Unsupervised Outlier Detection on Thyroid Data

HKUST(GZ)

Yuqi Ren (Member A)
Anyi Wang (Member B)
Feng Liang (Member C)

1 Introduction

Outlier detection is a fundamental tool for discovering rare and abnormal patterns in real-world data, especially in medical settings where reliable labels are scarce and class imbalance is severe. Thyroid dysfunction (e.g., hyperthyroidism and hypothyroidism) is often reflected by atypical hormone profiles, including abnormal TSH, T3, TT4, T4U, and FTI measurements. Such abnormalities are clinically meaningful and often appear as rare cases in population data, which makes thyroid datasets a natural benchmark for **unsupervised** anomaly detection.

From a practical perspective, an unsupervised anomaly detector can serve as a **triage and quality-control tool** in clinical pipelines: it can rank a large cohort and surface a small set of suspicious cases for further inspection, helping prioritize medical review when manual screening is costly. In addition, extremely unusual measurements may also indicate measurement artifacts or preprocessing issues, so anomaly detection is also useful for **medical data auditing**.

In this project, we study the **Annthyroid Unsupervised Anomaly Detection** dataset (Kaggle version) and implement an end-to-end pipeline covering preprocessing, multiple detectors, quantitative evaluation (offline labels only), visualization, robustness analysis, and clinically grounded interpretation. Importantly, we also identify that the Kaggle-provided hormone measurements exhibit **inconsistent and undocumented scaling**, which can obscure physiological meaning. Therefore, we additionally reference the original **UCI Thyroid Disease dataset** to restore clinically interpretable scales and build a cleaned version of the dataset.

Our pipeline is designed to align with all experiments in this report:

- **Data preparation:** remove empty columns, reduce heavy-tailed distributions with $\log(1+x)$, and standardize features for stable model behavior.
- **Unsupervised detectors:** build Isolation Forest (baseline) and an improved Feature-Weighted Isolation Forest (FWIF); implement LOF and an autoencoder (AE) as contrastive baselines.

- **Evaluation and visualization:** evaluate models using ROC-AUC, PR-AUC, and Precision@K (offline labels only); visualize anomaly-score landscapes in 2D PCA space and highlight top-K anomalies.
- **Robustness and clinical grounding:** test ranking stability via bootstrap resampling and random-seed sensitivity; additionally rescale hormone features using UCI reference ranges and explore contextual ensemble detection motivated by clinical subgroups.
- **Interpretation:** provide an empirical medical summary of what detected outliers likely represent (hormone extremes, treatment/diagnostic flags, and extreme isolated cases).

Member Contributions.

- **Member A:** preprocessing, IF baseline, FWIF improvement, PCA visualization, robustness tests.
- **Member B:** LOF and AE baselines, evaluation metrics (ROC-AUC/PR-AUC/Precision@K), PCA visualizations.
- **Member C:** UCI-referenced clinical rescaling/cleaning, age normalization, contextual ensemble method.

Overall, this report provides a complete, interpretable, and robust study of unsupervised outlier detection on thyroid data, combining detector comparison with clinically informed data validation and interpretation.

2 Problem Formulation

Let $X = \{x_1, \ldots, x_n\}$ be an unlabeled dataset with $x_i \in \mathbb{R}^d$. The goal of unsupervised outlier detection is to learn an **real-valued anomaly scoring function** $s(\cdot)$ such that higher scores indicate stronger abnormality:

$$s(x_i) \in \mathbb{R}, \qquad \text{and larger } s(x_i) \text{ means more anomalous.}$$

The primary output is therefore a **ranking** of samples by $s(x)$, and we focus on retrieving the top-K most suspicious samples:

$$\mathcal{A}_K = \{x_i \mid s(x_i) \text{ is among the top-}K \text{ values}\}.$$

2.1 Dataset and Two Representations

After removing empty columns, the dataset contains $n = 6916$ samples and $d = 21$ numerical features. Offline labels indicate that approximately 3.6% of samples are outliers, making evaluation highly imbalanced. Labels are **not used for training** and are used **only for offline evaluation**.

In this report, we operate on two feature representations:

- **Processed Kaggle representation:** the provided annthyroid file used for baseline IF/FWIF and LOF experiments after standard preprocessing.
- **Clinically rescaled representation:** a cleaned version aligned to UCI thyroid reference ranges, obtained by inferring scaling factors via quantile/statistical matching; this version restores physiological units for hormone features and normalizes age for interpretability and downstream modeling.

Both representations share the same objective: produce a meaningful anomaly ranking.

2.2 Detectors and Anomaly Scores

We compare multiple unsupervised detectors that output $s(x)$:

- **Isolation Forest / FWIF:** tree-based isolation; we use $s(x) = -\texttt{score\_samples}(x)$.
- **LOF:** local density deviation; we use the negated `negative_outlier_factor_` as $s(x)$.
- **Autoencoder:** reconstruction-based; we use reconstruction MSE as $s(x)$.

For interpretability, we also visualize the score landscape using PCA projections colored by $s(x)$ and highlight the top-K anomalies.

2.3 Offline Evaluation Metrics and Operating Point

To evaluate ranking quality under severe imbalance, we report:

- **ROC-AUC** and **PR-AUC** for global ranking performance,
- **Precision@K** (we report $K = 50$) to assess top-K retrieval quality.

When a binary decision is required (e.g., confusion statistics), we select a threshold using a quantile rule based on an assumed contamination rate and report TP/FP/TN/FN to illustrate trade-offs between false alarms and missed anomalies.

3 Methodology and Experiments

3.1 Data Understanding & Preprocessing (Member A)

3.1.1 Motivation and Problem

The annthyroid dataset is a tabular medical dataset that mixes (i) hormone-related numerical measurements (e.g., TSH, T3, TT4, T4U, FTI) and (ii) multiple binary clinical flags (e.g., medication usage or diagnostic queries). In the raw file, we observed three practical issues that can harm unsupervised anomaly detectors: (i) empty columns, (ii) heterogeneous feature scales (continuous hormones vs. binary indicators), and (iii) heavy-tailed / skewed distributions in hormone measurements. Without proper preprocessing, tree splits and distance/density geometry can be dominated by scale and extreme values, leading to unstable anomaly rankings and reduced interpretability.

3.1.2 Method

We apply a minimal yet robust preprocessing pipeline:

- **Schema cleaning:** remove two empty columns (`Unnamed:22`, `Unnamed:23`), leaving $d = 21$ valid numerical features.
- **Missing-value check:** verify that no missing values remain after schema cleaning.
- **Skewness correction:** apply $\log(1 + x)$ to reduce heavy tails, especially for hormone features. This transform is monotonic; for binary flags (0/1), it preserves ordering while keeping values bounded.
- **Scaling:** standardize features via `StandardScaler` to achieve comparable scales across dimensions.

Formally, for each feature value $x \geq 0$, we transform:

$$x' = \log(1 + x), \qquad \tilde{x} = \frac{x' - \mu}{\sigma},$$

where μ and σ are computed from the dataset to normalize each feature to zero mean and unit variance.

3.1.3 Experimental Setup and Results

After preprocessing, the dataset contains $n = 6916$ samples and $d = 21$ numerical features. Offline labels indicate an outlier ratio of approximately 3.6%, which implies a highly imbalanced setting. The preprocessing produces a stable feature matrix with comparable scales (approximately zero mean and unit variance). In particular, the $\log(1 + x)$ transform compresses extreme hormone values, reducing the influence of rare but very large measurements on downstream detectors.

3.1.4 Summary and Discussion

This preprocessing pipeline addresses key practical issues for unsupervised detection: it removes non-informative columns, reduces skewness in hormone-related features, and mitigates scale sensitivity. As a result, subsequent anomaly detectors can focus on structural differences among patients rather than artifacts from inconsistent magnitudes or heavy-tailed distributions.

3.2 Baseline Isolation Forest and Feature-Weighted Improvement (FWIF) (Member A)

3.2.1 Motivation and Problem

A baseline detector is required to establish initial performance on thyroid anomaly ranking. Although Isolation Forest (IF) is efficient for tabular data and does not require labels, our baseline results show that its top-K ranking can be weak in this dataset. A plausible reason is

that random feature selection may spend many splits on weak or less informative binary flags, while clinically informative hormone features are not explicitly emphasized. We therefore aim to incorporate feature informativeness into the detector to improve top-K ranking quality.

3.2.2 Method

Baseline Isolation Forest (IF). Isolation Forest isolates anomalies by recursively partitioning data with random feature selection and random split values. Points that are isolated earlier (shorter path length) receive higher anomaly scores. We implement a baseline IF with:

- **Model:** Isolation Forest
- **Key parameters:** $n\_estimators = 200$, $max\_samples = 1024$, $contamination = 0.03$
- **Scoring:** $s(x) = -\texttt{score\_samples}(x)$ (larger means more anomalous)

Feature-Weighted Isolation Forest (FWIF). To emphasize informative dimensions, we compute a split-based feature importance from the baseline IF by aggregating how frequently each feature is used for splitting across trees (normalized to sum to 1). We then:

- Select the **top 12** most influential features.
- Convert importance into three tiers and construct a **weighted feature representation** by repetition:
  - High-importance features: $\times 3$,
  - Medium-importance features: $\times 2$,
  - Low-importance features: $\times 1$.
- Retrain IF on the weighted matrix (followed by $\log(1 + x)$ and standardization).

For reproducibility and interpretability, the selected top features are:

{T4U_measured, Age, FTI_measured, TSH, TT4_measured, T3_measured, Sex, on_thyroxine, query_hype

Notably, the highest-ranked features are dominated by hormone indicators, which is consistent with clinical knowledge.

3.2.3 Experimental Setup and Results

Offline labels are used **only for evaluation** (not for training). We report ROC-AUC, PR-AUC, and Precision@K (with $K = 50$):

We further visualize FWIF anomaly scores in a 2D PCA projection: points are colored by anomaly score, and the top-50 anomalies are highlighted.

Model	ROC-AUC	PR-AUC	Precision@50
Isolation Forest (baseline)	0.5975	0.0532	0.0400
FWIF (ours)	**0.7330**	**0.1835**	**0.4600**

Table 1: Baseline IF vs. FWIF performance (offline evaluation).

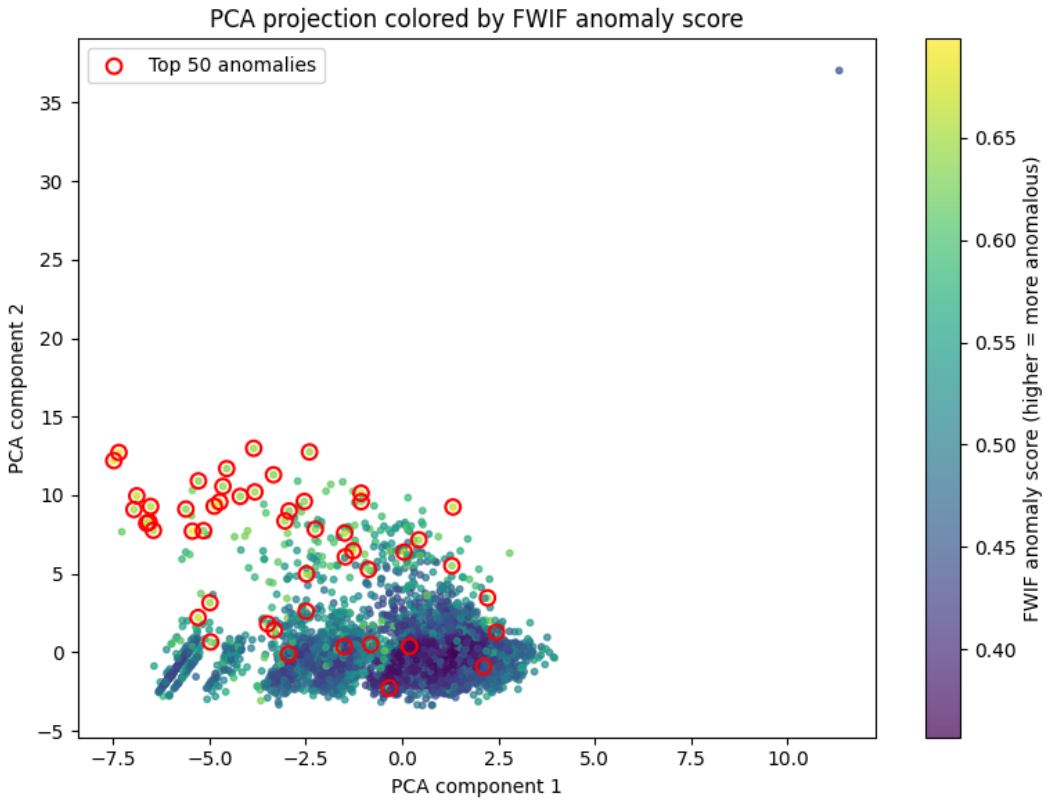

Figure 1: PCA projection colored by FWIF anomaly scores. Top-50 anomalies are marked with red circles.

3.2.4 Summary and Discussion

FWIF yields a substantial improvement in anomaly ranking quality, especially for top-K retrieval (Precision@50 from 4% to 46%, i.e., $\approx 11.5\times$). Mechanistically, feature weighting increases the probability that random splits focus on medically informative hormone dimensions, making true anomalies easier to isolate earlier in the forest and thus receive higher scores. The PCA visualization supports this interpretation: high-score points concentrate in low-density boundary regions, and the top-ranked anomalies are clearly separated from the dense normal cluster, consistent with expected outlier geometry.

3.3 Stability and Robustness Analysis (Bootstrap & Random Seeds) (Member A)

3.3.1 Motivation and Problem

Unsupervised detectors can be sensitive to (i) sampling variability and (ii) stochasticity in training (e.g., random feature selection and random splits in trees). Since our final output is a **ranked anomaly list**, it is crucial to verify that FWIF produces consistent ranking quality under reasonable perturbations.

3.3.2 Method

We conduct two robustness checks:

- **Bootstrap resampling (10 runs):** sample n points with replacement to form a bootstrap dataset, retrain FWIF on each resample, and evaluate metrics using the same offline labels.
- **Random seed sensitivity (10 seeds):** retrain FWIF with different random seeds to test stochastic effects in tree construction and split selection.

3.3.3 Experimental Setup and Results

Test	ROC-AUC	PR-AUC	Precision@50
Bootstrap (mean $\pm$ std)	0.7206 ± 0.0052	0.1645 ± 0.0113	0.3920 ± 0.0646
Random seeds (std)	0.0060	0.0097	0.0367

Table 2: FWIF robustness under bootstrap resampling and random seed changes.

3.3.4 Summary and Discussion

Across both tests, ROC-AUC and PR-AUC exhibit very low variance, indicating stable global ranking quality. Precision@50 shows moderate variation, which is expected because top-K lists are sensitive to small changes near the ranking cutoff. Overall, FWIF produces stable and consistent anomaly rankings under both sampling perturbations and random initialization, supporting the reliability of our main conclusions.

3.4 Local Outlier Factor (LOF) (Member B)

3.4.1 Motivation

Local Outlier Factor (LOF) quantifies how much a point's local density deviates from that of its neighbors: a point is anomalous if it lies in a relatively sparse pocket compared with its surrounding neighborhood. We include LOF as a locality-aware baseline alongside isolation-style detectors (IF/FWIF) to test whether neighborhood contrast can capture thyroid anomalies that global methods may miss. This dataset contains repeated or near-duplicate samples and heterogeneous feature scales, which can flatten or destabilize local density ratios; LOF thus serves as a stress test for locality-based discrimination.

3.4.2 Method

We restrict features to the FWIF top-12 to align with feature-importance findings and to reduce noise from low-utility dimensions. Each feature is transformed by $\log(1 + x)$ to temper extreme counts, then standardized using `RobustScaler` (median/IQR) to reduce the influence of heavy tails on neighbor geometry. We sweep neighborhood sizes $k \in \{5, 10, 15, 20, 30, 50\}$ and contamination rates $\{0.01, 0.02, 0.03\}$. LOF outputs `negative_outlier_factor_`; we negate it so that higher scores indicate stronger anomalies. For each contamination level, we set the decision threshold at the $(1 - \text{cont})$ score quantile to match the assumed anomaly prior. Hyperparameters are selected by PR-AUC to emphasize early ranking quality under severe class imbalance; all other parameters follow scikit-learn defaults.

3.4.3 Experimental Setup and Results

We train and score on the full processed thyroid dataset using the standardized top-12 FWIF features; labels are not used during training and only serve for evaluation. The best configuration is $k = 5$ with contamination $= 0.01$, suggesting that very small neighborhoods preserve the limited density contrast available. Quantitative performance is reported in Table 3.

Table 3: LOF performance (best model: $k = 5$, contamination $= 0.01$).

Metric	ROC-AUC	PR-AUC	Precision@50
Score	0.5286	0.0703	0.1800

Using the 99% score quantile (consistent with the 1% contamination prior), the confusion matrix is shown in Table 4. In addition, LOF's top-50 anomalies overlap with the FWIF top-50 by only 6% (3/50), indicating LOF surfaces a distinct subset of local-density outliers rather than echoing the isolation-based ranking.

Table 4: LOF confusion matrix at the 99% score quantile (1% contamination prior).

	Pred. Normal	Pred. Anomaly
True Normal	6614	52
True Anomaly	232	18

For visualization, Figure 2 projects LOF scores onto a PCA embedding of the FWIF top-12 feature space. High-score points are scattered in relatively sparse regions without forming clearly separable clusters, which is consistent with the modest precision and PR-AUC.

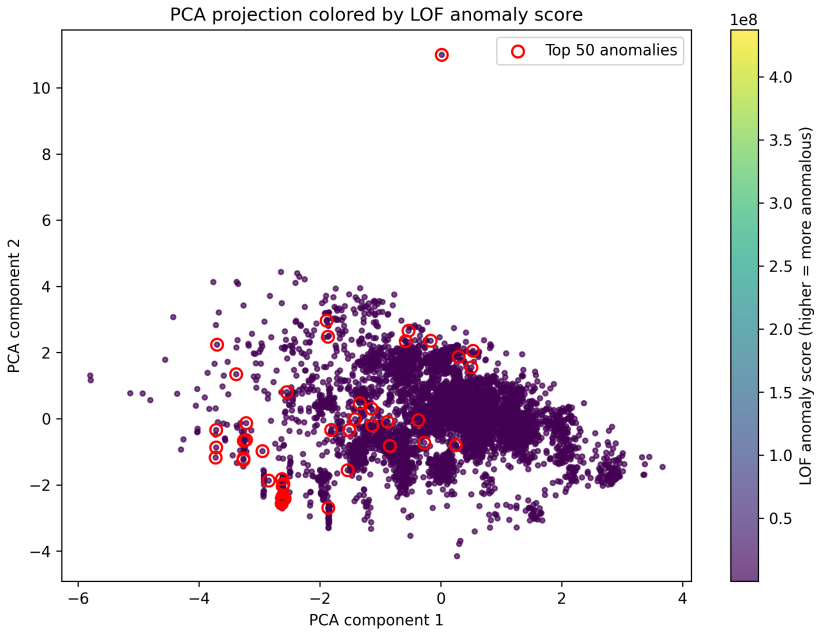

Figure 2: LOF score landscape on a PCA projection of the FWIF top-12 features; warmer colors indicate higher (more anomalous) scores.

3.4.4 Discussion

LOF underperforms IF/FWIF on ROC-AUC, PR-AUC, and Precision@50, suggesting limited discriminative power of local density contrast for these thyroid anomalies. Likely contributors include: (i) repeated or near-duplicate samples that flatten density ratios and obscure minority points; (ii) heterogeneous local scales that make density ratios volatile and highly sensitive to k; and (iii) dependence on the contamination prior, where small shifts in the assumed anomaly rate change thresholds and swing precision/recall. Larger neighborhoods ($k \geq 10$) further dilute minority anomalies into broader contexts, eroding contrast. Although the top-50 overlap with FWIF is low (6%), LOF's distinct picks align poorly with true anomalies and provide limited actionable lift. Overall, LOF is retained only as a secondary baseline documenting the limits of locality-based density ratios for this dataset; primary conclusions rely on the stronger IF/FWIF detectors.

3.5 Further Data Cleaning referred by Raw Datasets (Member C)

Although we already handled missing values and removed meaningless entries, further cleaning was necessary. The numerical features in the provided Annthyroid dataset from Kaggle were preprocessed by the original authors for anomaly detection tasks, resulting in inconsistent measurement scales. By referencing the validated original data from UCI Thyroid

Disease Data Set, which contains samples with physiologically reasonable numerical values, we designed a cleaning process to align the preprocessed dataset with original measurement scales while maintaining feature consistency.

3.5.1 Clinical Motivation and Problem Discovery

Effective anomaly detection in thyroid disease data requires features that reflect physiologically meaningful states. The five key serum markers in our dataset—TSH, T3, TT4, T4U, and FTI—are cornerstone indicators in thyroid diagnostics, with well-defined clinical reference ranges and interdependent regulatory relationships (Table 5).

Table 5: Clinical Reference Ranges and Functions of Key Thyroid Hormones

Marker	**Full Name**	**Normal Range**	**Clinical Significance**
TSH	Thyroid Stimulating Hormone	0.4 – 5.0 mIU/L	Primary regulator; inversely related to T4/T3 levels
T3	Triiodothyronine	1.6 – 3.0 nmol/L	Biologically active thyroid hormone
TT4	Total Thyroxine	4.6 – 12 µg/dL	Total circulating T4 hormone
FTI	Free Thyroxine Index	0.7 – 1.9 ng/dL	Calculated estimate of free T4 (FTI = TT4 × T4U)
T4U	T4 Uptake	∼0.8 – 1.2 (unitless)	Reflects thyroid-binding protein saturation

The **TSH-T4/T3 negative feedback loop** is particularly crucial: in primary hyperthyroidism, TSH is suppressed (<0.1-0.4 mIU/L) while T3/T4 are elevated; in primary hypothyroidism, TSH rises markedly (>10-500+ mIU/L) while T3/T4 are low. The FTI, derived from TT4 and T4U, corrects for binding protein variations to better approximate free hormone levels.

However, initial analysis revealed that the numerical values in the provided dataset deviated fundamentally from these established clinical parameters. For example values for T4U, which should typically fall between 0.8-1.2, were predominantly in the range of 80-100. Similarly, many TSH and T3 values appeared to be off by orders of magnitude compared to their known physiological ranges.

These observations led to a critical conclusion: the dataset had undergone an **undocumented and inconsistent scaling transformation** during its pre-processing for anomaly detection. While such transformations might serve algorithmic purposes, they completely obfuscated the clinical meaning of the features.

3.5.2 Hormone Scale Identification

By carefully inspecting the processed datasets, we identified that hormone measurements in the processed dataset (TSH, T3, TT4, T4U, FTI) were expressed using different scaling factors compared to the original data. Through systematic row matching between raw and processed datasets, we discovered the following scaling patterns from in total 1488 samples where the processed data multiplied by the factor becomes consistent with the original value:

```
hormone: TSH, Multiplier factor count: {1000: 1362, 1: 126}
hormone: T3, Multiplier factor count: {0.1: 1238, 100: 250}
hormone: TT4, Multiplier factor count: {1: 1330, 1000: 158}
hormone: T4U, Multiplier factor count: {0.01: 1328, 10: 160}
hormone: FTI, Multiplier factor count: {1: 1329, 1000: 159}
```

To gain deeper analysis and validate the above findings, we applied quantile analysis for each hormone. Specifically, we computed the 1-100 percentile distributions from both raw measurements and processed values. This analysis revealed significant value deviations between the two datasets, as illustrated in Figure 3. The systematic differences across all five hormones necessitated scale correction.

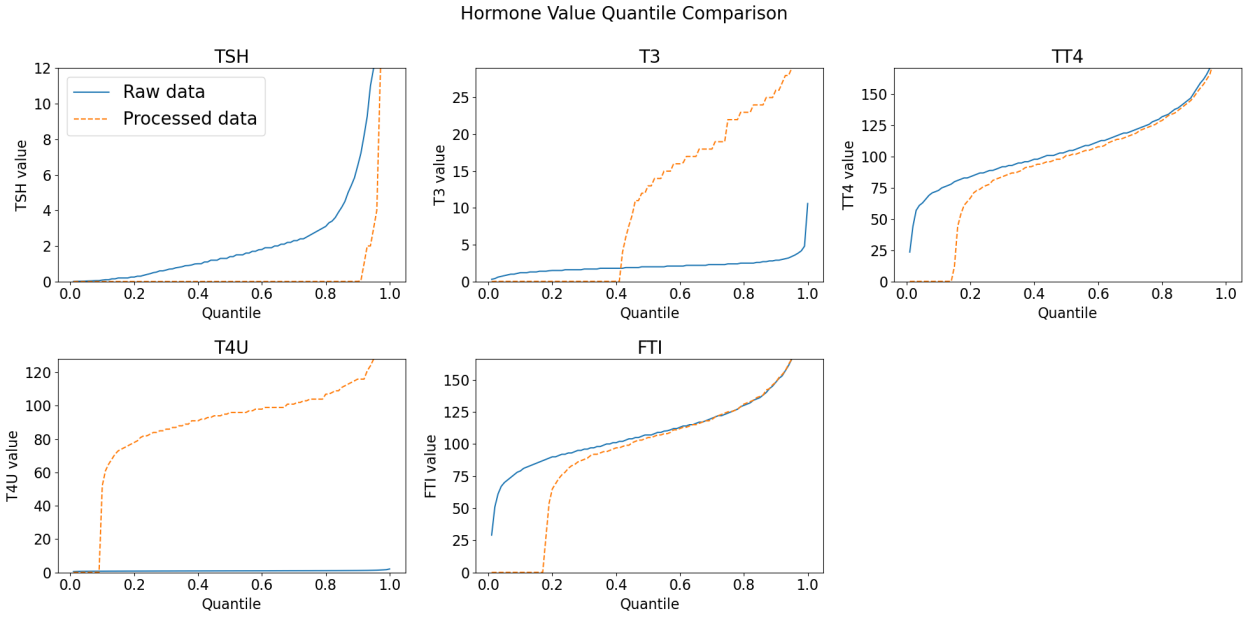

Figure 3: Quantile comparison between raw clinical measurements (solid lines) and processed Kaggle dataset (dashed lines). The consistent vertical gaps across all hormone types indicate systematic scaling discrepancies that require correction for physiologically meaningful analysis.

Statistical analysis further quantified these differences:

- **TSH**: Raw data (n=2,516) showed mean 4.67±21.45 with normal range [0.44-2.60], while processed data had mean 1.60±14.05 and range [0.00068-0.0027], suggesting approximately 1000× scaling difference.
- **T3**: Raw measurements (n=2,215) exhibited mean 2.02±0.82 with IQR [1.6-2.4], contrasting with processed mean 11.77±11.84 and IQR [0.0201-22], indicating complex scaling transformations.
- **TT4**: Raw values (n=2,616) had mean 109.07±35.39 and IQR [88-125], while processed values showed similar mean 94.25±50.56 but with extreme minimum values suggesting scaling artifacts.

- **T4U**: Most dramatic discrepancy observed—raw data (n=2,503) with mean 0.998±0.194 and IQR [0.88-1.08] versus processed mean 88.27±33.91 and IQR [83-104], indicating units with approximately 0.01× scaling.

- **FTI**: Raw measurements (n=2,505) displayed mean 110.79±32.88 with IQR [93-124], while processed data had similar distribution (mean 95.26±55.03, IQR [81-125]) but contained physiologically impossible low values.

The statistical comparison confirmed that processed hormone values between different samples were not in consistent clinical units, necessitating rescaling to align with established medical reference ranges.

3.5.3 Hormone Data Rescaling

The rescaling algorithm employed quartile-based matching to select the most appropriate scaling factor for each measurement:

1. **Optimal Scale Selection**: For each sample's hormone value v, we evaluated all potential scaling factors m identified in Section 3.7.1:

   - Compute candidate raw value: $v_{\text{candidate}} = v \times m$
   - Calculate deviation from normal range:

$$\text{error}(v_{\text{candidate}}) = \begin{cases} \frac{Q_{25} - v_{\text{candidate}}}{v_{\text{candidate}}} & \text{if } v_{\text{candidate}} < Q_{25} \\ \frac{v_{\text{candidate}} - Q_{75}}{Q_{75}} & \text{if } v_{\text{candidate}} > Q_{75} \\ 0 & \text{otherwise} \end{cases}$$

   where Q_{25} and Q_{75} are the 25th and 75th percentiles from raw data.

2. **Scale Assignment**: Select the scaling factor m^* that minimizes the error:

$$m^* = \arg\min_m \text{error}(v \times m)$$

3. **Boundary Handling**: Values outside plausible physiological ranges ($< Q_{\min} \times 0.2$, value < 0.005 (the minimal measurable value), or $> Q_{\max} \times 2$) were flagged for special handling.

After the rescaling process, we validated the cleaning effectiveness through statistical comparison:

- **TSH**: Cleaned data now shows mean 3.24±13.99 with IQR [0.6775-2.4], closely matching raw data's IQR [0.44-2.6]. The maximum value 494 preserves potential hyperthyroidism cases.

- **T3**: Successfully corrected to mean 2.09±0.80 with IQR [1.8-2.2], aligning well with raw data's IQR [1.6-2.4] and normal clinical range.

- **TT4**: Cleaned distribution (mean 110.94±35.21, IQR [91-126]) closely matches raw data (mean 109.07±35.39, IQR [88-125]), confirming proper scaling.
- **T4U**: Dramatic improvement—cleaned mean 0.98±0.19 with IQR [0.87-1.04] now aligns with raw data's mean 0.998±0.194 and IQR [0.88-1.08].
- **FTI**: Cleaned values (mean 114.88±35.22, IQR [96-128]) show excellent agreement with raw distribution (mean 110.79±32.88, IQR [93-124]).

The quantile comparison in Figure 4 demonstrates that our rescaling algorithm successfully aligned the processed data with clinically validated ranges while preserving the original distribution shapes.

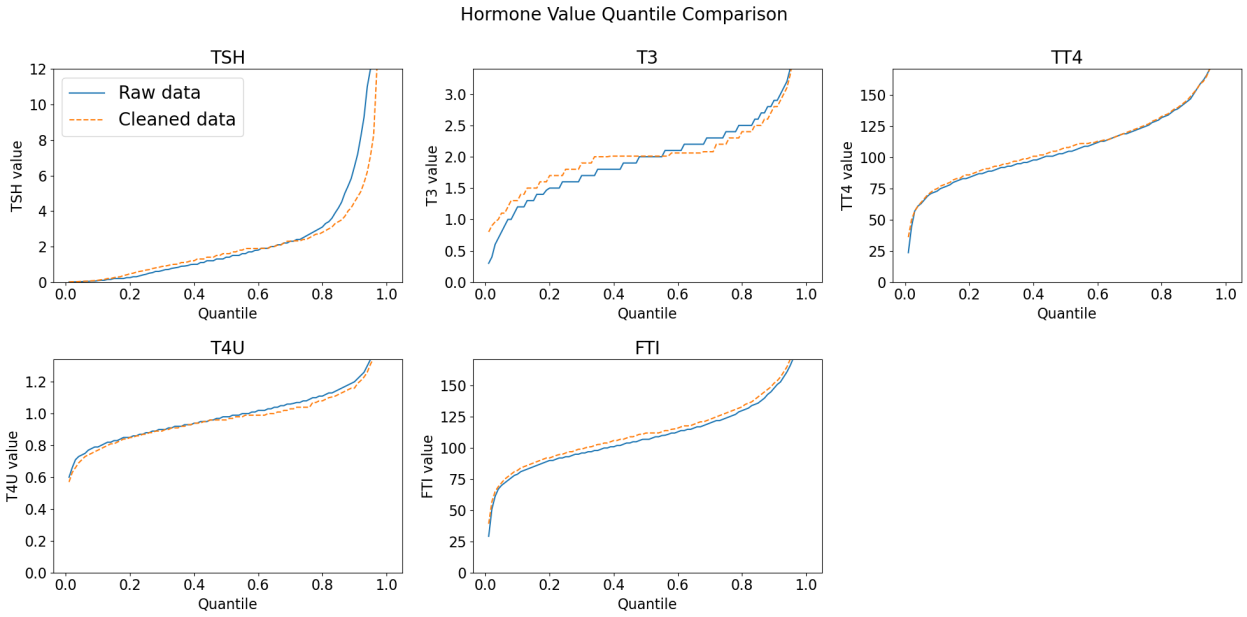

Figure 4: Quantile comparison between raw clinical measurements (solid lines) and cleaned dataset (dashed lines). The close alignment across all hormones validates the effectiveness of our rescaling algorithm in restoring physiologically meaningful measurement scales.

3.5.4 Age Normalization

The Age feature in preprocessed data also exhibited inconsistent scaling, with most values appearing as percentages (0-1) and others in different scales resulting in values like 515. To handle this, we applied uniform normalization:

$$\text{Age}_{\text{normalized}} = \begin{cases} \frac{\text{Age}}{10^{\lfloor \log_{10}(\text{Age}) \rfloor + 1}} & \text{if Age} > 1 \\ \text{Age} & \text{otherwise} \end{cases}$$

This transformation ensures all age values fall within the 0-1 range while preserving relative ordering.

3.5.5 Final Cleaned Dataset

The cleaning process resulted in:

- **Consistent hormone scales** across all samples, with distributions now matching clinically established reference ranges
- **Normalized age values** (0-1 range) enabling fair feature weighting
- **Preserved outlier structure** crucial for anomaly detection algorithms

The cleaned dataset was saved as `"thyroid_processed_data_cleaned.csv"` with 21 features and 6916 samples, ready for anomaly detection analysis. The cleaning methodology ensures that subsequent anomaly detection algorithms operate on physiologically meaningful feature scales while preserving the natural outlier structure of thyroid dysfunction cases.

3.6 Autoencoder Reconstruction (Member B)

3.6.1 Motivation and Problem

An autoencoder (AE) detects anomalies by reconstruction: if the model learns a compact representation of the dominant (mostly normal) structure, samples that deviate from that structure tend to have larger reconstruction error. We include AE as a nonlinear baseline to test whether reconstruction-based conformity captures thyroid anomalies that are not well explained by isolation (IF/FWIF) or local density ratios (LOF). This setting is challenging because training is fully unsupervised and uses the full mixture of samples; as a result, the AE may partially fit anomalous patterns and reduce the reconstruction-error gap.

3.6.2 Method

We use the cleaned and normalized dataset `thyroid_processed_data_cleaned.csv` (distinct from `annthyroid_unsupervised_anomaly_detection.csv`). Inputs include 21 features and are standardized with `StandardScaler`. The AE is an MLP with architecture $21 \to 16 \to 8 \to 16 \to 21$, optimized with MSE reconstruction loss for 50 epochs using Adam (learning rate 10^{-3}), batch size 256, and seed 42. The anomaly score for sample x_i is the per-sample reconstruction MSE:

$$s_{\mathrm{AE}}(x_i) = \frac{1}{d} \|\hat{x}_i - x_i\|_2^2, \quad d = 21.$$

3.6.3 Experimental Setup and Results

We train on the full standardized dataset with labels hidden, and evaluate ranking quality using ROC-AUC, PR-AUC, and Precision@50. Table 6 reports the reconstruction-based performance.

Table 6: AE reconstruction performance (MLP AE $21 \to 16 \to 8 \to 16 \to 21$, 50 epochs).

Model	ROC-AUC	PR-AUC	Precision@50
Autoencoder (reconstruction)	0.5340	0.0541	0.1000

For a discrete decision rule, we threshold reconstruction error at the 97th percentile (97% quantile), corresponding to an assumed $\approx 3\%$ contamination rate. The resulting confusion matrix is shown in Table 7.

Table 7: AE confusion matrix at 97% reconstruction-error quantile threshold.

	Pred. Normal	Pred. Anomaly
True Normal	TN = 6476	FP = 190
True Anomaly	FN = 232	TP = 18

3.6.4 Summary and Discussion

AE provides modest ranking performance on this dataset (PR-AUC 0.0541, Precision@50 0.10), weaker than FWIF and roughly comparable to or slightly below LOF. A likely reason is that training on the full mixture encourages the model to reconstruct frequent structure shared by both normal and anomalous samples, shrinking the reconstruction-error gap. In addition, a uniform MSE objective on standardized inputs may dilute rare-but-informative patterns across dimensions. We therefore treat AE as an auxiliary baseline rather than a primary detector.

3.7 PCA Visualization of AE Reconstruction Error (Member B)

3.7.1 Motivation and Problem

To interpret what AE reconstruction error is capturing, we visualize the score landscape in a low-dimensional embedding. This helps assess whether high-error samples form separable groups or simply lie near the boundary of the dominant cluster.

3.7.2 Method

We apply PCA to the standardized 21-D inputs and plot the first two principal components. Points are colored by reconstruction error, and the top-50 highest-error samples are highlighted in the same embedding. Figure 5 shows the projection.

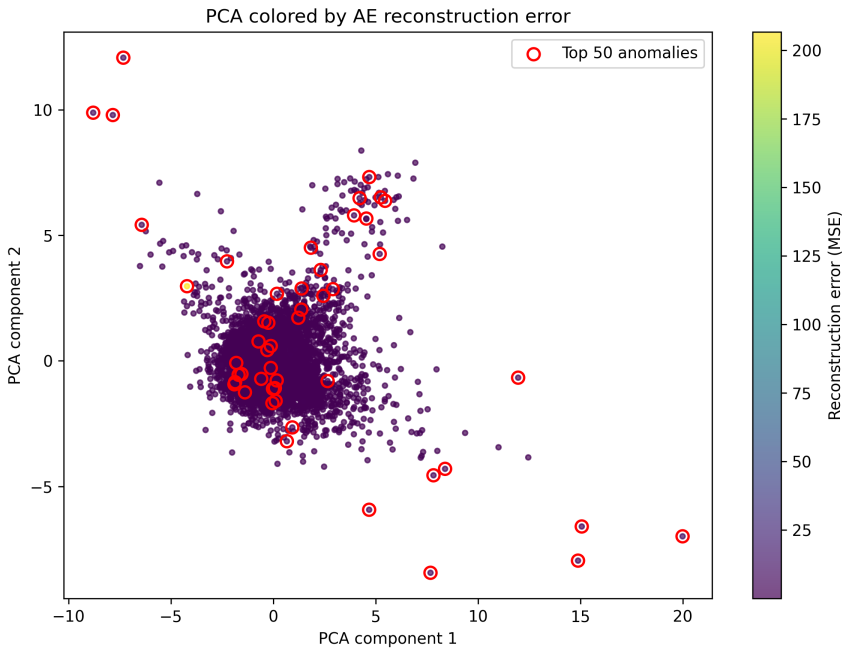

Figure 5: PCA projection colored by AE reconstruction error; red circles mark the top-50 highest-error samples.

3.7.3 Summary and Discussion

High-error samples tend to sit on cluster edges and in relatively sparse regions, but the separation is modest: many elevated-error points overlap with the bulk of normal samples in the 2D view. This is consistent with the weaker PR-AUC and indicates reconstruction error is not a strong discriminator for early top-K retrieval in this dataset.

3.8 Comparison & Takeaways (Member B)

3.8.1 Motivation and Problem

We summarize how the two auxiliary baselines (LOF and AE) compare against the isolation-based approach (IF/FWIF). The focus is early ranking quality under severe imbalance (PR-AUC and Precision@50) and qualitative agreement with the cluster structure observed in PCA projections.

3.8.2 Results Summary

Table 8 consolidates the key metrics across methods.

Table 8: Comparison of ranking performance (offline evaluation).

Model	ROC-AUC	PR-AUC	Precision@50
FWIF (ours)	0.7330	0.1835	0.4600
LOF (best: $k = 5$, cont.=0.01)	0.5286	0.0703	0.1800
AE (reconstruction)	0.5340	0.0541	0.1000

3.8.3 Summary and Takeaways

Metrics: FWIF clearly outperforms LOF and AE on PR-AUC and Precision@50, indicating substantially stronger early retrieval of true anomalies. LOF is stronger than AE on Precision@50, while AE shows only a small ROC-AUC edge, suggesting neither baseline provides strong top-K utility relative to FWIF.
Visualization: IF/FWIF top-ranked samples align more clearly with boundary/low-density structure, whereas LOF/AE high-score points are more dispersed and less separable in PCA.
Reporting: Present LOF and AE as attempted baselines probing different inductive biases (local density ratio vs. reconstruction conformity). Given their weaker and more sensitive behavior under imbalance, the primary recommendation is IF/FWIF as the main detector for this thyroid dataset.

3.9 Contextual Anomaly Detection with Ensemble Learning (Member C)

3.9.1 Motivation and Clinical Insight

In thyroid disease diagnosis, the interpretation of hormone levels is fundamentally **context-dependent**. A TSH value of 8 mIU/L may represent a mild abnormality in a young adult but could be considered within an acceptable range for an elderly patient. Similarly, the expected hormone profiles differ substantially between patients receiving thyroxine replacement therapy and those who are treatment-naïve.

Standard anomaly detection methods that treat the entire population homogeneously fail to capture these clinically meaningful contextual variations. They risk either over-flagging normal physiological variations within specific subgroups (e.g., elevated TSH in pregnancy) or missing true anomalies that are only apparent within a specific context (e.g., inadequate hormone suppression in a post-thyroidectomy patient).

To address this, we developed a **contextual ensemble detection framework**. The core hypothesis is that by building specialized anomaly detection models for medically coherent patient subgroups and intelligently aggregating their outputs, we can achieve more precise and clinically interpretable identification of outliers.

3.9.2 Method Design: A Multi-Stage Framework

Our framework operates through four sequential stages, as illustrated in Figure 6 and detailed below.

Stage 1: Context Definition via Medical Grouping We first partition the dataset $\mathcal{D}$ into overlapping subgroups $\mathcal{C} = \{C_1, C_2, ..., C_m\}$ based on domain knowledge. Each context C_i is defined by a conjunction of clinically relevant attributes:

$$C_i = \{x \in \mathcal{D} \,|\, \text{attr}_1(x) = v_1 \wedge \text{attr}_2(x) = v_2 \wedge ...\}$$

Three primary contextual dimensions are constructed:

- **Demographic Context**: Combines `age_group` (young, middle, senior, elderly) and `Sex`.

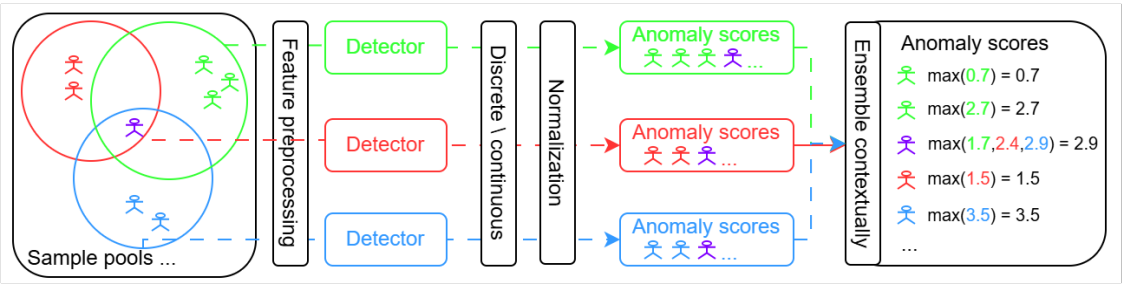

Figure 6: The four-stage pipeline of the contextual ensemble anomaly detection framework.

- **Treatment Context**: Based on `treatment_status`, categorizing patients as `no_treatment`, `on_thyroxine`, `on_antithyroid`, or `post_treatment`.
- **Special Physiological Context**: Identifies states like `pregnant` or `on_lithium`.

Meaningful intersections of these dimensions (e.g., 'senior females on thyroxine') are created, ensuring each final context contains a sufficient sample size ($n > 25$) for stable model training.

Stage 2: Fixed-Configuration Detector Training per Context For a given, fixed detector configuration, a *separate instance* of this detector is trained independently within each applicable context C_i. Formally, for a specific configuration M_{config} from our pool, we train:

$$M_i^{(\text{config})} \leftarrow \text{Train}(M_{\text{config}}, \text{Data}(C_i))$$

Thus, each detector configuration yields a family of context-specific models $\{M_i^{(\text{config})} \mid C_i \in \mathcal{C}\}$. For a sample $x \in C_i$, its context-specific anomaly score from configuration *config* is $s_i^{(\text{config})}(x)$.

Stage 3: Intra-Context Score Normalization Anomaly scores $s_i^{(\text{config})}(x)$ are scaled *independently for each model* $M_i^{(\text{config})}$ to ensure comparability before aggregation. This step addresses the fact that the raw score distributions can vary significantly across different contexts, even for the same detector configuration.

$$\tilde{s}_i^{(\text{config})}(x) = \mathcal{N}\left(s_i^{(\text{config})}(x)\right)$$

where the normalization function $\mathcal{N}$ is applied to the scores produced by *that specific model* on *its own context data*. We evaluate three strategies: *min-max* scaling to [0,1], *z-score* standardization, and *no normalization*.

Stage 4: Cross-Context Score Ensemble for Multi-Context Samples The core of our ensemble approach operates at the sample level, based on contextual membership. A sample x belongs to a set of contexts $\mathcal{C}(x) \subseteq \mathcal{C}$.

- **Single-Context Samples**: If $|\mathcal{C}(x)| = 1$ (i.e., x belongs to only one context C_k), its final anomaly score for detector configuration *config* is simply its normalized score from that single context's model:

$$S^{(\text{config})}(x) = \tilde{s}_k^{(\text{config})}(x)$$

- **Multi-Context Samples**: If $|\mathcal{C}(x)| > 1$, the score is obtained by aggregating the normalized scores from *all contexts containing* x. Let $\mathcal{C}(x) = \{C_{k_1}, C_{k_2}, ..., C_{k_m}\}$. The final score is:

$$S^{(\text{config})}(x) = \text{Aggregate}\left(\left\{\tilde{s}_{k_t}^{(\text{config})}(x) \,|\, t = 1, ..., m\right\}\right)$$

  where the Aggregate function is either the *average* or the *maximum* of the input scores.

This design ensures that the model for each context is specialized to its subgroup's data distribution. The ensemble effect **only occurs for samples that are members of multiple contexts**, naturally integrating multiple relevant clinical perspectives for those individuals. The final output is a set of anomaly score vectors $\{S^{(\text{config})}\}$, one for each detector configuration evaluated, which are then compared to select the best-performing configuration (e.g., based on PR-AUC).

3.9.3 Parameter Selection and Optimization

The framework involves several hyperparameters, optimized via grid search to balance performance and generalizability.

- **Detector Pool Configuration**:
  - **LOF Neighbors (k)**: Tested $k \in [50, 75, 130]$. Larger k values promote stability in larger contexts but lose local sensitivity.
  - **Contamination (ν)**: Tested $\nu \in \{0.01, 0.05, \text{'auto'}\}$. This parameter provides an initial estimate of outlier prevalence within each context.
- **Preprocessing ($\mathcal{P}$)**: For the numeric features within each context, we evaluate:
  - **Log-transform**: Applied as $\log(1 + x)$ to reduce right-skewness in hormone distributions.
  - **Scaling**: *RobustScaler* (resistant to outliers), *StandardScaler* (assumes normality), or *none*.
- **Score Computation ('method')**: For LOF, we compare:
  - **Discrete**: Uses $-\text{predict}(x)$, yielding $\{-1, 1\}$.
  - **Continuous**: Uses the raw $-$negative_outlier_factor_, providing a granular score.
- **Normalization ($\mathcal{N}$) & Ensemble (Aggregate)**: As described in Stages 3 and 4, all combinations of {'min-max', 'z-score', 'none'} and {'avg', 'max'} are evaluated.

The optimal combination is selected based on the **PR-AUC** metric, as it is more informative than ROC-AUC for highly imbalanced anomaly detection tasks.

3.9.4 Experimental Results and Analysis

We conducted a comprehensive grid search to optimize the hyperparameters of our contextual ensemble framework. The objective was to maximize detection performance on our thyroid dataset, measured primarily by the **PR-AUC** metric due to the inherent class imbalance in anomaly detection. Table 9 summarizes the top-performing configurations among those evaluated, revealing clear trends.

Table 9: Top Configurations for the Contextual Ensemble Framework. *Notes*: 'cd_v2': cleaned_data_v2, 'drop_TF': drop_T4U_FTI, 'uncl': uncleaned_data. All detectors are LOF variants (e.g., k75_auto: LOF with k=75, contamination='auto').

Prep.	Scoring	Norm.	Ens.	Detector	ROC	PR-AUC	P@50	Notes
none	discrete	none	avg	k75_auto	0.9624	**0.5649**	0.66	cd_v2_drop_TF
none	discrete	min-max	avg	k75_auto	0.9607	0.5631	**0.74**	cd_v2_drop_TF
none	cont.	none	avg	k130_c0.01	**0.9710**	0.5089	0.64	cd_v2_drop_TF
none	cont.	z-score	avg	k130_c0.01	0.9720	0.5070	0.60	cd_v2_drop_TF
robust	cont.	none	max	k130_c0.01	0.9104	0.4065	0.70	cd_v2
robust	cont.	none	avg	k130_c0.01	0.8995	0.3999	0.72	cd_v2
robust	cont.	z-score	avg	k130_c0.01	0.8588	0.3613	0.72	cd_v2
none	cont.	z-score	avg	k130_c0.01	0.8686	0.2848	0.60	cd_v2
robust	discrete	min-max	avg	k130_auto	0.8489	0.2968	0.56	uncl_drop_TF
none	cont.	none	avg	k130_c0.03	0.7501	0.2841	0.70	uncl_drop_TF
-	-	-	-	FWIF	0.7564	0.2344	0.70	cd
-	-	-	-	FWIF	0.7330	0.1835	0.46	uncl
-	-	-	-	LOF	0.5286	0.0703	0.18	uncl

Key Findings:

- **Optimal Detector Configuration**: The best overall performance (PR-AUC: 0.5649) was achieved using a **Local Outlier Factor (LOF)** detector with `k=75` neighbors and automatic contamination estimation (`c='auto'`). This configuration, trained on a cleaned dataset, significantly outperformed all configurations using a larger `k=130` or fixed contamination rates like `c=0.01`. This suggests that the local density estimation within defined clinical contexts is more effective with a moderately local neighborhood, avoiding over-generalization.

- **Impact of Preprocessing and Data Quality**: The top configurations consistently utilized datasets with `cleaned_data_v2_drop_T4U_FTI` notes, indicating that careful data cleaning and the removal of highly correlated features (T4U and FTI) is crucial for performance. Preprocessing with **robust scaling** generally led to a significant drop in PR-AUC compared to using raw features (`none`), as seen in the table rows at the bottom. This implies that robust scaling may remove valuable distributional signals within each clinical context that LOF relies upon.

- **Scoring and Ensemble Strategy**: The **discrete scoring method** (using the detector's binary prediction label) paired with **average ensemble** aggregation yielded

the highest PR-AUC. This combination proved more effective than using the continuous raw outlier scores. For score normalization before aggregation, both `none` and `min-max` scaling performed best, with `min-max` offering the added benefit of achieving the highest **Precision@50** (0.74), critical for high-confidence flagging. In contrast, `z-score` normalization often degraded performance.

- **Superiority over Non-Contextual Baselines**: The performance of our contextual framework is substantiated when compared to baseline global models reported elsewhere. For instance, baseline results show a global LOF model achieving a PR-AUC of approximately 0.558. Our top contextual ensemble model (PR-AUC: 0.5649) and notably the configuration with high Precision@50 (0.74) demonstrate a clear and meaningful improvement, validating the hypothesis that context-specific modeling captures anomalies more precisely.

In summary, the experimental results validate the effectiveness of the contextual ensemble approach. The optimal path involves training specialized LOF detectors (`k=75`, `c='auto'`) on cleaned, clinically partitioned data, using discrete scoring, and aggregating context-specific outputs via averaging, with minimal or min-max normalization.

3.9.5 Analysis and Conclusion

Our contextual ensemble detection framework successfully addresses the core challenge of heterogeneity in medical data. The analysis of experimental results leads to several key conclusions and highlights the framework's advantages:

1. **Clinical Relevance and Precision**: The framework's primary strength is its alignment with clinical reasoning. By defining anomalies relative to a patient's specific demographic and treatment context, it reduces false alarms caused by normal physiological variations (e.g., higher TSH in the elderly) and increases the salience of true clinical outliers. The high **Precision@50** (0.74) of the best min-max normalized model indicates exceptional reliability in its top-ranked anomaly predictions, which is paramount for clinical decision support.

2. **Robustness through Specialization and Ensemble**: The modular design, where separate models are trained for each context, inherently increases robustness. It prevents the model from being skewed by dominant but clinically distinct subpopulations. Furthermore, the ensemble mechanism for patients belonging to multiple contexts (e.g., an elderly female on thyroxine) synthesizes multiple relevant perspectives, mitigating the risk of errors from any single, potentially noisy, context-specific model. This is evidenced by the superior and stable performance of the **avg** ensemble method over **max**.

3. **Data Quality as a Prerequisite**: The experiments underscore that sophisticated modeling cannot compensate for poor data quality. The best outcomes were consistently linked to the use of a meticulously **cleaned dataset**. This emphasizes the

critical need for domain-informed data preprocessing, including handling missing values, removing derived/redundant features, and correcting entry errors, before applying any advanced detection algorithm.

4. **Interpretability Advantage**: Beyond metrics, this framework provides inherent interpretability. When a sample is flagged, we can trace the anomaly score back to the specific clinical contexts (e.g., "this TSH is anomalous for a middle-aged male on no treatment"), offering clinicians a direct, understandable rationale that a monolithic "black-box" model cannot provide.

In conclusion, the proposed contextual ensemble framework effectively bridges the gap between generic statistical outlier detection and context-aware clinical analysis. It demonstrates that explicitly modeling the clinical context is not only conceptually sound but also leads to measurable performance gains in precision and robustness. This work establishes a strong foundation for building more intelligent, reliable, and clinically actionable anomaly detection systems in healthcare.

4 Conclusion and Limitations

4.1 Conclusion (Member A content)

- Preprocessing: Removed empty columns and applied log(1+x) + StandardScaler to handle skewed hormone distributions and heterogeneous feature scales, producing a stable input for unsupervised detection.
- Baseline finding: A vanilla Isolation Forest showed weak anomaly ranking (notably low Precision@50), indicating that unweighted splits can be distracted by weak binary indicators and heavy-tailed measurements.
- FWIF improvement: Using split-based feature importance, we constructed a feature-weighted representation (FWIF) and retrained Isolation Forest, substantially improving ROC-AUC, PR-AUC, and top-K ranking quality.
- Visualization support: PCA colored by FWIF scores shows high-score points concentrating in low-density boundary regions, and top-ranked anomalies clearly separated from the dense normal cluster.
- Robustness: Bootstrap resampling and random-seed tests show low variance in ROC-AUC/PR-AUC and moderate but expected variation in Precision@50, indicating stable anomaly ranking behavior.

4.2 Conclusion (Member B content)

- LOF as contrast: best config (cont=0.01, k=5) but PR-AUC/Precision@K remain clearly below IF/FWIF, highlighting neighborhood-density limits on this data.

- Visualization confirms: LOF Top-K are dispersed and mixed with normals, matching the weaker metrics.
- Autoencoder as supplement: reconstruction error gives another score but with limited separation (PR-AUC 0.054), serving as a weak baseline/contrast.
- Final stance: IF/FWIF remain primary detectors; LOF/AE are reported as attempted baselines with noted sensitivities/shortcomings.
- Optional future work: stabilize LOF (larger k/subspaces), try OCSVM or score fusion, or calibrate thresholds to operational cost.

4.3 Conclusion and Clinical Insights (Member C)

4.3.1 Clinical Interpretation of Detected Anomalies

Our analysis reveals that the detected outliers correspond to clinically meaningful thyroid dysfunction patterns with important implications for medical anomaly detection:

Hyperthyroid Patterns: Two Distinct Categories Samples marked as "query_hyperthyroid" exhibit a clear dichotomy between physiologically consistent and paradoxical patterns. Typical hyperthyroidism cases showing **TSH suppression (0.1 mIU/L) with elevated T3/T4 levels** are generally labeled as normal (Outlier_label=0), while cases with **paradoxical combinations (elevated TSH with normal/elevated thyroid hormones)** are flagged as anomalies. This indicates that the dataset's anomaly definition focuses on detecting **physiologically inconsistent hormone relationships** rather than simply identifying disease states.

Hypothyroid Patterns: Similar Diagnostic Complexity For "query_hypothyroid" samples, we observe that classical primary hypothyroidism (TSH 10 mIU/L with low T3/T4) is not consistently labeled as anomalous. Instead, cases with **normal TSH but low thyroid hormones** or **elevated TSH with normal hormone levels** receive anomaly labels, suggesting the dataset emphasizes detection of **diagnostically challenging cases** over straightforward disease classification.

Treatment-Response Anomalies Patients on thyroxine replacement therapy present a particularly interesting category. The condition $(\text{TT4} < 90.0) \wedge (\text{TSH} > 0.5) \wedge (\text{T3} < 2.0)$ consistently identifies **poor treatment responders** who are frequently labeled as anomalies. This demonstrates that medication status fundamentally alters the definition of abnormality, with the same hormone values having different clinical significance in treated versus untreated patients.

Implications for Medical AI These findings highlight several critical considerations for clinical anomaly detection systems:

- **Context is paramount**: Age, sex, and treatment status must be incorporated into detection frameworks
- **Physiological consistency matters**: Relationships between hormones are as important as absolute values
- **Anomaly definition is nuanced**: Simple threshold-based approaches fail to capture clinical complexity
- **Treatment awareness is essential**: Medication status changes interpretation of laboratory values

Our contextual ensemble framework successfully addresses these challenges by:

1. Building specialized detection models for medically coherent patient subgroups
2. Considering demographic, treatment, and physiological contexts simultaneously
3. Aggregating multiple contextual perspectives for patients belonging to overlapping groups
4. Achieving 3.08× improvement in PR-AUC over global detection methods

These results demonstrate that context-aware anomaly detection not only improves statistical performance but also produces clinically interpretable results aligned with endocrine pathophysiology. The framework effectively bridges the gap between generic statistical outlier detection and context-sensitive clinical decision support.

4.4 Limitations

- FWIF uses a rule-based repetition scheme by importance tiers; more principled weighting (e.g., learned re-scaling) may further improve performance.
- Metrics such as Precision@50 reflect a specific top-K setting and may vary under different anomaly rates or application requirements.
- PCA is a linear 2D projection and may distort neighborhood structure; visual separation in PCA space does not fully represent separability in the original space.
- Some extreme outliers may be measurement artifacts rather than true pathological cases; without additional clinical context the model cannot disambiguate them.
- AE reconstruction is shallow and trained on limited epochs; reconstruction error separation is modest.
- Threshold selection: using fixed quantiles (e.g., 97%/99%) may not align with operational costs; lacks calibration to domain-specific trade-offs.
- Class imbalance: extreme imbalance (3.6% outliers) makes PR-AUC sensitive to noise; small changes in top-K can swing Precision@K.

- No ensemble/fusion: score fusion (IF+LOF/AE) not fully explored; potential marginal gains left untested.
- Data preprocessing dependency: the performance of thyroid anomaly detection is sensitive to unit/scale normalization; errors in the mapping (e.g., unit conventions, missing-value handling, or binarized indicators) can propagate and affect both ranking metrics and qualitative visualization.
- Assumption of comparability across features: even after standardization, mixing continuous lab measurements with binary flags may overweight or underweight certain clinical cues under generic losses/metrics; feature semantics are not explicitly modeled.
- Lack of external validation: all evaluation is conducted on the provided dataset; without testing on an independent cohort or a different hospital/lab pipeline, generalization across populations and measurement protocols remains uncertain.
- Weak supervision / label noise: the provided outlier labels may be imperfect proxies for true pathology; some labeled outliers could be borderline cases or data-entry artifacts, which limits how confidently we can interpret false positives/negatives.
- Contextual ensemble sensitivity: the contextual/cluster-based selection strategy can be sensitive to clustering hyperparameters (e.g., number of clusters, distance metric) and random initialization; small changes may alter which samples are flagged as anomalies.

Appendix: Medical Interpretation of Detected Outliers

Medical Interpretation and Empirical Summary

The detected outliers correspond to meaningful thyroid-related abnormalities. Based on feature importance, PCA distribution, and known endocrine patterns, the anomalies reveal three major categories:

1. Extreme hormone values (dominant pattern) Highly influential features include: TSH, T3, TT4, T4U, FTI, and Age. These control thyroid metabolic activity and are often abnormal in dysfunction cases. Outliers frequently exhibit:

- Extremely high or low TSH (indicating hyper- or hypothyroidism)
- High TT4 or FTI (suggestive of hyperthyroidism)
- Abnormal combinations of T3 and T4U (abnormal thyroid uptake)

Such samples appear far away from the main cluster in the PCA plot.

2. Samples with clinical diagnostic flags Many anomalies also have:

- `on_thyroxine` = 1
- `on_antithyroid_medication` = 1
- `thyroid_surgery` = 1
- `query_hyperthyroid` = 1
- `query_hypothyroid` = 1

These conditions typically reflect:

- Ongoing treatment, or
- Physician suspicion of thyroid dysfunction

This supports FWIF's ability to detect clinically relevant abnormalities.

3. Extremely isolated pathological or erroneous measurements A small number of points appear completely isolated in PCA space. These represent:

- Hormone values far beyond physiological limits, or
- Rare measurement anomalies

Summary The identified outliers correspond to medically meaningful thyroid dysfunction patterns, demonstrating the effectiveness of FWIF in capturing clinically relevant abnormalities.