ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

Outlier Detection in Medical Diagnostic Data: A Comparative Study on the Wisconsin Breast Cancer Dataset

Bahadır Aydın
bahadir.aydin@metu.edu.tr

May 29, 2025

Outline

Context and Motivation

- Labeled medical records are scarce and expensive to obtain; unsupervised outlier detection can flag risky cases and assist medical professionals.
- Early flagging of malignant outliers can speed up treatment, improving patient outcomes.
- In high-dimensional data the curse of dimensionality makes distance unreliable, so we need structure-based ways to spot outliers.

Survey Scope and Method Choice

- We study unsupervised and semi-supervised outlier detection for structured, high-dimensional medical data.
- We compare three established methods—Isolation Forest, Local Outlier Factor, and One-Class SVM—that stand for diverse approaches.
- Isolation Forest splits data at random, LOF calculates local density, and One-Class SVM draws a boundary in kernel space.
- Each method lives in standard libraries and appears in real studies, so the set gives a clear, reproducible benchmark.

Dataset Selection

- Dataset: Breast Cancer Wisconsin (Diagnostic) – 569 samples, 30 numeric features drawn from fine-needle aspirate images.
- Each record carries a label: benign $=1$ or malignant $=0$.
- Data are complete and clean; we apply z-score standardisation and a stratified train–test split using **scikit-learn**.
- Training uses benign cases to reflect semi-supervised practice; evaluation covers both classes.

Isolation Forest

- Ensemble-based partitioning — exploits random partitioning of the feature space.
- Builds many random trees and measures how fast each point gets isolated.
- Anomalies need fewer splits, so they sit near the tree roots.
- Runs in $\mathcal{O}(n \log n)$ time and keeps memory use low.
- Few knobs: number of trees and sub-sample size.

Local Outlier Factor (LOF)

- Density-based locality — focuses on deviations in neighbourhood density.
- Compares the local density of each point to that of its neighbours.
- A point with much lower density than its k neighbours scores as an outlier.
- Handles clusters of different shapes because it stays local.
- Needs k and a distance metric; sensitive to both.

Example

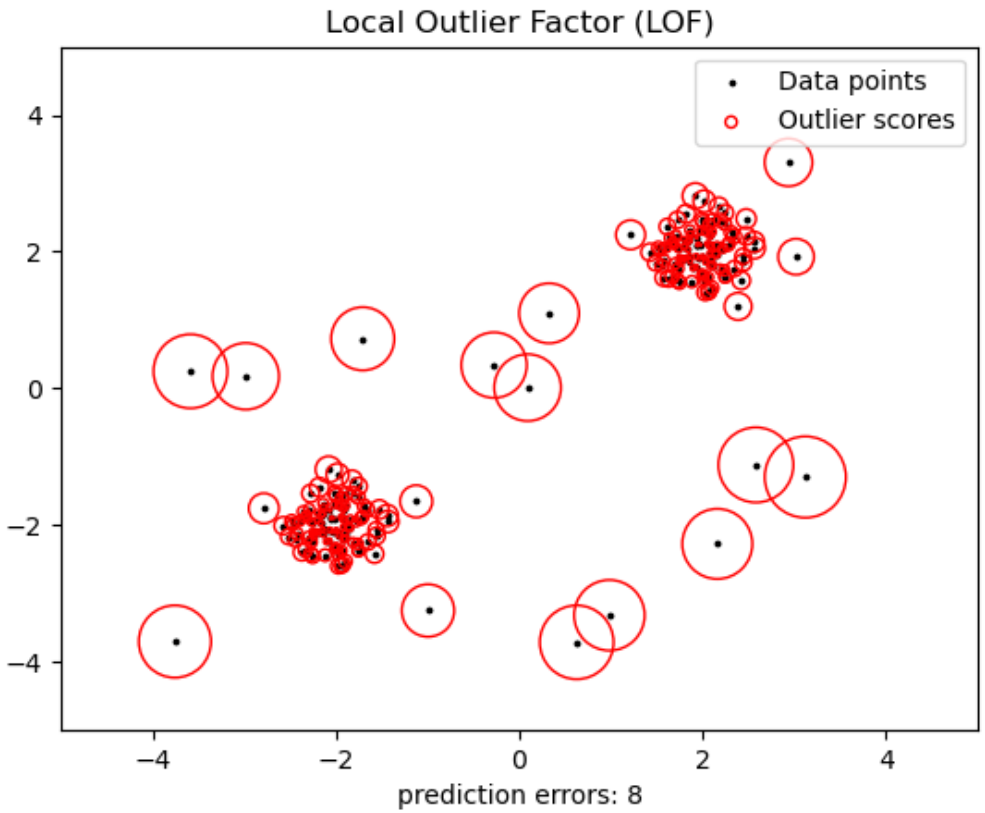


Figure: Taken from scikit-learn documentation.

One-Class SVM

- Kernel boundary learning — treats outlier detection as a one-class classification task.
- Learns a boundary that encloses normal data in high-dimensional kernel space.
- Points outside the boundary count as anomalies.
- Works with any kernel; we use the Radial Basis Function.
- Training is $\mathcal{O}(n^2)$ in practice, so scaling is harder.

Implementation Details (I)

- **Environment**: Python 3.11 with `scikit-learn` 1.4 Pedregosa et al. (2011).
- **Dataset load**: `load_breast_cancer()` from `sklearn.datasets` (Dua and Graff, 2019).
- **Pre-processing**
  - Standardise features using `StandardScaler`.
  - Stratified 80 / 20 train–test split.
- **Workflow**: separate scripts for data, models, and plots; each step can run end-to-end from a single shell command.
- **Reproducibility**: fixed `random_state = 61`; repository on GitHub.

Implementation Details (II)

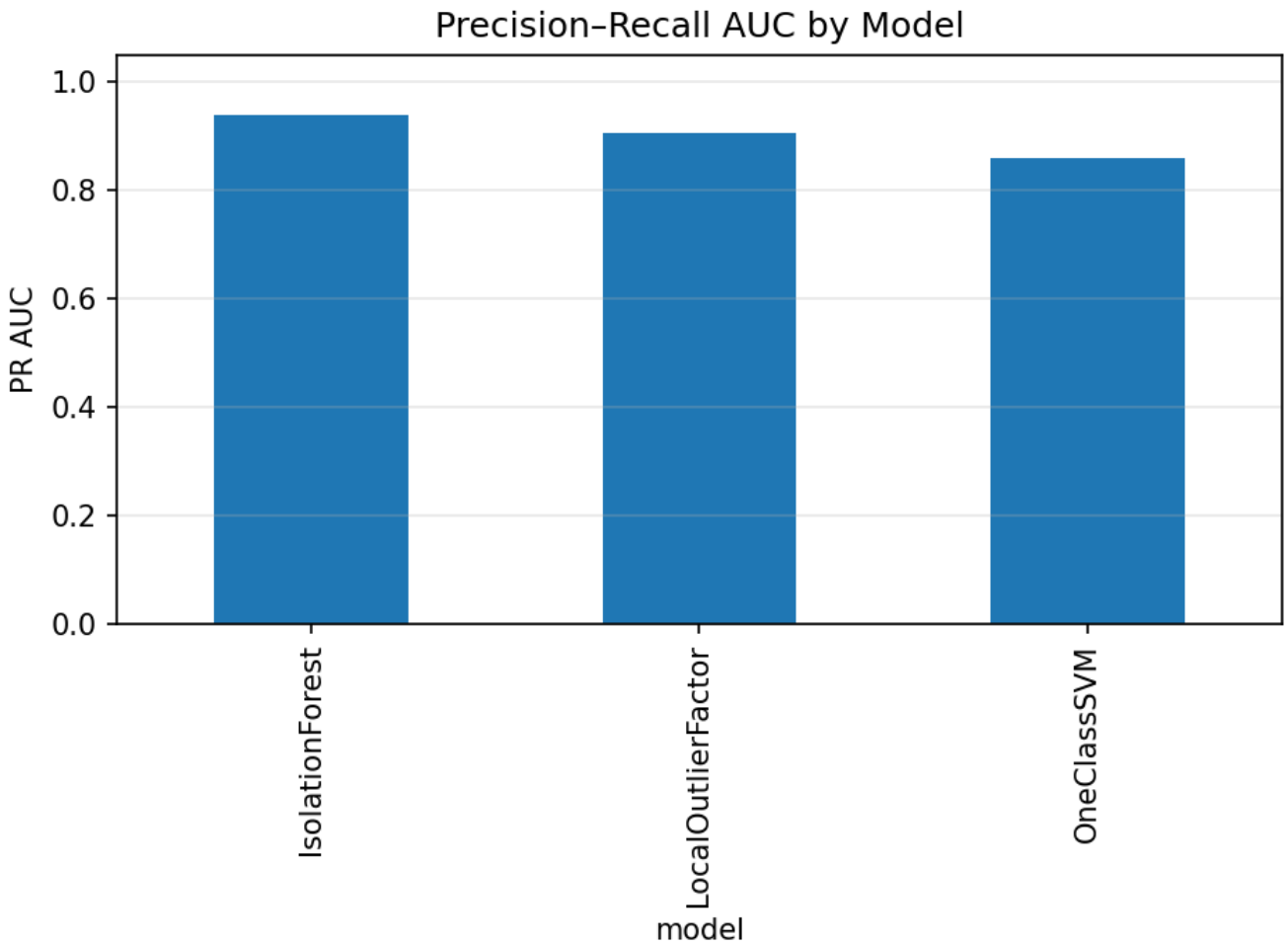
Training protocol

- Fit each model on benign training data.
- Predict anomaly scores on the held-out set (benign + malignant).

Evaluation metrics

- ROC AUC — separates classes across thresholds .
- PR AUC — focuses on the minority class, more informative under imbalance (Saito and Rehmsmeier (2015)).

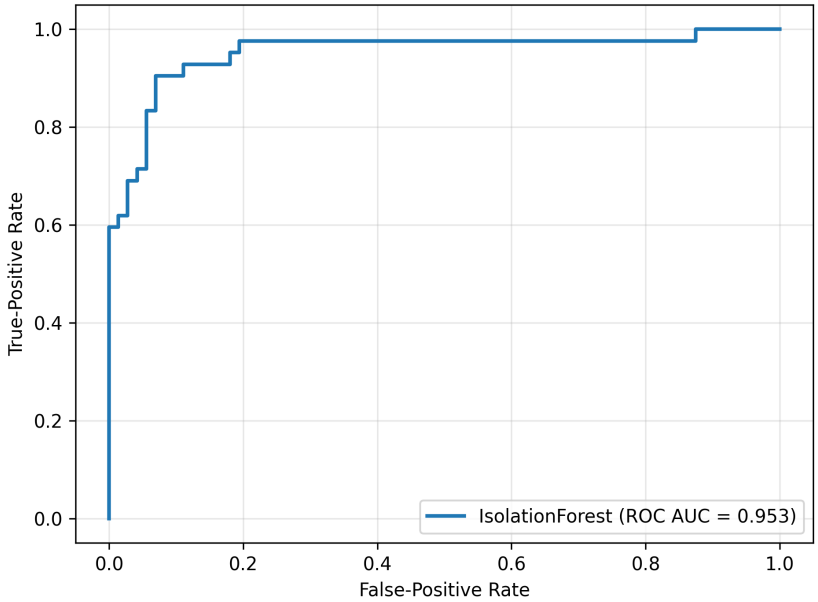
Results Overview

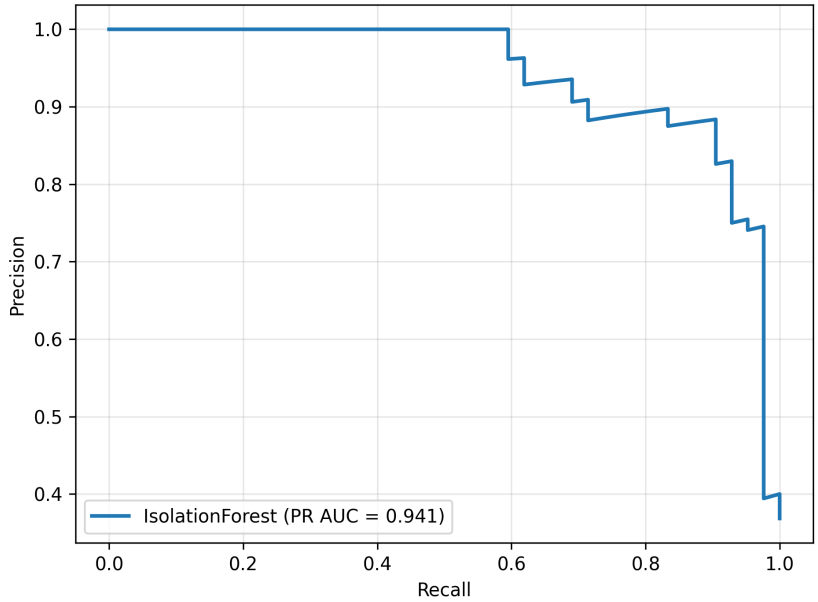


Key Findings

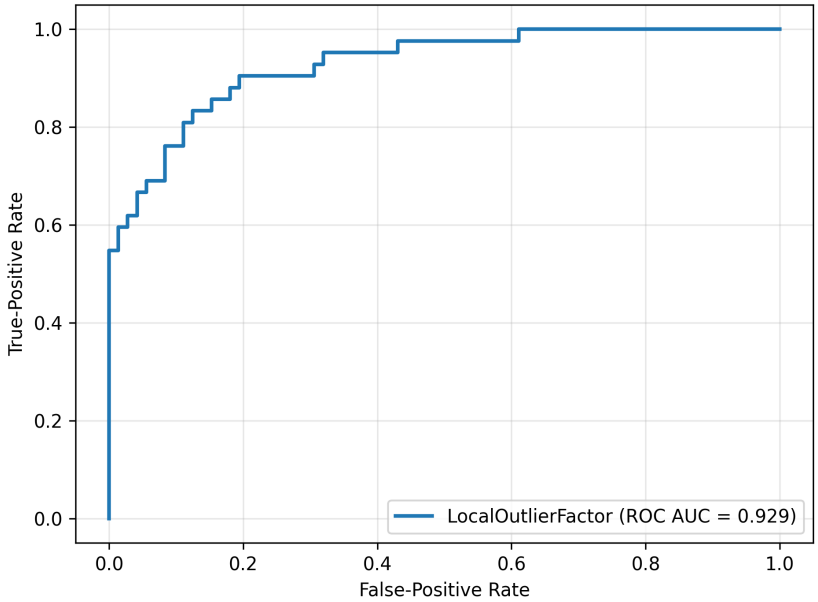
- **Isolation Forest** ranks first — PR AUC 0.94, ROC AUC 0.95. best performer.
- **Local Outlier Factor** follows — PR AUC 0.91, ROC AUC 0.93. Density checks work, yet high dimensionality blurs neighbourhoods and costs 4 pp in precision–recall.
- **One-Class SVM** trails — PR AUC 0.86, ROC AUC 0.92. The kernel boundary fits the overall data well, but it lets more malignant samples slip past the boundary.
- All three beat random guessing by a wide margin.

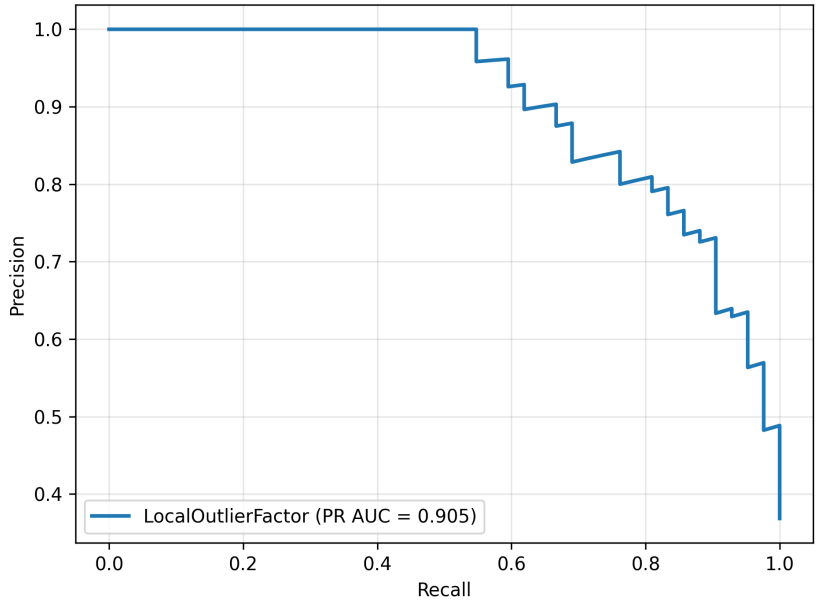
Isolation Forest – ROC and PR Curves



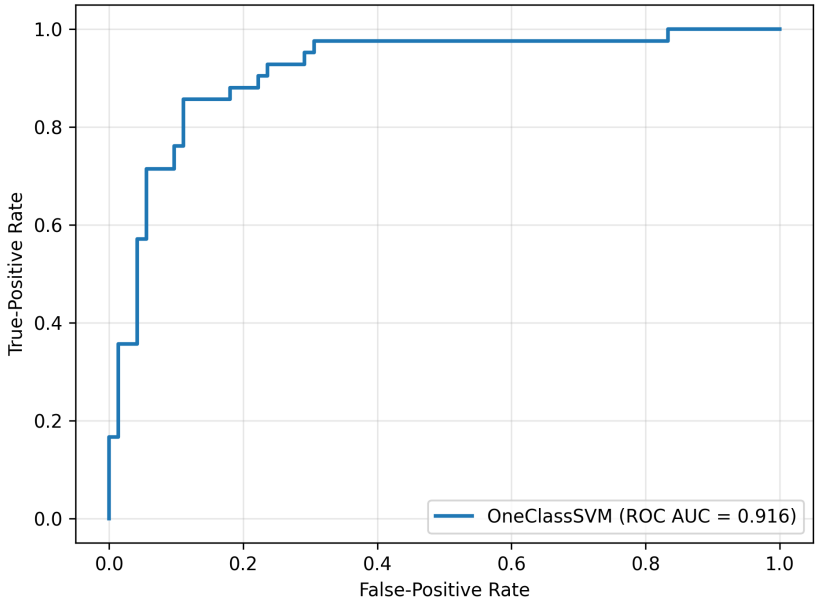


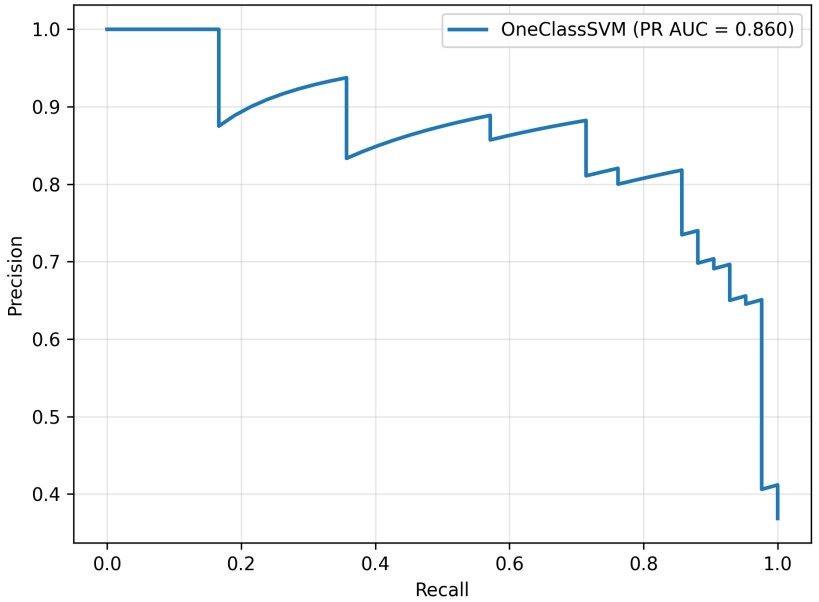
Local Outlier Factor – ROC and PR Curves





One-Class SVM – ROC and PR Curves





Open Problems

- **Dimensionality reduction** — the current pipeline keeps all 30 features; techniques such as PCA could cut noise and reveal stronger patterns (Aggarwal and Yu, 2001).
- **Hyperparameter tuning** — every model still runs on default settings; a grid search could lift accuracy and stability, especially for LOF and One-Class SVM.
- **Interpretability** — clinicians need clear reasons for each alert; today's scores explain little, so methods that map anomalies back to patient features are still critical. (Zimek et al., 2012).

Project Repository

BahadirAydin/
ceng562-ml

ceng562-ml

1 Contributor | 0 Issues | 0 Stars | 0 Forks

- **Code**: full pipeline and figures
- **Run**: `python main.py`
- **Reproduce**: same metrics, same plots

References I

Aggarwal, C. C., & Yu, P. S. (2001).Outlier detection for high dimensional data. *ACM SIGMOD Record, Volume 30, Issue 2*, 37–46.

Aktepe, S. C. (2022). Middle east technical university unofficial presentation template.

Dua, D., & Graff, C. (2019). Uci machine learning repository: Breast cancer wisconsin (diagnostic) data set [Accessed: 2025-05-11]. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011).Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*.

References II

Saito, T., & Rehmsmeier, M. (2015).The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS One*, *10*(3), e0118432.

Zimek, A., Schubert, E., & Kriegel, H.-P. (2012).A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *5*(5), 363–387.

Thank You

for your attention.

Do you have any question?