**Outlier Detection in ICU Patient Data Using Stratified Mahalanobis Distance**

**1. Introduction:** The detection of clinical outliers is a crucial task in healthcare analytics, particularly in identifying patients at higher risk of adverse outcomes. In this study, we utilize Mahalanobis distance to identify outliers in a medical dataset containing patient records from an intensive care unit (ICU). Given that physiological and lab parameters vary across gender, we implement a stratified approach to ensure fairness and statistical rigor.

**2. Dataset Overview:** The dataset includes 7,886 patient records with both continuous and categorical features. Key columns include vitals, lab tests, gender, ICU admission types, and the target variable In-hospital\_death.

**3. Preprocessing Pipeline**

* Columns with >50% missing data were dropped.
* Categorical columns (Gender, CCU, CSRU, SICU, MechVent) were imputed with mode and cast to category dtype.
* Continuous columns were standardized using StandardScaler.
* Missing values in continuous features were imputed using KNNImputer.

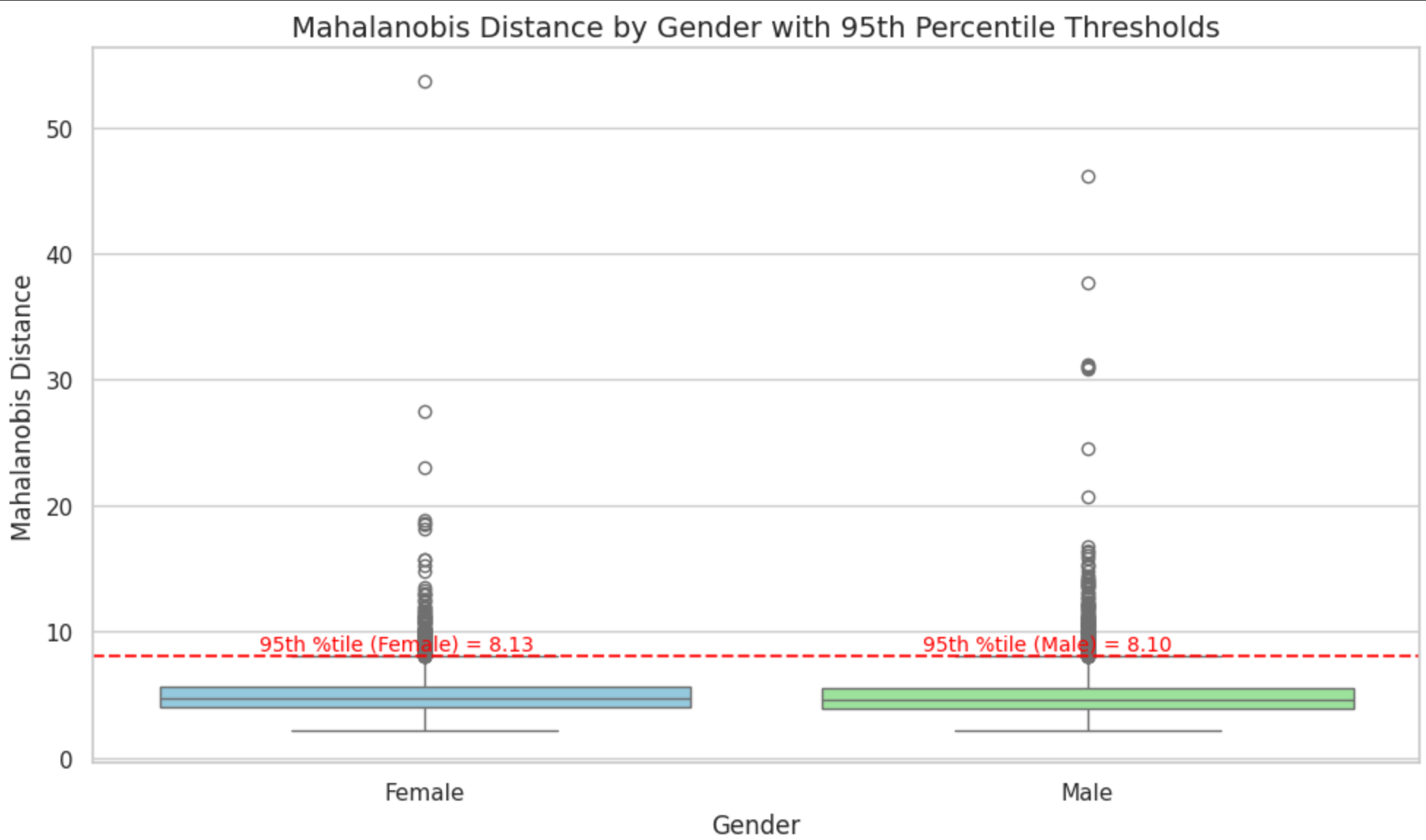
**4. Outlier Detection with Mahalanobis Distance:** Mahalanobis distance is a multivariate measure that accounts for correlations between variables. For each patient, Mahalanobis distance was computed **separately by gender** using group-specific mean vectors and covariance matrices. This stratification ensures that each individual is evaluated against a distribution that reflects their biological context.

**5. Outlier Thresholds**

* 95th and 99th percentile Mahalanobis distance thresholds were calculated **separately for males and females**.
* Patients above these thresholds were flagged as outliers.

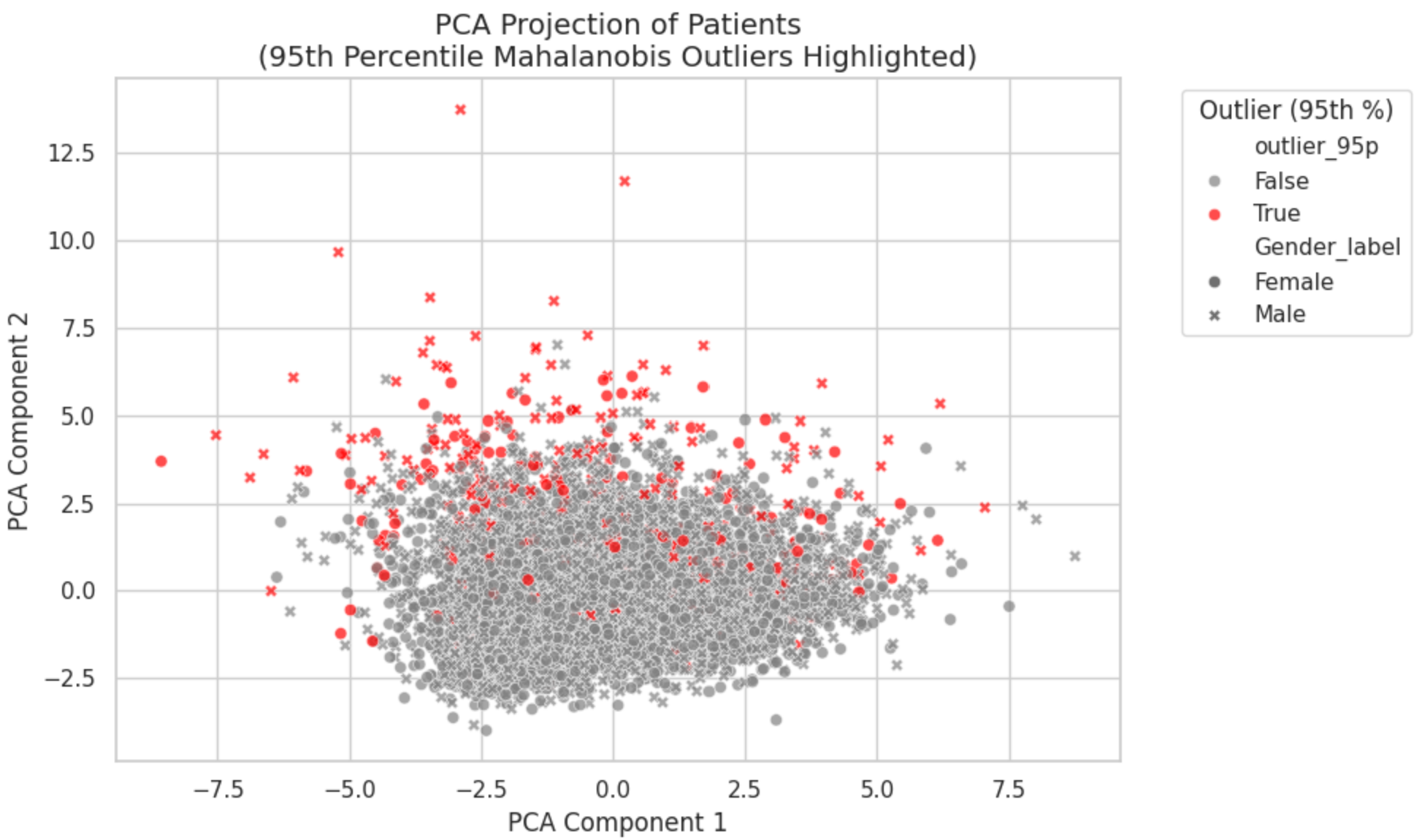
**6. Results and Visualization**

**6.1. Boxplot of Mahalanobis Distance by Gender:** A boxplot showed that both males and females had a similar spread, with notable extreme outliers above the 95th percentile threshold.

(Figure 1) 

**6.2. PCA Projection:** A PCA plot reduced the dataset to two principal components, showing how 95th percentile outliers were spatially separate from the main population cluster. This validated that Mahalanobis distance was effective in identifying unusual multivariate patterns.

(Figure 2)



**6.3. Mortality Rate Comparison:** Outliers had significantly higher in-hospital death rates:

**Overall Comparison (95th Percentile Outliers vs Others):**

**Total Patients Deaths Mortality Rate**

**outlier\_95p**

False 7490 1021 0.136315

True 396 101 0.255051

👩‍⚕️ **By Gender:**

**Total Patients** **Deaths** **Mortality Rate**

**Gender\_label** **outlier\_95p**

Female False 3269 475 0.145304

True 173 37 0.213873

Male False 4221 546 0.129353

True 223 64 0.286996

A graph showing a number of patients with a number of patients with their eyes closed

AI-generated content may be incorrect.

**7. Conclusion:** Stratified Mahalanobis distance is a powerful tool for outlier detection in healthcare datasets. By accounting for gender-specific distributions, we ensure fair and meaningful anomaly detection. Outliers flagged by this method show a higher risk of mortality, suggesting that Mahalanobis distance may aid in early risk identification.

**8. References**

* De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1), 1-18.
* Pedregosa et al., (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
* Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065), 20150202.