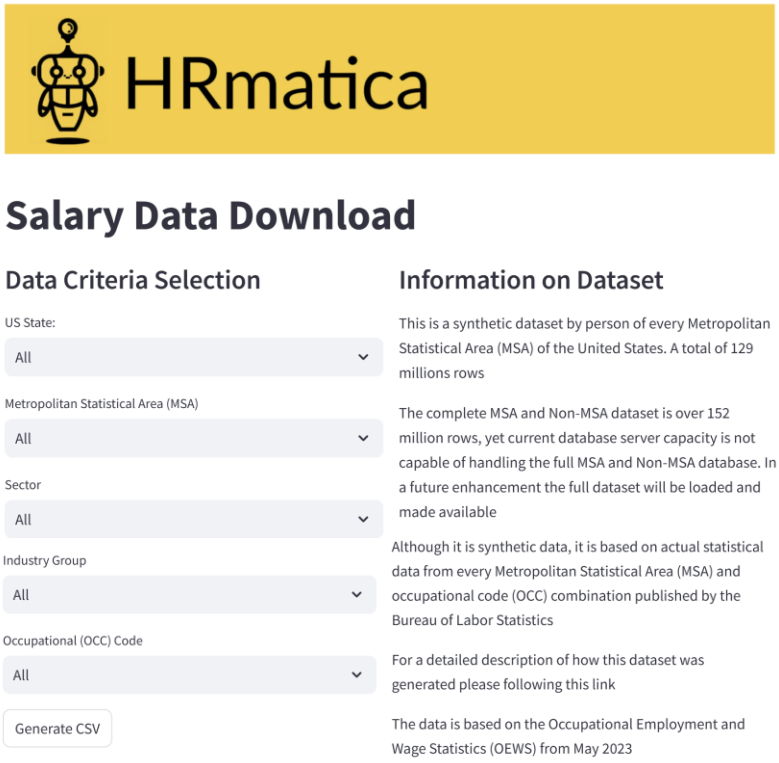
## CALDM Architecture (Parameters)

| Parameter | Contextual Encoder | | VAE Encode/Decoder | | Diffusion Model | |
|---|---|---|---|---|---|---|
| | Low-Dimensional Dataset (D<=30) | High-Dimensional Datasets (D>30) | Low-Dimensional Dataset (D<=30) | High-Dimensional Datasets (D>30) | Low-Dimensional Dataset (D<=30) | High-Dimensional Datasets (D>30) |
| Latent Dimensions | n/a | n/a | 4 | 16 | n/a | n/a |
| Contextual Dimensions | 10 | 20 | n/a | n/a | n/a | n/a |
| Feed Forward Layers | 2 | 2 | 2 | 2 | 2 | 2 |
| Nodes per Layer | 64 | 64 | 64 | 256 | 256 | 256 |
| Attention Heads | 4 | 4 | n/a | n/a | n/a | n/a |
| Layer Normalization | no | no | yes | yes | yes | yes |
| Dropout Layer rate | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| Learning Rate | 0.001 | 0.001 | 0.01 | 0.01 | 0.001 | 0.001 |
| Optimizer | Adam | Adam | Adam | Adam | Adam | Adam |
| Weight Decay | 1E-04 | 1E-04 | 1E-04 | 1E-04 | 1E-04 | 1E-04 |
| Epochs | 20 | 50 | 20 | 50 | 200 | 200 |
| Batch Size | 256 | 256 | 256 | 256 | 256 | 256 |
| Difusion Steps | n/a | n/a | n/a | n/a | 1000 | 1000 |
| Diffusion Scheduler Type | n/a | n/a | n/a | n/a | Linear | Linear |
| Positional Encoder Type | n/a | n/a | n/a | n/a | Sinusoidal | Sinusoidal |

## Appendix D

**Frontpage of public website created to access and download synthetic datasets utilized for this Praxis.**

**https://data.hrmatica.com/wagedata/**



## Appendix E

**Characteristics of the final dataset used for training and testing.**

| | |
|---|---|
| Total US States | 3 (Florida, Georgia, Alabama |
| Total Number of Rows | 16,143,118 |
| Rows selected for Training & Testing | 596,856 |
| Total Number of MSAs in Dataset | 52 |
| Total number of Job Codes (OCC Code) | 387 |
| Total number of Industries | 164 |
| Total number of Sectors/Sub-Sectors | 20/47 |
| Mean Salary | $55,502 |
| Max Salary | $1,592,918 |
| Minimum Salary | $14,810 |
| Standard Deviation of Salary | $36,156 |
| Total Number of outliers | 44,133 (7.4%) |
| Train/Test split | 80:20 |

# Appendix F

| OCC_CODE - Description | Count | Mean Salary | Median Salary | Outlier Rate |
|---|---|---|---|---|
| Physicians, All Other | 314,319 | 230,437 | 187,842 | 38.79 |
| Family Medicine Physicians | 120,539 | 227,211 | 217,384 | 41.00 |
| Airline Pilots, Copilots, and Flight Engineers | 86,351 | 225,407 | 211,008 | 32.02 |
| Psychiatrists | 27,619 | 215,509 | 193,175 | 42.04 |
| Chief Executives | 198,843 | 210,448 | 199,378 | 49.05 |
| Nurse Anesthetists | 51,894 | 209,142 | 207,402 | 79.82 |
| General Internal Medicine Physicians | 75,204 | 204,631 | 190,159 | 33.49 |
| Ophthalmologists, Except Pediatric | 20,616 | 187,328 | 151,101 | 31.31 |
| Pediatricians, General | 48,491 | 185,972 | 182,946 | 16.85 |
| Computer and Information Systems Managers | 602,502 | 177,302 | 173,690 | 18.45 |
| Dentists, All Other Specialists | 10,192 | 169,023 | 195,590 | 24.56 |
| Architectural and Engineering Managers | 212,261 | 168,487 | 166,653 | 22.56 |
| Dermatologists | 15,078 | 167,496 | 85,294 | 28.01 |
| Obstetricians and Gynecologists | 30,040 | 165,590 | 153,146 | 28.59 |
| Natural Sciences Managers | 102,370 | 164,666 | 163,187 | 16.89 |
| Neurologists | 12,701 | 163,526 | 92,002 | 22.49 |
| Financial Managers | 801,760 | 163,483 | 160,976 | 20.42 |
| Dentists, General | 125,472 | 160,917 | 160,955 | 20.98 |
| Marketing Managers | 367,684 | 158,842 | 157,416 | 18.04 |
| Orthopedic Surgeons, Except Pediatric | 19,560 | 156,700 | 80,065 | 25.94 |