Anomaly Detection in Healthcare

Yuqi Chen, Henry Chien, Carlos Torres Angleró, Enyue Lu

The work is funded by NSF CNS-2149591 under Research Experiences for Undergraduates (REU) Program.

Abstract

Anomaly detection is a critical task in various domains, including healthcare, where identifying outliers can significantly impact patient outcomes and operational efficiency. This research report focuses on the progression of anomaly detection techniques applied to several healthcare-related datasets: Wisconsin Prognostic Breast Cancer (WPBC), Heart Disease, Annthyroid, and Pima Indians Diabetes.

Purpose

The primary objective of this study is to evaluate and determine the most effective machine learning methods for detecting anomalies in these datasets, providing insights into the performance and suitability of different algorithms in the context of healthcare data.

Methods

1. Random Forest
– Random Forest is an ensemble learning method used for classification, regression, and other tasks that operates by constructing multiple decision trees during training.
– It outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
– Random Forest corrects the habit of overfitting to their training set inherent in decision trees

2. Logistic Regression
– Logistic Regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable.
– It is used for binary classification problems, predicting the probability that a given input belongs to a certain class.
– Logistic Regression is the go-to method for binary classification due to its simplicity and effectiveness

3. Vote Classifier
– A Vote Classifier (or Voting Classifier) is an ensemble machine learning model that combines the predictions of multiple sub-models.
– It aggregates the predictions through voting, either majority voting (for classification) or averaging (for regression).
– Voting can be hard (majority voting) or soft (weighted voting based on predicted probabilities)

Results

	Accuracy						
	LOF	iForest	Sp	iNNE	RF	LR	VC
WPBC	0.73	0.69	0.7	0.68	0.83	0.77	**0.85**
Heart Disease	0.55	0.56	**0.63**	0.55	0.56	0.57	0.56
Annthyroid	0.9	0.9	0.86	0.84	0.93	0.92	**0.97**
Pima	0.65	0.67	**0.68**	0.65	0.65	0.65	0.65

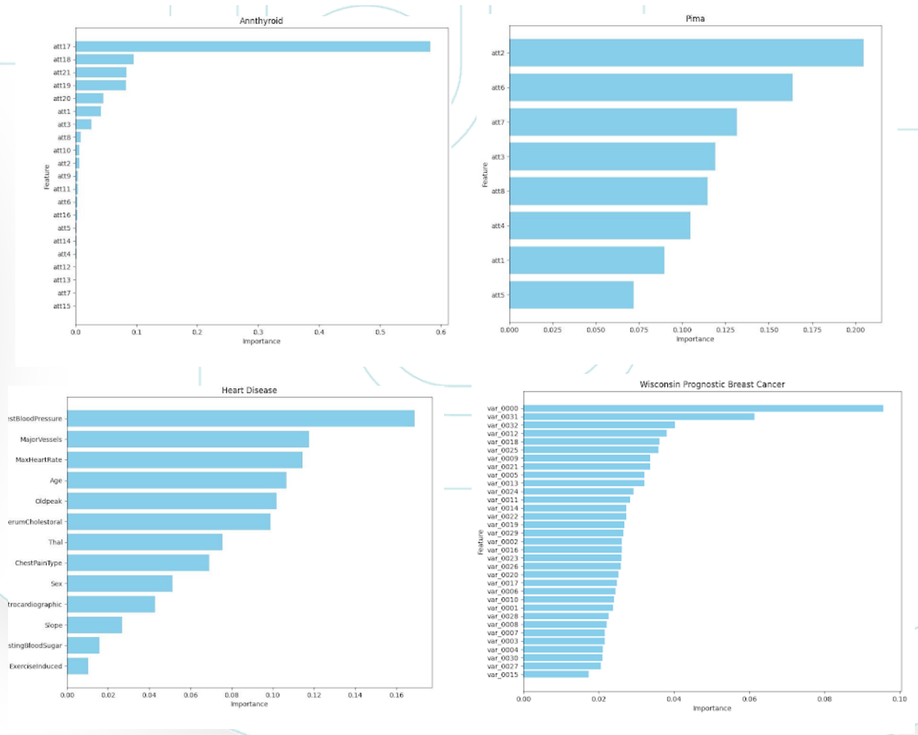
Table 2: Accuracy for Various Algorithms

	AUC						
	LOF	iForest	Sp	iNNE	RF	LR	VC
WPBC	0.5	0.48	0.48	0.5	0.63	**0.78**	0.73
Heart Disease	0.47	0.6	0.3	0.41	0.72	0.63	**0.82**
Annthyroid	0.33	0.63	0.31	0.49	0.95	0.74	**0.99**
Pima	0.42	0.68	0.35	0.31	0.52	0.80	**0.83**

Table 1: AUC Score for Various Algorithms

	Runtime						
	LOF	iForest	Sp	iNNE	RF	LR	VC
WPBC	0.11s	0.04s	0.13s	0.05s	0.04s	**0.0137s**	0.14s
Heart Disease	<**0.01s**	0.04s	0.13s	0.05s	0.03s	<**0.01s**	0.08s
Annthyroid	0.10s	**0.06s**	0.14s	0.24s	0.13s	0.14s	1.96s
Pima	<**0.01s**	0.05s	0.13s	0.05s	0.04s	<**0.01s**	0.21s

Table 3: Runtime for Various Algorithms



Conclusion

Our experiments with Random Forest, Logistic Regression, and our Voting Classifier have demonstrated significant improvements in anomaly detection within healthcare datasets. Random Forest, known for its robustness and ability to handle complex data structures, provided excellent performance in terms of accuracy and resilience to overfitting. On the other hand, Logistic Regression offered efficient computation and clear probabilistic interpretations, making it an effective tool for understanding and predicting anomalies. Additionally, our Voting Classifier achieved superior results on some datasets, proving to be the most formidable and useful method overall.
These results underscore the importance of leveraging diverse machine learning techniques to address the unique challenges presented by healthcare data. By exploring and validating these methods, we have laid a solid
foundation for more sophisticated and integrated approaches in the future.

References

[1] Samariya D., Ma J., Aryal S., Zhao X. (2023). Detection and explana-
tion of anomalies in healthcare data. Health Information Science and
Systems, 11(1):20. https://doi.org/10.1007/s13755-023-00221-2.
PMID: 37035724; PMCID: PMC10079801.

[2] Rayana, S., & Akoglu, L. (2016). Collective opinion spam detection using
active inference. Proceedings of the 2016 SIAM International Conference
on Data Mining. https://doi.org/10.1137/1.9781611974348.71