Semi-supervised Feature Selection for Efficient Detection of Systemic Deviations to Develop Trustworthy AI

Girmaw Abebe Tadesse*, **William Ogallo, Aisha Walcott-Bryant, Skyler Speakman**
IBM Research - Africa

Abstract

Trustworthy AI aims to achieve systems that support decision-making with data-driven insights while satisfying fundamental requirements such as explainability and fairness. Identifying systemic deviations in datasets and model outputs helps to validate fairness issues, such as bias to a certain subgroup. Multiple techniques have been proposed in the state-of-the-art to detect systemic deviations, but computational complexity grows with the dimension of the feature space. Thus, feature selection could be employed for efficient detection process. However, existing feature selection techniques are often conducted by optimizing the performance of prediction outcomes rather than systemic deviations. In this paper, we propose a sparsity-based and automated feature selection (SAFS) framework for efficient discovery of anomalous patterns, by encoding systemic outcome deviations via the sparsity of feature-driven odds ratios, without a supervised-training of a particular model. SAFS achieves more than $3\times$ reduction in computation time while maintaining detection performance, using just half of the original feature space, when validated on a publicly available critical care dataset. SAFS also results in superior performance when compared against multiple baselines for feature selection.

Introduction

Detection of systemic deviations or anomalous samples is a field of active research in trustworthy AI, and it aims to identify observations (a subgroup of samples) in a given data that deviate from some concept of normality (Ruff et al. 2021). The detection and characterization of anomalous samples helps to validate the fundamental requirements of trustworthy AI systems (such as explainability, transparency, robustness, and fairness) across different domains including healthcare, which is characterized by large differences in disease patterns, patient response to interventions, and cost of care across patient subpopulations (Ogallo et al. 2021; Kim et al. 2021; Senn 2016; Zhao et al. 2019). The challenges associated with anomalous detection primarily span three themes. First, the lack of representative examples of anomalous cases results in a significant imbalance that limits the choice of detection approaches. Second, the variations among in-distribution samples might be equivalent (even worse) compared to the deviation of anomalous samples, resulting in a rise in type-I and type-II errors. Finally, the scope of anomalousness could be too wide to model, with extreme variations of anomalous cases.

Nevertheless, a plethora of methods has been proposed for anomalous detection. The methods could be mainly categorized into *reconstruction*, *classification* and *probabilistic* groups (Ruff et al. 2021). The well-known principal component analysis and autoencoders are examples of reconstruction-based methods that, first, transform the data (e.g., to a latent space) so that anomalousness could be detected from failing to reconstruct the data back from the transformed data (Hawkins et al. 2002). Classification-based approaches, particularly one-class classification is often employed due to the lack of examples representing anomalous cases (Tax 2002; Khan and Madden 2014). Furthermore, the traditional probabilistic models have also been used to identify anomalous samples using estimation of the normal data probability distribution, e.g., Gaussian mixture models (Roberts and Tarassenko 1994) and Mahalanobis distance evaluation (Laurikkala et al. 2000). Moreover, there are purely distance-based methods, such as k-nearest neighborhood (Gu, Akoglu, and Rinaldo 2019), that do not require a prior training phase nor data transformations. Of note is that most existing methods infer anomalousness by exploiting individual sample characteristics rather than group-based characteristics. To this end, researchers have developed the Multi-dimensional subset scanning (MDSS), a method that aims to identify subsets of anomalous samples by exploiting group-level characteristics (McFowland III, Somanchi, and Neill 2018; Cintas et al. 2021).

A significant gap in the state-of-the-art anomalous subgroup discovery concerns the lack of a principled and scalable feature selection step for efficient discovery. Most discovery techniques require manual selection of features (e.g., using domain experts) or use the whole input space, often resulting in inefficient discovery, particularly in cases of larger input feature spaces, since the computational complexity grows significantly with the number of features resulting in less optimal anomalous discovery. Furthermore, the lack of pre-discovery feature selection step might result in less

* Corresponding Author
Email: girmaw.abebe.tadesse@ibm.com

interpretable anomalous characterization (because of two many features characterizing the anomalous subset). Automated feature selection steps could be employed to solve the above problem by selecting useful features for the subsequent discovery step. However, existing feature selection techniques mainly optimize over higher outcome prediction performance of a trained model (Molina, Belanche, and Nebot 2002; Miao and Niu 2016; Wanjiru et al. 2021), and hence they are limited for encoding systemic outcome deviations among subsets of the data. Moreover, model training results in computational overhead, and the feature selection output is prone to model hyper-parameters, class imbalance, and underfitting or overfitting.

In this paper, we proposed a sparsity-based automated feature selection (SAFS) framework, which is *model-free* as it does not require training a particular model, i.e., semi-supervised. SAFS encodes systemic outcome deviations using the sparsity of the feature-driven odds ratios. The proposed feature selection framework is simple and generalizable as it could be applied as a simple pre-processing step to any anomalous discovery technique for tabular data formats. Specifically, the contributions of the paper are as follows.

1. We propose a semi-supervised feature selection method optimized to achieve efficient detection of systemic deviations (not improving prediction outcome), without training a particular model.
2. We validate the proposed method on the publicly available MIMIC-III (Medical Information Mart for Intensive Care) dataset(Johnson et al. 2016), and results show that SAFS achieves similar anomalous subgroup discovery using just half of the features selected while providing more than $3\times$ reduction in computational time.
3. Furthermore, we demonstrate that SAFS results in superior detection performance when compared with existing feature selection techniques including Filters (Molina, Belanche, and Nebot 2002; Vergara and Estévez 2014), Wrappers (Miao and Niu 2016), and Embedded techniques that use tree-based models, such as XGBoost (Chen and Guestrin 2016) and Catboost (Hancock and Khoshgoftaar 2020).

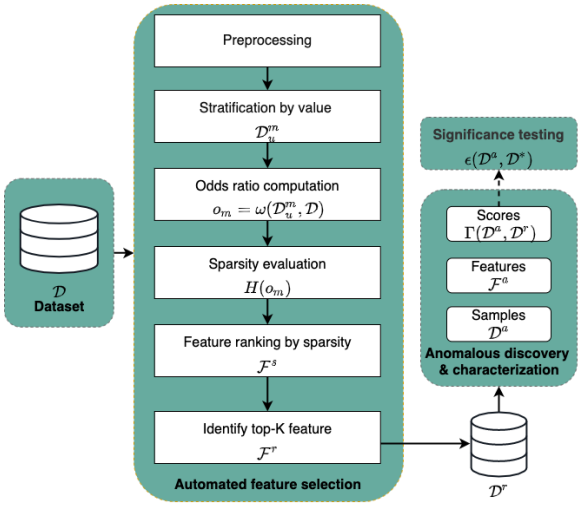
Proposed framework

The proposed framework is shown in Fig. 1 and it contains two main components: *automated feature selection (SAFS)* and *anomalous discovery and characterization*. The automatic feature selection exploits the sparsity of odd ratios computed per each feature value. The anomalous discovery step scans across all possible combinations of the values of the selected features and characterizes those identified to be divergent from the expectation (or normalcy). Each component of the framework is described below in detail.

Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i)|i = 1, 2, \cdots, N\}$ denotes a dataset containing N samples, and each sample x_i is characterised by a set of M discretized features $\mathcal{F} = [f_1, f_2, \cdots, f_m, \cdots, f_M]$ and y_i represents the outcome label. Note that each feature

Figure 1: SAFS framework for automatically selecting important features for anomalous subgroup discovery.



f_m has C_m unique values, $\hat{f}_m = \{\hat{f}_m^u\}_{u=1}^{C_m}$. The proposed automated feature selection process is defined as $\mathcal{R}(\cdot)$ that takes $\mathcal{D}$ as input and provides $\mathcal{D}^r$ represented with a top K features, i.e., $\mathcal{D}^r = R(\mathcal{D}, K) = \{(x_i^r, y_i)|i = 1, 2, \cdots, N\}$ and x_i^r is represented by $\mathcal{F}^r = \{f_1^r, f_2^r, \cdots, f_k^r, \cdots, f_K^r\}$, where $K \leq M$. Then the anomalous subgroup discovery, $S(\cdot)$, takes $\mathcal{D}^r$ as input and identifies the anomalous subgroup ($\mathcal{D}^a$) represented by the logical (AND and OR) combinations of anomlaous feature values as $\hat{\mathcal{F}}^a = \bigcap_{z=1}^{Z}(\bigcup_{h=1}^{H_z} \hat{f_{zh}^a})$, where $\hat{f_{zh}^a}$ represents the h^{th} value of the f_z^a and $H_z < C_z$. The anomalous subgroup $\hat{\mathcal{D}}^a$ contains samples from $\mathcal{D}^r$ whose feature values are characterized by $\hat{\mathcal{F}}^a$, i.e., $\mathcal{D}^a = \{(x_j^a, y_j^a)|j = 1, 2, \cdots, P\}$, where $P < N$. The anomalousness of the identified subgroup is evaluated based on the anomalous score.

Automated Feature Selection

The sparsity-based automated feature selection (SAFS) component in Fig. 1 is tasked with selecting the top K features from a given M-dimensional feature space that are more useful for the follow-up anomalous subgroup discovery. To this end, SAFS employs sparsity of the odds ratios of feature values to rank and select features. A highly-ranked feature for anomalous discovery is assumed to have high variation of the odds ratios of its feature values. The variation is quantified using a sparsity metric. Given a feature f_m with C_m unique values, we manually stratify $\mathcal{D}$ per each feature value $\hat{f}_m^u \in \hat{f}_m$, i.e., $\mathcal{D}_m^u = \mathcal{D}|\hat{f}_m == \hat{f}_m^u$. The mean of the outcome in the stratified $\mathcal{D}_m^u$ is computed as $\mu_m^u = \frac{\sum_{j=1}^{N_m^u} y_j}{N_m^u}$, where N_m^u is the number of samples in $\mathcal{D}_m^u$. Similarly, the global average of the outcome is computed as $\mu_g = \frac{\sum_{i=1}^{N} y_i}{N}$, where N is the total number of samples in $\mathcal{D}$. Thus, ratios of the odds of the outcome in $\mathcal{D}_m^u$ and in $\mathcal{D}$

Algorithm 1: Pseudo-code for automated feature selection based on sparsity of odds ratios

input : Dataset: $\mathcal{D} = \{(x_i, y_i)|i = 1, 2, \cdots, N\}$, Set of features: $\mathcal{F} = [f_1, f_2, \cdots, f_m, \cdots, f_M]$, Required number of features: K.
output: Set of selected features: $\mathcal{F}^r$

$\eta \leftarrow$ `ZerosArray`(M);
for f_m *in* $\mathcal{F}$ **do**
  $\hat{f}_m \leftarrow$ `IdentifyUniqueValues`(f_m);
  $C_m \leftarrow |\hat{f}_m|$;
  $o_m \leftarrow$ `ZerosArray`(C_m);
  for $u \leftarrow 1$ **to** C_m **do**
    $\mathcal{D}_m^u \leftarrow$ `Stratification` $(\mathcal{D}, \hat{f}_m^u)$;
    $o_m^u \leftarrow$ `OddsRatio` $(\mathcal{D}_m^u, \mathcal{D})$;
  $\eta_m \leftarrow$ `MeasureSparsity` (o_m);
$I \leftarrow$ `SortDescendingIndices` (η);
$\mathcal{F}^s \leftarrow \mathcal{F}[I]$;
$\mathcal{F}^r \leftarrow$ `TopK` $(\mathcal{F}^s, K)$;
return $\mathcal{F}^r$

is computed as:

$$o_m^u = \frac{\mu_m^u/(1-\mu_m^u)}{\mu_g/(1-\mu_g)} \tag{1}$$

To compute the sparsity of the odd ratios in o_m, we use the Hoyer sparsity metric (Hoyer 2004) that was proven to satisfy key requirements of sparsity (Hurley and Rickard 2009) as follows:

$$\eta_m = (\sqrt{C_m} - \frac{\sum_{u=1}^{C_m} o_m^u}{\sum_{u=1}^{C_m} o_m^{2\ u}})(\sqrt{C_m} - 1)^{-1} \tag{2}$$

The summary of the steps for sparsity-based feature selection is shown in Algorithm 1.

Anomalous Discovery and Characterization

We employ Multi-Dimensional Subset Scanning (MDSS) (McFowland III, Somanchi, and Neill 2018; Cintas et al. 2021) from the anomalous pattern detection literature in order to identify significantly divergent subset of samples. Characterization of the identified samples includes quantifying the anomalousness score, the analysis of the anomalous features and their values, the time elapsed to identify them, and the statistical significance of these findings. MDSS could be posed as a search problem over possible subsets in a multi-dimensional array to identify systematic deviation between observation (i.e., y_i) and expectation of the outcomes, which could be set differently for variants of MDSS. In the simple automatic stratification setting, the expectation is the global outcome average in $\mathcal{D}^r$, i.e., μ_g. The deviation between the expectation and observation is evaluated by maximizing a Bernoulli likelihood ratio scoring statistic, $\Gamma(\cdot)$. The null hypothesis assumes that the likelihood of the outcome in each sample $x_i^r \in \mathcal{D}^r$ or subgroup is similar to the expected (μ_g), i.e., $H_0 : odds(y_i) = \frac{\mu_g}{1-\mu_g}$; while the alternative hypothesis assumes a constant multiplicative increase in the outcome odds for the anomalous subgroup, $H_1 : odds(y_i) = q\frac{\mu_g}{1-\mu_g}$ where $q \neq 1$ ($q > 1$ for extremely over observed subgroup; and $0 < q < 1$ for extremely under observed subgroup). The anomalous scoring function for a subgroup ($\mathcal{D}^s$) with reference $\mathcal{D}^r$ is formulated as, $\Gamma(\mathcal{D}^s, \mathcal{D}^r)$ and computed as:

$$\Gamma(\mathcal{D}^s, \mathcal{D}^r) = \max_q log(q) \sum_{i \in S} y_i - N_S * log(1 - \mu_g + q\mu_g), \tag{3}$$

where N_S is the number of samples in $\mathcal{D}^s$. Consequently, subsets in which average of outcome different from μ_g will have higher scores. Subset identification is iterated until convergence to a local maximum is found, and the global maximum is subsequently optimized using multiple random restarts. The subset ($\mathcal{D}^s$) with the highest score becomes anomalous subset $\mathcal{D}^a$ and it is characterized by its score of anomalousness, $\Gamma(\mathcal{D}^a, \mathcal{D}^r)$ and a combination of feature values $\hat{\mathcal{F}}^a$ describing the identified $\mathcal{D}^a$.

The statistical significance of SAFS is evaluated using a randomization testing. The null hypothesis suggests $\Gamma(\mathcal{D}^a, \mathcal{D}^r)$ is not significantly different from a set of $\Gamma(\mathcal{D}^*, \mathcal{D}^r)$, where $\mathcal{D}^*$ represents the anomalous subset obtained from $\mathcal{D}^{r*}$ by randomly selecting K features from $\mathcal{D}$. This experiment is performed iteratively $\sigma = 100$ times resulting $\delta = \{\Gamma(\mathcal{D}_t^*, \mathcal{D}^r)\}_{t=1}^{\sigma}$. We compute the empirical p-value as $\frac{(\xi+1)}{(\sigma+1)}$ where ξ is the number of scores in δ that are greater than or equal to the actual score $\Gamma(\mathcal{D}^a, \mathcal{D}^r)$.

Experiments

Dataset and Experimental Setup

We used the Medical Information Mart for Intensive Care (MIMIC-III) dataset(Johnson et al. 2016) to validate the proposed framework. We selected a study cohort of adult patients (16 years or older) who were admitted to the ICU for the first time, where the length of stay was greater than a day, and with no hospital readmissions, no surgical cases, and having at least one chart events. The final cohort consisted of $N = 18,761$ patients. We constructed $M = 41$ features based on observations made on the first 24 hours of ICU admission. We defined the target outcome as a binary indicator variable y_i such that $y_i = 1$ for patients who who died within 28 days of the onset of their ICU admission, and $y_i = 0$ otherwise. For the automatic selection step, we set top K to different values $\{0.1 \times l \times 41\}_{l=1}^{10}$ resulting $K \in [4, 8, 12, 16, 20, 25, 29, 33, 37, 41]$.

Results and Discussion

We selected different top K features using SAFS and apply MDSS to identify the anomalous subset characterized by more observations of deaths compared to the global average.

Detection performance vs. Elapsed time: Figure 2 illustrates the anomalous score and the elapsed time to complete the scanning across the top K selected features. These results show that it is possible to achieve comparable anomalous scores by scanning over the top 20 features (half of the

Table 1: Subpopulation size, odds ratio, and significance (empirical p-value) of the most anomalous subgroup when scanning over the top K features identified by SAFS.

Top K	Size	Odds Ratio (95% CI)	P-Value
4	4312	2.48 (2.3, 2.67)	< 0.01
8	2811	3.00 (2.75, 3.26)	< 0.01
12	4383	2.48 (2.31, 2.67)	< 0.01
16	4383	2.48 (2.31, 2.67)	< 0.01
20	4383	2.48 (2.31, 2.67)	< 0.01
25	3078	2.93 (2.7, 3.18)	< 0.01
29	3078	2.93 (2.7, 3.18)	< 0.01
33	3078	2.93 (2.7, 3.18)	< 0.01
37	3078	2.93 (2.7, 3.18)	< 0.01
41	4218	2.54 (2.36, 2.73)	< 0.01

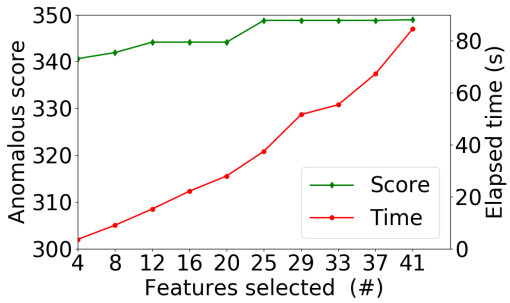


Figure 2: Anomalous subgroup scores (green) and elapsed scanning time (red) across top K features selected by SAFS.

original) identified by SAFS, with more than $3\times$ reduction in elapsed scanning time.

Consistency of discovered subgroups: We also evaluated the consistency of the anomalous group across different top K values as shown in Table 1 and Fig. 3. The results demonstrate consistently competitive performance across different top K values without a loss of performance in anomalous discovery. Fig. 3 shows that consistent features are identified across these K values. Features *angus (severe sepsis)* and *curr_service (type of current ICU service)* represented the anomalous group in all top K values, and *urine output* appeared in eight cases out of 10 different K values. Moreover, these most frequently occurring features are ranked higher during the feature selection step, validating the effectiveness of SAFS in selecting features that would be useful for anomalous pattern discovery.

Comparison with existing methods: Table 2 shows the comparison of SAFS with multiple state-of-the-art feature selection approaches including a filter method using mutual information gain (Molina, Belanche, and Nebot 2002; Vergara and Estévez 2014), a wrapper method using recursive feature elimination (Guyon et al. 2002), and embedded methods using XGBoost (Chen and Guestrin 2016) and Catboost (Hancock and Khoshgoftaar 2020) and committee vote based on the average of feature importance from XGBoost and Catboost. The results demonstrate the superior performance of SAFS in achieving the highest anomalous scores, particularly with $K \leq 29$ features. All methods become competitive for larger K values.

Table 2: Comparison of the proposed SAFS to existing feature selection methods for different top K values. The existing methods are used as baselines and include Filter (Molina, Belanche, and Nebot 2002), Wrapper (Miao and Niu 2016), XGB: Extreme Gradient Boosting (Chen and Guestrin 2016), CatB:Categorical Boosting (Hancock and Khoshgoftaar 2020) and Committee: average of XGB and CatB. SAFS is shown to outperform others with its highest anomalous scores for $K \leq 25$ features. The methods become competitive afterwards.

	Baseline methods					**Proposed**
			Embedded methods			
K	Filter	Wrapper	XGB	CatB	Committee	SAFS
4	337.19	311.27	337.19	339.23	337.19	**340.61**
8	340.61	311.27	337.19	340.61	340.61	**341.90**
12	340.61	311.27	340.61	340.61	340.61	**344.13**
16	340.61	316.09	340.61	**344.13**	340.61	**344.13**
20	340.61	340.86	340.61	**344.13**	**344.13**	**344.13**
25	340.61	346.64	348.53	348.45	348.53	**348.77**
29	344.59	346.64	**348.77**	**348.77**	**348.77**	**348.77**
33	**348.77**	346.64	**348.77**	**348.77**	**348.77**	**348.77**
37	348.45	**348.91**	**348.91**	**348.91**	**348.91**	348.77
41	**348.91**	**348.91**	**348.91**	**348.91**	**348.91**	**348.91**

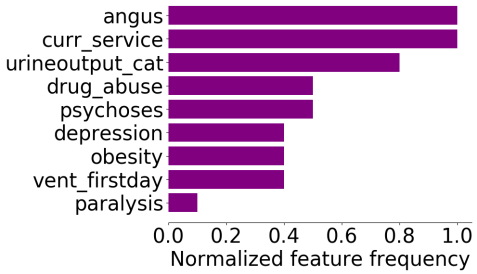


Figure 3: Overlapping of anomalous features detected under different top K values.

Conclusion and Future work

We proposed a sparsity-based automated feature selection (SAFS) framework for efficient detection of systemic deviations (anomalous subgroups) without training a particular model. Efficiency is achieved by reducing the feature space thereby the amount of time required to complete the detection, which often grows exponentially with the number of features. The reduction in feature space also reduces the number of optimization steps to approximate global optima. SAFS uses the feature-driven deviation of outcome likelihood via the sparsity of the odds ratios to encode systemic deviations. SAFS outperformed multiple baseline feature selection methods and achieved more than $3\times$ reduction in computational time but with competitive detection performance using just half of the features. Future work aims to extend SAFS to select layers and nodes in deep learning frameworks that employ activations and reconstruction errors to identify anomalous patterns in other modalities.

References

Chen, T., and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Cintas, C.; Speakman, S.; Tadesse, G. A.; Akinwande, V.; McFowland III, E.; and Weldemariam, K. 2021. Pattern detection in the activation space for identifying synthesized content. *arXiv preprint arXiv:2105.12479*.

Gu, X.; Akoglu, L.; and Rinaldo, A. 2019. Statistical analysis of nearest neighbor methods for anomaly detection. *arXiv preprint arXiv:1907.03813*.

Guyon, I.; Weston, J.; Barnhill, S.; and Vapnik, V. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning* 46(1):389–422.

Hancock, J. T., and Khoshgoftaar, T. M. 2020. Catboost for big data: an interdisciplinary review. *Journal of Big Data* 7(1):1–45.

Hawkins, S.; He, H.; Williams, G.; and Baxter, R. 2002. Outlier detection using replicator neural networks. In *Proceedings of International Conference on Data Warehousing and Knowledge Discovery*, 170–180.

Hoyer, P. O. 2004. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research* 5(9).

Hurley, N., and Rickard, S. 2009. Comparing measures of sparsity. *IEEE Transactions on Information Theory* 55(10):4723–4741.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Li-Wei, H. L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data* 3(1):1–9.

Khan, S. S., and Madden, M. G. 2014. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review* 29(3):345–374.

Kim, H.; Tadesse, G. A.; Cintas, C.; Speakman, S.; and Varshney, K. 2021. Out-of-distribution detection in dermatology using input perturbation and subset scanning. *arXiv preprint arXiv:2105.11160*.

Laurikkala, J.; Juhola, M.; Kentala, E.; Lavrac, N.; Miksch, S.; and Kavsek, B. 2000. Informal identification of outliers in medical data. In *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology*, volume 1, 20–24.

McFowland III, E.; Somanchi, S.; and Neill, D. B. 2018. Efficient discovery of heterogeneous treatment effects in randomized experiments via anomalous pattern detection. *arXiv preprint arXiv:1803.09159*.

Miao, J., and Niu, L. 2016. A survey on feature selection. 91:919–926.

Molina, L.; Belanche, L.; and Nebot, A. 2002. Feature selection algorithms: a survey and experimental evaluation. *Proceedings of IEEE International Conference on Data Mining*.

Ogallo, W.; Tadesse, G. A.; Speakman, S.; and Walcott-Bryant, A. 2021. Detection of anomalous patterns associated with the impact of medications on 30-day hospital readmission rates in diabetes care. In *AMIA Annual Symposium Proceedings*, volume 2021, 495. American Medical Informatics Association.

Roberts, S., and Tarassenko, L. 1994. A probabilistic resource allocating network for novelty detection. *Neural Computation* 6(2):270–284.

Ruff, L.; Kauffmann, J. R.; Vandermeulen, R. A.; Montavon, G.; Samek, W.; Kloft, M.; Dietterich, T. G.; and Müller, K.-R. 2021. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*.

Senn, S. 2016. Mastering variation: variance components and personalised medicine. *Statistics in Medicine* 35(7):966–977.

Tax, D. M. J. 2002. One-class classification: Concept learning in the absence of counter-examples.

Vergara, J. R., and Estévez, P. A. 2014. A review of feature selection methods based on mutual information. *Neural Computing and Applications* 24(1):175–186.

Wanjiru, C.; Ogallo, W.; Tadesse, G. A.; Wachira, C.; Mulang, I. O.; and Walcott-Bryant, A. 2021. Automated supervised feature selection for differentiated patterns of care. *arXiv preprint arXiv:2111.03495*.

Zhao, R.; Yan, R.; Chen, Z.; Mao, K.; Wang, P.; and Gao, R. X. 2019. Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing* 115:213–237.