Open-Set Recognition of Breast Cancer Treatments

Alexander Cao[a,*], Diego Klabjan[a] and Yuan Luo[b,**]

[a]Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, USA
[b]Department of Preventive Medicine, Northwestern University, Chicago, IL, USA

ARTICLE INFO

Keywords:
Breast cancer treatments
Open-set recognition
Deep neural networks
Autoencoder

ABSTRACT

Open-set recognition generalizes a classification task by classifying test samples as one of the known classes from training or "unknown." As novel cancer drug cocktails with improved treatment are continually discovered, classifying patients by treatments can naturally be formulated in terms of an open-set recognition problem. Drawbacks, due to modeling unknown samples during training, arise from straightforward implementations of prior work in healthcare open-set learning. Accordingly, we reframe the problem methodology and apply a recent Gaussian mixture variational autoencoder model, which achieves state-of-the-art results for image datasets, to breast cancer patient data. Not only do we obtain more accurate and robust classification results (14% average F1 increase compared to recent methods), but we also reexamine open-set recognition in terms of deployability to a clinical setting.

1. Introduction

Most prior work in classification is closed-set; meaning the classes are assumed to be the same for both training and testing. Only relatively recently have classifiers designed for open-set evaluation, where unknown classes appear only in testing, gained attention as a real-world necessity. In particular, open-set image recognition arises from increasingly automated computer vision systems such as those in self-driving cars. It would certainly be impossible to include every object class that could possibly be seen while driving in model training [1]. Given the inherent dynamism of healthcare, one can argue a greater need for open-set classifiers. Some diseases are too rare to include sufficient samples in training [2]. There is also a consistent cycle of identifying novel diseases and developing treatments for them (recently Coronavirus 2019, for instance). Both of these circumstances necessitate generalizing medical classification tasks to open-set recognition. We base this paper on a more prevalent circumstance: new combinations of known drugs for existing diseases. Accordingly, we seek to classify patients by cocktail treatments based on medical and demographic features.

Personalized medicine through quantitative, phenotypic profiling shows promise in medical care by guiding drug combination strategies [3, 4]. In cancer treatments, these drug combinations are becoming the standard of care and many drug combination therapies have been approved or are under clinical trials [5, 6, 7]. The landscape of cancer drug combinations, or "cocktails," evolves with discoveries of novel drugs as well as novel combinations of existing drugs with improved treatment and lessening side effects. Although some guidelines exist for certain cancer types, individual patients' responses to various drug combinations are still not well understood. For instance, which drug combinations would likely benefit a specific patient the most is still a critical, open question. In this vein, we formulate classifying patient encounters by cancer cocktail treatments as an open-set problem. Figure 1 illustrates this via graphical flowchart.

Again, our goal is to classify patients by cocktail treatments based on medical and demographic features. In addition, sufficiently unique patients unlike those historically associated with known cocktails (i.e., cocktails in the training set) should be classified as "novel." This "novel" class is an indication that different or new cocktails may be more suitable for those patients' treatments. While "novel" classification does not provide any information regarding what might that novel cocktail be, it can still be of use to a physician. A "novel" classification can indicate that a different combination of existing drugs can lead to a better treatment or lessen side effects. Our open-set recognition is not motivated by drug discovery but rather by personalized medicine via individual attention treatment plans, prompting physicians to consider accessible cocktail modifications based on recent lab tests or symptoms. In this vein, it is apt to cast this as drug repurposing. We may have U.S. Food and Drug Administration (FDA) approved individual drugs that we may not be aware of their combined effects, or there may be new individual drugs approved, that when added to existing drugs form an effective combination. A data driven approach like open-set recognition can help us find such a needle in a haystack. To our knowledge, this is the first application of open-set recognition to cancer treatment classification.

In this paper, we focus on the open-set learning variant of training (and validating) on only the C known classes for $(C+1)$-class classification during inference. The $(C+1)$-th class aggregates all novel test samples not belonging to the known classes. To reflect a real-world scenario, we do not have samples from "novel" cocktails during the training and validation phases. For this pilot study, we focus on predicting

[*]Corresponding author at: Technological Institute C138, 2145 Sheridan Rd, Evanston, IL, 60208, USA.

[**]Corresponding author at: Rubloff Building 11th floor, 750 N Lake Shore Dr, Chicago, IL, 60611, USA.

✉ a-cao@u.northwestern.edu (A. Cao); d-klabjan@northwestern.edu (D. Klabjan); yuan.luo@northwestern.edu (Y. Luo)

ORCID(s):

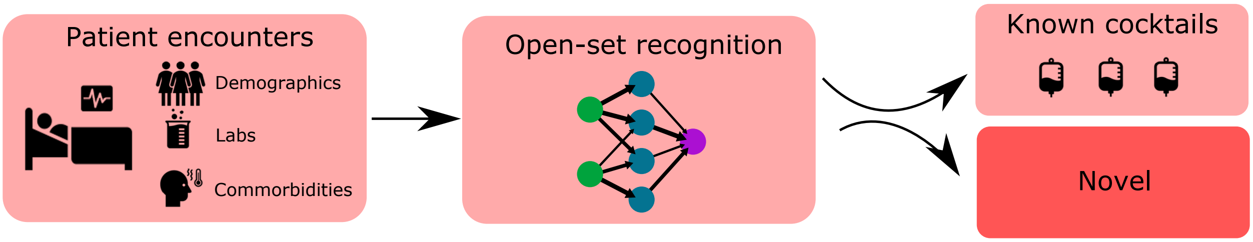

Figure 1: Information from cancer patient encounters, such as demographics, lab results, and commorbidities should inform which drug cocktail treatment is best. The goal is to use these features in an open-set recognition model to discern if the patient should receive one of the known, existing drug cocktail treatment, and if so which one. However, if the open-set recognition model instead returns "novel," this is an indication that a novel cocktail treatment (novel drug or novel combination of existing drugs) may be beneficial.

if a patient will benefit from a novel cocktail treatment versus known cocktails. Previous healthcare open-set studies rely on the use of fabricated or auxiliary data and standard softmax classifiers [2, 8]. Bypassing this data necessity, we adapt the existing Gaussian Mixture Variational Autoencoder (GMVAE) model and related "uncertainty" threshold [9], which achieves state-of-the-art results in open-set image recognition, to our open-set cancer treatment recognition task. We apply these methods to breast cancer patients' electronic health records (EHRs) from Northwestern Memorial Hospital. In doing so, this study achieves a step towards implementing a system to help physicians identify cancer patients who may benefit from a novel drug cocktail in a real-time clinical setting.

Our paper is organized as follows. In §2, we compare related work along with a comprehensive summary of the benchmark models [10, 11, 12]. Following in §3, we first provide background on the GMVAE model coupled with the "uncertainty" threshold [9]. In particular, we emphasize the intuition behind dual reconstruction-classification learning and "uncertainty" for open-set recognition. Next, we present the complete experimental design from data feature engineering to model evaluation. Subsequently in §4, we conduct open-set recognition experiments on our breast cancer patient dataset.

From these experimental results, we stress two findings, which are our main contributions. First, GMVAE outperforms other open-set classifiers both in terms of accuracy and robustness to an increasing number of unknown cocktails. Second, relevant prior methods [2, 8, 10] bypass selecting a single optimal threshold for rejecting unknowns by reporting area under the curve (AUC) or receiver operating characteristic (ROC) metrics or simply assuming a binary known-unknown false positive rate. However, all of these are uninformative for actual model deployment where a single threshold would be used for decisions. In contrast, GMVAE combined with "uncertainty" showcases an intuitive heuristic for selecting a single, optimal threshold. We emphasize this is a more practical model evaluation comparison. Summary ROC metrics can be useful in comparing different models in a holistic sense. However, in terms of real-world model usage in a clinical setting, it is more apt to compare actual decision accuracies which are only apparent after choosing a threshold. Finally in §5 and §6, we end with a discussion on limitations and future work, and conclude.

2. Related work

For literature placement, it is important to note that open-set recognition reduces to outlier detection in the case where the number of known classes $C = 1$ (viewed as a "normal" class). Outlier detection, or the related novelty or anomaly detection, is a longer studied topic [1, 13, 14, 15, 16]. Such methods are utilized in healthcare to detect outliers in breast cancer survivability predictions [17] and anomalous activity in EHRs [18]. Outlier detection, however, does not generally extend to differentiating between multiple known classes hence open-set recognition. For instance, in our breast cancer patient experiments there are three and four known cocktail classes.

There is an immense body of existing work concerning traditional closed-set classification. Open-set recognition, on the other hand, is only recently receiving more consideration. Earlier examples of $(C+1)$-class classification employ support vector machines (SVMs) [12, 19, 20] or sparse representation [21]. Open-set recognition in conjunction with deep neural networks is the current trend [2, 11, 22, 23]. However, these methods are almost exclusively designed solely for image recognition; network architecture reliance on image patching, channel activation, spatial pooling, feature map modulation, and pixel reconstruction inhibit usability for non-image-based tasks (such as ours).

While image classification benefits from a well-rounded surge in open-set recognition, applications to general healthcare data are wanting. Specifically in [2], eye diseases are open-set classified using optical coherence tomography (OCT) images but the method is contingent on a patchGAN-derived model [24] to generate synthetic, "boundary" images that are deliberately difficult to classify with a pre-trained, closed-set softmax classifier. These manufactured outliers are then added to the original dataset and used to train a standard $(C + 1)$-class classifier. The multi-phase training, known complications of training generative adversarial networks (GANs), and assumed image-based data limit the generalization of this work to other healthcare applications. Furthermore, the authors in [2] visualize the generated "unknown" class images to verify they are "different yet reasonable," it is not clear how to apply this criterion to patient demographics or abnormal lab tests that comprise our data. Relatedly, [8] proposes framing medical diagnosis classification in terms of open-set recognition. Their method treats samples from less common conditions as a proxy for the unknown classes and instead maximizes their cross-entropy during classic softmax training. During inference, a simple threshold is applied to closed-set, softmax probabilities to reject unknown samples. A shortcoming of this method is the restricting assumption that one's dataset can afford such a large enough and representative residual subset. Indeed in [8], the authors have a known training set of 160 diagnosis classes and a counterpart set of another 160 diagnoses (each with at least 10% of training diagnosis' samples) to model the unknown classes. For our breast cancer patient dataset, there are orders of magnitude differences in the number of samples per cocktail, as well as a drug approval timeline. Both reasons render our residual samples inadequate and possibly time-inconsistent for such a procedure.

In contrast, GMVAE naturally serves non-image data and entirely circumvents the need for artificial "unknown" or "novel" samples. Accordingly, for comprehensive comparison, we benchmark GMVAE against the so-called ii-loss with outlier score from [10], OpenMax from [11], and SVM with local outlier factor (LOF) from [12]. We briefly summarize each benchmark method in the following subsections.

2.1. ii-loss with outlier score [10]

This benchmark is fitting because it (i) attains state-of-the-art open-set recognition accuracies on two non-image-based datasets, and (ii) makes use of similar latent space, distance-based thresholding to reject "novel" samples. It is worth noting that [10] demonstrates that naive thresholding on closed-set, softmax classifiers can lead to significantly poorer open-set recognition. The ii-loss is still wholly classification-based and by contrast GMVAE has the advantage of dual classification-reconstruction learning.

For completeness, we now summarize the ii-loss and outlier score. The authors in [10] argue that open-set recognition is most amenable in a data embedding that clusters samples from the same known class tightly together (low intra-spread) but pushes samples from different known classes far apart from each other (high inter-spread). To directly produce such a neural network mapping z of the data x, they minimize the following loss function:

$$\begin{aligned}\text{ii-loss} = &\left(\frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{|C_i|}||z(x_j)-\mu_i||_2^2\right)\\ &-\left(\min_{i\neq j}||\mu_i-\mu_j||_2^2\right)\end{aligned} \tag{1}$$

where N is the total number of samples, $|C_i|$ is the number of samples in class $i = 1, ..., C$ and

$$\mu_i = \frac{1}{|C_i|}\sum_{j=1}^{|C_i|} z(x_j)$$

is the centroid of each class i. The first term in (1) measures intra-spread and so aims to minimize the distance between each latent z and its own centroid. The second term in (1) quantifies inter-spread and seeks to maximize the minimum distance between class centroids. Batch normalization layers prevent this term from diverging to infinity. The neural network projection $z(x)$ has no set architecture and can be composed of any architectural designs.

With a trained latent representation $z(x)$ in hand, the outlier score, or squared distance to the nearest centroid, is given by

$$\text{outlier score}\left(\hat{x}\right) = \min_i ||\mu_i - z\left(\hat{x}\right)||_2^2$$

for a test sample $\hat{x}$. Consequently, distances to centroids also naturally emit a softmax posterior class probability

$$\mathbb{P}\left(y = i|\hat{x}\right) = \frac{\exp\left\{-||\mu_i - z\left(\hat{x}\right)||_2^2\right\}}{\sum_{j=1}^{C}\exp\left\{-||\mu_j - z\left(\hat{x}\right)||_2^2\right\}}\ .$$

Finally, thresholding on the outlier score, the open-set predicted class $\hat{y}$ is

$$\begin{aligned} c^* &= \arg\max_i \mathbb{P}\left(y = i|\hat{x}\right)\\ \hat{y} &= \begin{cases} c^* & \text{if outlier score}\left(\hat{x}\right) \leq \epsilon_{\text{OS}}\\ C+1 & \text{otherwise.}\end{cases}\end{aligned}$$

We argue that a drawback of this entire procedure is the unsystematic, ad-hoc method of selecting the threshold ϵ_{OS}. It is assumed that some percentage, a so-called contamination ratio α, of the training set are outliers. Correspondingly, the threshold ϵ_{OS} is set to the $1-\alpha$ percentile of all training outlier scores. In experiments, [10] finds that a 1% contamination ratio is broadly suitable. While this is certainly easily understood for the user, it lacks any guidance from the embedding clustering and simply follows from the early presumption. In §4, we illustrate a more deliberate selection for GMVAE's "uncertainty" threshold τ.

2.2. OpenMax [11]

While there are several deep open-set recognition baselines, for instance Classification-Reconstruction learning for Open-Set Recognition (CROSR) [22], Class Conditioned Auto-Encoder (C2AE) [23], and Conditional Gaussian Distribution Learning (CGDL) [25], all of these were specifically designed for images. Embeddings and outputs explicitly depend on convolution and spatial pooling layers, cropping and feature map modulation. Thoughtfully adapting these models to non-image data is outside the scope of this work. Instead we focus on OpenMax; a ubiquitous deep open-set recognition baseline that works by extending softmax activations to alleviate its closed-set nature. Again, this method is still wholly classification-based without incorporating data structure information like GMVAE.

Next, we briefly overview OpenMax. Given a trained classifier's softmax activation vector $v(x) \in \mathbb{R}^C$, a Weibull distribution parametrized by ρ_i is fitted to the tail of distances $||v(x) - \mu_i||$ for correctly classified samples x in class i where μ_i is the respective mean activation vector. The value of WeibullCDF $\left(||v\left(\hat{x}\right) - \mu_i||; \rho_i\right)$ is then a proxy outlier measure for a test sample $\hat{x}$ belonging to class i.

With each class's fitted Weibull distribution, the activation vector function $v\left(\hat{x}\right)$ is modified to include an "unknown" probability $\widetilde{v}_{C+1}\left(\hat{x}\right)$. Letting s correspond to the indices of a reverse sort of the elements of $v\left(\hat{x}\right)$, we have that

$$\begin{aligned}
w_{s_i} &= 1 - \frac{C-i}{C}\text{WeibullCDF}\left(||v\left(\hat{x}\right) - \mu_{s_i}||; \rho_{s_i}\right) \\
\widetilde{v}\left(\hat{x}\right) &= v\left(\hat{x}\right) \odot w \\
\widetilde{v}_{C+1}\left(\hat{x}\right) &= v\left(\hat{x}\right) \cdot (1-w)
\end{aligned}$$

for $i = 1, ..., C$. Essentially OpenMax weights the "unknown" activation by the proxy outlier measures of the known classes and respectively scales those activations down. The factor $\frac{C-i}{C}$ in w is used to place more emphasis on top classes and vice-versa. Finally these modified activations $\widetilde{v}\left(\hat{x}\right)$ are used in softmax

$$\mathbb{P}\left(y = i|\hat{x}\right) = \frac{\exp\left\{\widetilde{v}_i\left(\hat{x}\right)\right\}}{\sum_{j=1}^{C+1}\exp\left\{\widetilde{v}_j\left(\hat{x}\right)\right\}}.$$

Note that $\mathbb{P}\left(y = C+1|\hat{x}\right)$ is included in the above. After a threshold is applied, the open-set predicted class $\hat{y}$ is

$$\begin{aligned}
c^* &= \arg\max_i \mathbb{P}\left(y = i|\hat{x}\right) \\
\hat{y} &= \begin{cases} c^* & \text{if } \mathbb{P}\left(y = c^*|\hat{x}\right) \geq \epsilon_{\text{OM}} \\ C+1 & \text{otherwise.} \end{cases}
\end{aligned}$$

To determine the threshold ϵ_{OM}, the authors in [11] state "we can do a grid search calibration procedure using a set of training images plus a sampling of open set images, optimizing F-measure over the set." As previously discussed, we do not assume access to open-set samples (real or fake) since it may not be feasible. Fortunately [11] finds that $\epsilon_{\text{OM}} = 0.2$ consistently performs best across multiple experiments. We emphasize again that GMVAE's "uncertainty" threshold selection is more broadly applicable and data derived, and not just an experimental byproduct.

2.3. SVM with LOF [12]

Considering our features are structured cancer patient data, we also include a non-neural network baseline. Several open-set recognition methods utilize SVMs as their root model and employ different procedures for rejecting samples as "unknown" [12, 19, 20]. We choose the recent SVM with LOF methodology presented in [12] because it is directly applicable and distinct from the other, centroid-based methods in this paper.

Before summarizing the overall methodology, we first recap LOF as an outlier detection method itself for completeness [26]. For sample a, let k-distance(a) be the distance from a to its k-th nearest neighbor. The reachability distance for two samples a and b is then given by

$$\text{rdist}_k(a, b) = \max\left\{k\text{-distance}(b), ||a-b||\right\}.$$

So reachability distance of a from b is the greater of the their distance and the k-distance of b. In other words, the k nearest neighbors of b are treated as equidistant. [26] justifies this as a smoothing effect. This leads to the local reachability density (lrd) defined as

$$\text{lrd}_k(a) = \left(\frac{\sum_{b \in N_k(a)} \text{rdist}_k(a, b)}{|N_k(a)|}\right)^{-1}$$

where $N_k(a)$ is the set of k nearest neighbors of a. It is the reciprocal average reachability distance of a from its neighbors. LOF then compares local reachability densities with those of its neighbors with

$$\text{LOF}_k(a) = \frac{\sum_{b \in N_k(a)} \text{lrd}_k(b)}{|N_k(a)|\text{lrd}_k(a)}.$$

It is the average local reachability density of the neighbors divided by a's own local reachability density. If $\text{LOF}_k(a) \leq 1$, a has at least the same density as the neighbors so it is not considered an outlier. If $\text{LOF}_k(a) > 1$, a has a lower density as the neighbors so it may be an outlier. The authors in [26] find that a threshold of $\epsilon_{\text{LOF}} = 1.5$ is broadly suitable. This value is not derived and just experimentally works; again, this is a stark contrast to GMVAE's "uncertainty" threshold.

From this, open-set recognition using SVM with LOF in [12] is a straightforward procedure. First a closed-set, multi-class SVM f is trained on the training data X. For a test sample $\hat{x}$, let the winning SVM class be $f\left(\hat{x}\right) = c$. Score $\text{LOF}_k\left(\hat{x}\right)$ is then calculated where $N_k\left(\hat{x}\right) \subseteq X_c$ and X_c is all training samples in class c. The open-set class prediction $\hat{y}$ is then

$$\hat{y} = \begin{cases} c & \text{if } \text{LOF}_k\left(\hat{x}\right) \leq \epsilon \\ C+1 & \text{otherwise.} \end{cases}$$

In this way, LOF utilizes data structure information and is not solely classifier based.

3. Methodology

In this section we outline the complete methodology used for the open-set recognition of breast cancer treatments experiments. First, we describe the GMVAE model; second, each step of the dataset construction is detailed. Finally, we summarize model training and evaluation procedures.

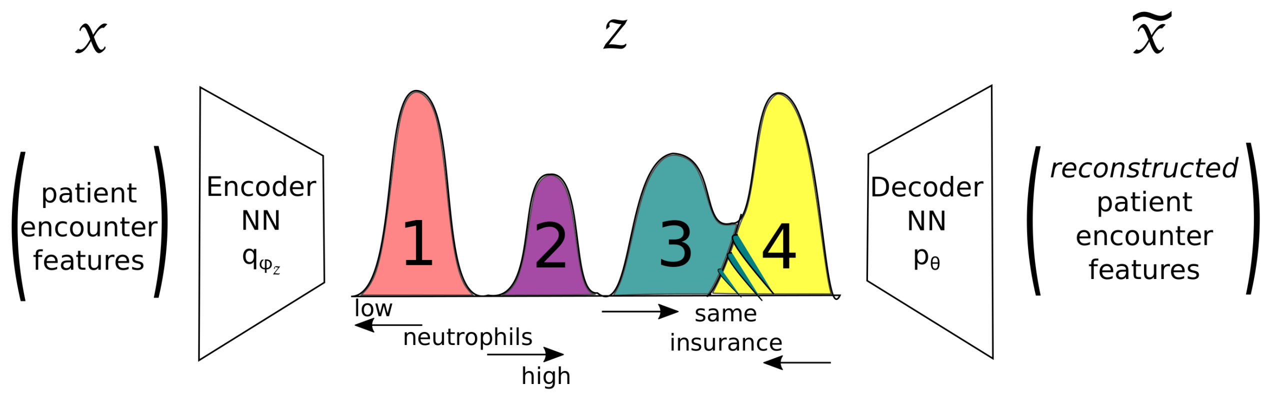

Figure 2: One-dimensional depiction of GMVAE's class-based and reconstruction bottleneck for our breast cancer patients dataset. Neural network q_{ϕ_z} encodes patient encounter features x into learned embedding z. Classes (represented by different numbers and colors) correspond to drug cocktail treatments prescribed for patients' encounters. Samples in each class are clustered together by class but also by (dis)similar features. Latent variable z is then used to reconstruct the original data $\widetilde{x} \sim x$ via network p_θ.

3.1. GMVAE and "uncertainty" algorithm

While the detailed derivation and technical mathematics of GMVAE and the "uncertainty" algorithm can be found in [9], we also briefly overview the model in Appendix A. Additionally, in [9], the authors spend a great time justifying the need for a Gaussian mixture embedding per class for images as well as a procedure for identifying the number of components. However, for our breast cancer dataset, initial assessments indicate that we cannot discern enough patient encounter heterogeneity to warrant multiple components per drug cocktail treatment class. Accordingly, we utilize $K = 1$ (one cluster per class) for all experiments and simplify GMVAE to a single Gaussian prior for each class. This, however, does not reduce GMVAE down to a standard variational autoencoder (VAE) as GMVAE is supervised and clusters based on class too.

The bulk of this section focuses on an intuitive understanding of GMVAE and "uncertainty," as it relates to the open-set recognition of breast cancer cocktail treatments. While GMVAE's structure is more intricate than a standard VAE, its essence can still be understood as the encoder-decoder composition. The principal difference with unsupervised VAEs is that the latent, bottleneck layer cooperatively performs class-based clustering (clinically can be thought of as endophenotyping) and learns reconstruction. We illustrate this duality in Figure 2.

Patient encounter features x are projected to latent space z of significantly fewer dimensions, hence the bottleneck, with neural network q_{ϕ_z}. In z-space, GMVAE's log evidence lower bound's (ELBO's) latent covering term clusters drug cocktail treatment classes together as depicted with class numbers and colors. However, these class clusters are translated and scaled by the reconstruction term (see Appendix A) which promotes patient encounters with similar features to be closer and vice-versa. For example, drug cocktails 1 and 2 may be separated based on neutrophil levels and this characteristic further discriminates these classes. Conversely, a subset of patients of drug cocktails 3 and 4 may share the same insurance, forcing the class clusters to overlap (shown as the yellow-green striped region). While this overlap may be seen as counterproductive, we believe one should not weight or select features to maximize class separation. The reason being that one does not know apriori which features will best separate the "novel" samples. Therefore, an embedding z which most accurately represents data features will naturally separate those distinguishing "novel" samples. Embeddings z are lastly used to reconstruct the samples' features $\widetilde{x}$ via network p_θ. Of course this cooperative, multi-task learning occurs across the entire multi-dimensional z-space. It is important to note here that because of GMVAE's tendency to overlap classes based on reconstruction, the benchmark models exhibits better discrimination among the known classes. However, this closed-set weakness becomes an open-set strength as this behavior shrinks high-risk open-space between the known clusters.

Similar to the outlier score and OpenMax, [9] then applies a distance measure to carve the "novel" decision boundaries around each known centroid. The "uncertainty" threshold quantity is defined as the ratio between the distance to the nearest centroid and the average distance to all other centroids. For our $K = 1$ case with test sample $\hat{x}$, let us denote $\overline{z}_c$ as each known class's training latent centroid and $c^* = \arg\min_c ||\mu(\hat{x}; \phi_z) - \overline{z}_c||_2$. Then "uncertainty" U is mathematically expressed as

$$U = \frac{||\mu(\hat{x}; \phi_z) - \overline{z}_{c^*}||_2}{\frac{1}{C-1}\sum_{c \neq c^*} ||\mu(\hat{x}; \phi_z) - \overline{z}_c||_2}$$

with the corresponding classification rule

$$\hat{y} = \begin{cases} c^* & \text{if } U \leq \tau \\ C+1 & \text{otherwise.} \end{cases}$$

The key differences are that this threshold captures orientation with respect to known centroids (unlike the rotationally symmetric outlier score) and is scale invariant. We visualize these attributes in Figure 3.

For non-trivial open-set recognition, we may assume the "unknown" or "novel" samples are comparable to the known

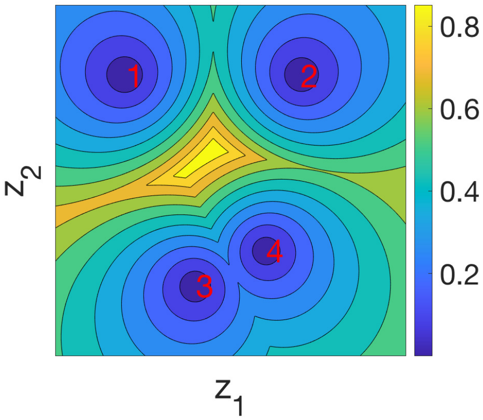

Figure 3: Two-dimensional heat-map visualization of the "uncertainty" U with four known centroids labeled with red numbers.

classes. As such, there exists a large risk of incorrectly classifying a "novel" sample as one of the known classes. Thresholding upon U seeks to minimize this risk by penalizing the open space between known centroids more heavily, as perceived in Figure 3. If $U = 0$ then the test sample's latent embedding is exactly one of the known centroids with no doubt of its classification. However, if $U = 1$ then the test sample's embedding is equidistant to all known centroids and is unclassifiable among the known classes. In addition, U approaches 1 if the test sample's embedding is sufficiently far from all known centroids. Finally, the metric U is designed to be standardized between 0 and 1, making it more universally applicable as opposed to the outlier score's raw distance.

3.2. Dataset construction

The dataset consists of breast cancer patient records at Northwestern Memorial Hospital spanning from 2000 to 2015. We consider patient encounters as independent samples. While this removes the longitudinal aspect of the data, it allows for a more direct application of existing, timeless open-set recognition methods and also matches the cancer treatment narrative because physicians can adjust drug cocktail treatments as patients respond differently or side effects flare.

The samples' classes are specific drug cocktails and are assembled by simply aggregating the prescribed medications for each patient encounter. We only include drugs principally related to treating cancer (as listed by the National Cancer Institute). For the purposes of our breast cancer-related task, drugs like Acetaminophen are extraneous and therefore excluded. In addition, we ultimately only take those cocktails with more than 1,000 encounters to maintain reasonably-sized classes. Table 1 below summarizes the drug cocktail classes of our dataset. Each cocktail's FDA approval year is set to its latest component drug's FDA approval year. Perhaps not so surprising is that most encounter-level cocktails are composed of only a single drug (for convenience, we still refer to these as cocktails). The cocktail data used in the study has to be considered in a hindsight perspective because otherwise a clinical trial would have to be established in order to deal with timely data. For instance, Cocktail 4 of Trastuzumab as a monoclonal antibody is effective only in a subset of breast cancer patients. Years later, we can see Cocktail 6 contains breast cancer chemo-therapy drugs Cyclophosphamide and Doxorubicin, along with Palonosetron which prevents nausea. The 2003 approval of Palonosetron made the already existing Cyclophosphamide and Doxorubicin chemo-therapy more tolerable and feasible in more patients. Open-set recognition experiments with Cocktail 6 as "unknown" (as we have in §4) can broach questions such as: Can we prospectively recognize when patients should receive cocktail treatments containing drugs to combat side effects?

Phenotypic feature engineering for the patient encounters is relatively straightforward with few transformations. We enumerate and categorize demographic and physical characteristic features in Table 2, diagnoses (International Classification of Diseases, Ninth Revision codes or ICD-9 codes) in Table 3, and lab features in Table 4. All demographics except 'age at encounter' are the same for a single patient across encounters. Physical characteristics, having little variance, are averaged across encounters for each patient. In addition, we cutoff ICD-9 codes with less than 1,000 total encounters so that diagnoses are relevant to the dataset as a whole.

After physical characteristic averaging, the lab results are the only remaining features with missing values. The initial missing rates for lab results, resulting from a naive encounter-based merge between cocktails and lab results, are as great as 94.9%. While this is very large, it is consistent with our understanding of patient encounters. Physicians often do not re-order lab tests if there is no reason to expect a change in results. Accordingly, our first-pass data imputation is performing carry forward on all encounters (even those where a cancer drug was not prescribed). After this procedure, nearly all of the missing rates for cocktail-prescribed encounters' lab results fall drastically. We list these missing rates for each lab in Table 4. As the final step, we use Multiple Imputation by Chained Equations (MICE) [27] to impute all outstanding missing lab results. We run five MICE trials and average the results to create the final, holistic dataset.

To conduct the open-set recognition experiments, we must regard a subset of the breast cancer drug cocktails as the "novel" class. Novel drug development signifies an obvious chronology (hence the FDA approval years in Table 1) and so we designate the more recent cocktails as "novel." For the purposes of considering well-balanced "known" and "novel" class splits (in terms of both the number of classes and samples), we conduct two separate experiments. In the first experiment, we designate cocktails 1, 2, 3, and 4 as "known" and cocktails 5, 6, and 7 as "novel" (cocktail numbers in Table 1). The second experiment has cocktails 1, 2, and 3 as

Table 1
Breast cancer drug cocktail classes ordered by FDA approval year.

Cocktail number	Drugs	Number of encounters	FDA approval year
1	Dexamethasone, Ondansetron	2,714	1991
2	Ondansetron	1,868	1991
3	Paclitaxel	3,553	1992
4	Trastuzumab	5,454	1998
5	Pegfilgrastim	1,977	2002
6	Cyclophosphamide, Doxorubicin, Palonosetron	2,416	2003
7	Denosumab	2,616	2010

Table 2
Distributions of demographic and physical characteristic features of breast cancer patient encounters. **Bolded** features are consistent per patient. Categorical demographics are enumerated with indents.

Demographics/physical characteristics	Count (percent)/median (IQR)
Race	
Native American	44 (0.2%)
Asian	1,053 (5.1%)
Black	3,246 (15.8%)
Hispanic	255 (1.2%)
White	13,111 (63.7%)
Unknown	2,889 (14.0%)
Ethnicity	
Hispanic or Latino	1,488 (7.2%)
Not Hispanic or Latino	17,650 (85.7 %)
Unknown	1,460 (7.1%)
Marital status	
Divorced	1,777 (8.6%)
Married	11,699 (56.8%)
Separated	59 (0.3%)
Significant Other	3 (0.01%)
Single	5,645 (27.4%)
Widowed	959 (4.7%)
Unknown	456 (2.2%)
Gender	
Female	20,497 (99.5%)
Male	101 (0.5%)
Insurance	
Private	12,456 (60.5%)
Medicare	3,626 (17.6%)
Medicaid	1,648 (8.0%)
Unknown	2,868 (13.9%)
Age at encounter	53.24 (45.4 - 62.7)
BMI	26.60 (23.3 - 30.9)
Height	64.0 (62.2 - 66.0)
Weight	155.46 (135.9 - 181.0)

"known" and cocktails 4, 5, 6, and 7 as "novel." More details are given in the respective experimental results subsections.

Finally for a model-ready dataset, the training, validation, and testing sets are created as follows. The training set is composed of 2/3 of each "known" cocktail's samples. The validation set is composed of 1/6 of each of the "known" cocktail's samples. These two sets have no novel samples. Finally, the test set is composed of 1/6 of each "known" cocktail's samples and a random subset of size 1/6 of each "novel" cocktail's samples. In this way, the class balances of each split reflects the population. We create 100 test sets by sampling without replacement within each "novel" cocktail. (There are no repeated "novel" samples within a test set, but there are across test sets.) Accordingly, we can present test evaluation minimum-to-maximum intervals and p-values.

Table 3
Diagnosis (ICD-9 code) features of breast cancer patient encounters grouped by disease type.

Disease group	ICD-9 codes
Viral diseases accompanied by exanthem	53.9, 54.9
Malignant neoplasms	153.9, 162.9, 174.8, 174.9, 182.0, 183.0 193.0, 196.3, 196.9, 197.0, 197.7, 198.2, 198.3, 198.4, 198.5, 199.1, 202.8
Benign neoplasms, carcinoma in situ	211.3, 217.0, 218.9, 233.0
Diseases of thyroid and other glands	241.0, 241.1, 244.9, 250.0
Nutritional deficiencies, metabolic and immunity disorders	266.2, 268.9, 272.0, 272.4, 276.8, 278.0
Blood diseases	280.9, 285.22, 285.9
Mental disorders	300.0, 305.1, 311.0
Diseases of nervous system and sense organs	327.23, 338.3, 346.9, 354.0, 355.9, 356.9, 362.5, 365.9, 366.9, 375.15, 389.9
Hypertensive, ischemic heart, pulmonary circulation diseases	401.1, 401.9, 414.0, 415.19
Other heart diseases	424.0, 424.1, 425.4, 427.31, 428.0, 429.9
Diseases of circulatory system	434.91, 443.9, 453.4, 453.9, 455.6, 457.1
Diseases of respiratory system	473.9, 477.9, 486.0, 493.9, 496.0, 511.81, 511.9, 518.89
Diseases of digestive system	530.81, 553.3, 562.1, 562.11, 564.0, 564.1, 571.8, 573.8, 574.2
Diseases of urinary system	592.0, 593.9, 599.0
Disorders of breasts	610.8, 611.72, 611.79, 612.1
Disorders of female genital tract	620.2, 627.2
Contact dermatitis and other eczema	692.9
Arthropathies, dorsopathies, rheumatism	714.0, 715.0, 715.9, 716.9, 719.41, 719.45, 719.46, 723.1, 724.0, 724.02, 724.2, 724.5, 729.1, 729.5
Other disorders of bone and cartilage	733.0, 733.9
Other symptoms	780.4, 780.52, 780.57, 780.79, 781.2, 782.1, 782.3, 782.62, 784.0, 785.1, 785.6, 786.05, 786.09, 786.2, 786.5, 787.01, 787.02, 787.91, 788.3, 788.41, 789.0
Nonspecific (abnormal) findings in blood, radiological examination	790.29, 793.19, 793.8, 793.89
Need for vaccine/inoculation against influenza	V04.81
Personal/family history of malignant neoplasm, other hazards	V10.3, V15.3, V16.3
Other postprocedural states, conditions influencing health	V45.71, V45.89, V49.81
Encounter for other procedure	V58.11, V58.69
Special examinations, screenings	V72.31, V72.83, V72.84, V77.91, V82.81
Genetic susceptibility to disease	V84.01, V84.02
Estrogen receptor positive status	V86.0
Personal history of antineoplastic chemotherapy	V87.41

3.3. Model training and evaluation

For inputs, the numerical features are normalized and the categorical features are one-hot encoded. Below we detail the model parameters for GMVAE, ii-loss, OpenMax, and SVM with LOF.

A table of network architectures for GMVAE is presented in Table 5. The θ network is the mirrored ϕ_z network. Sigmoid activations follow each hidden layer with no dropout nor batch normalizations. The ϕ_z network is pretrained on the known classes and the respective weights are then frozen. We minimize the loss over the training set (using Adam optimizer with learning rate 0.001) until the objective, evaluated on the known validation set, plateaus or begins to increase.

For ii-loss, the z network has the same fully connected layers as GMVAE's ϕ_z network with the same latent space dimension of 10. Each layer is followed by a batch normalization and the two hidden layers also have rectified linear unit (ReLU) activations and dropout with keep probability 0.9. This follow the authors' implementation in [10]. Adam optimizer with learning rate 0.001 is also used here with early stopping on the known validation set.

For OpenMax, the underlying classifier is a neural network that follows GMVAE's ϕ_z network except the last layer's dimension size is C. Again, we utilize the Adam optimizer with learning rate 0.001 and early stopping on the known validation set. We follow an author's own code (`https://github.com/abhijitbendale/OSDN`) in implementing the OpenMax algorithm; namely [11] found that a tail size of 20 for fitting the Weibull distribution preforms best.

For SVM with LOF, Python package `scikit-learn` is used for both algorithms. For validating the SVM classifier over the known validation set, the radial basis function

Table 4
Lab features of breast cancer patient encounters and corresponding carry forward missing rates.

Lab feature	Missing rate
Albumin in blood	11.0%
Alkaline phosphatase in blood	12.9%
Alanine aminotransferase in blood	12.6%
Aspartate aminotransferase in blood	12.6%
Basophils in blood	13.3%
Blood urea nitrogen	10.1%
Calcium level	12.4%
Carbon dioxide level	12.7%
Chloride level	12.4%
Creatinine level in blood	12.4%
Eosinophils in blood	12.0%
Estrogen receptor	0%
Glucose level	9.2%
Hematocrit level	9.8%
Hemoglobin level	9.8%
Human epidermal growth factor receptor 2 protein	0%
Ki-67 protein	0%
Lymphocytes in blood	13.2%
Mean corpuscular hemoglobin concentration	9.7%
Mean corpuscular volume	9.7%
Monocytes in blood	13.3%
Mean platelet volume	13.1%
Neutrophils in blood	13.3%
Pathological node of TNM stage	73.3%
Pathological tumor of TNM stage	92.9%
Platelets in blood	9.9%
P53 gene mutation	0%
Potassium level	12.4%
Progesterone receptor	0%
Red blood cell count	9.8%
Red cell distribution width	9.7%
Sodium level	12.4 %
Tumor grade	61.2%
Tumor size	60.2%
Bilirubin in blood	12.9%
Protein level	12.9%
White blood cell count	9.8%

Table 5
Network architectures for GMVAE.

ϕ_z	ϕ_w	β
Input: x	Input: x, y	Input: w
FC-100	Concatenate with y	FC-20
FC-50	FC-20 $(2 \times \dim(w))$	FC-20
FC-20 $(2 \times \dim(z))$		FC-80 $(2 \times C \times \dim(z))$

(RBF) kernel and regularization parameter of 3 for the first experiment and 2.5 for the second experiment is used. All other parameters are left at their default values. For LOF, all parameters are left at default values, i.e. 20 nearest neighbors and euclidean distance.

Unfortunately, Northwestern's cancer patient data cannot be made available as it contains individuals' protected health information. To our knowledge, there is no public cancer patient dataset with comprehensive treatment information that would allow for similar open-set experiments.

4. Results

From the experimental results given next, we clearly demonstrate that GMVAE outperforms the three benchmark open-set classifiers described in §2, both in terms of accuracy and robustness to an increasing number of novel

Table 6
Breast cancer drug cocktails split into four "known" classes and one "novel" class according to FDA approval year.

"Knowns"/"novel" split	Cocktail numbers	Number of encounters	FDA approval year
"Knowns"	1	2,714	1991
	2	1,868	1991
	3	3,553	1992
	4	5,454	1998
"Novel"	5	1,977	2002
	6	2,416	2003
	7	2,616	2010

cocktails (and samples). We attribute this to two primary reasons. First, the GMVAE model also considers reconstruction, which captures additional data structure information, as well as classifier information. Second, GMVAE is more deliberate in algorithmically selecting an "uncertainty" threshold τ based on the known validation set. Indeed, threshold selection for outlier score, OpenMax, and LOF is not optimal, nor even well-defined, without unknown samples in validation. For GMVAE's threshold selection as well as model comparison, we utilize the macro-averaged F1 score as our accuracy measure to account for class imbalance.

4.1. Four existing and three "novel" cocktails

For this first experiment, we divide the cocktails according to Table 6. While "known" and "novel" splits have a similar number of cocktails, here we are considering a scenario in which there are more "known" samples. There are approximately twice as many sample in the "known" cocktails as "novel" cocktails.

To contextualize this particular split, we can imagine we are in the year 2000. Can we identify if a patient encounter should receive one of the four "known" cocktails (and which one) or should a "novel" cocktail be prescribed? Again, the "novel" class (composed of three cocktails) is an indication these encounters are opportune for an original cocktail treatment. In reality, we must acknowledge that our study's experimental design is only a proxy to this scenario. We did not enforce samples to this timeline and so there may be "known" cocktail patient encounters after say 2003, who actually could have been prescribed one of the "novel" cocktails. Such is an inherent limitation of retrospective data used for simulation. Given our already small dataset, we lack the samples prior to 2000 to implement this true timeline. However, our study still captures the relevant, timeless task of classifying which patient encounters should be prescribed an existing versus novel cocktail.

As per the authors of [10, 11, 12, 26], the outlier score's threshold ϵ_{OS} corresponds with a $\alpha = 1\%$ contamination rate, OpenMax's threshold $\epsilon_{\text{OM}} = 0.2$, and LOF's threshold $\epsilon_{\text{LOF}} = 1.5$. We now detail our procedure for selecting the "uncertainty" threshold. Plotted in Figure 4 are the known validation F1 scores versus τ for GMVAE's U quantity. Work in [9] deduces (and empirically observes) that a consistently good threshold τ to pick for GMVAE's "uncertainty" is the saturation or plateau point of the known validation F1 curve. This is plotted with the red dashed line in Figure 4. Intuitively, this can be thought of as increasing the decision boundary around each class's centroid until diminishing classification accuracy returns. Further increasing τ is tantamount to overfitting the known validation samples and risks under-recognizing "novel" samples. Mathematically we define this saturation point in the following way. Letting

$$\widetilde{\tau} = \min\left\{\tau \,:\, \mathrm{F1}'(\tau) \geq \delta_1\right\} ,$$

the selected threshold is then set to the saturation point

$$\tau^* = \min\left\{\tau \,:\, \tau > \widetilde{\tau} \quad \text{and} \quad \mathrm{F1}'(\tau) \leq \delta_2\right\} .$$

For this experiment, we use $\delta_1 = 1$ and $\delta_2 = 0.25$ and approximate $\mathrm{F1}'(\tau)$ by using the simple forward difference scheme.

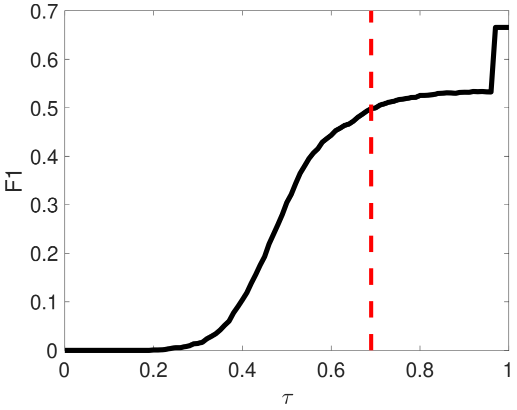

Figure 4: Known cocktails validation F1 scores versus GMVAE's "uncertainty" threshold τ and picked threshold $\tau^* = 0.69$ (dashed red line).

With the selected threshold τ^* for "uncertainty," we proceed to the testing phase with "novel" cocktails. To study robustness to increasing "novel" samples, as well as accuracy, we incrementally increase the number of "novel" cocktails (according to the order in Table 6) and measure F1 scores. These test F1 scores versus the number of "novel" cocktails are plotted in Figure 5. As previously discussed, GMVAE is not as accurate in the closed-set regime with no "novel" samples because ii-loss, OpenMax, and SVM

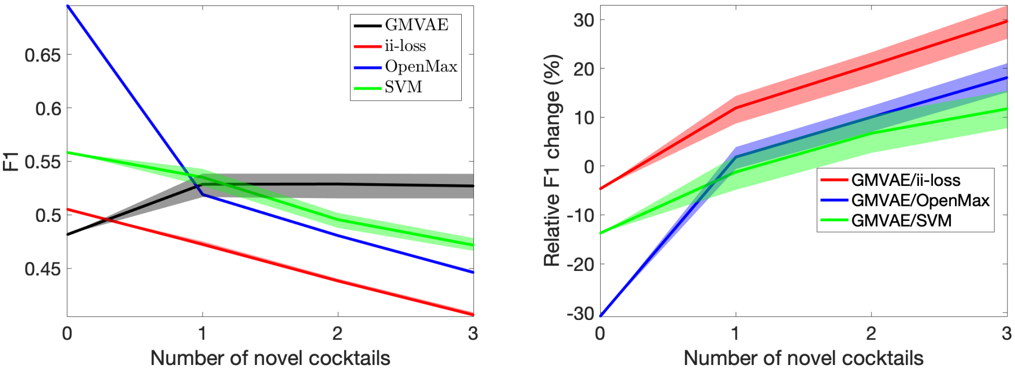

Figure 5: (Left) Open-set test F1 score intervals versus number of "novel" cocktails. Shaded bands show maximums and minimums for the respective points. Note that the y-axis scale is zoomed for detail. (Right) GMVAE's relative changes in F1 scores compared to benchmarks.

more directly optimize known-class discrimination. However, GVMAE and "uncertainty" quickly outperform all benchmarks with the introduction of "novel" cocktails. In addition, we clearly see that GMVAE's method remains more robust to an increasing number of "novel" cocktails and samples while the benchmarks' accuracies continuously diminish. This observation is magnified in the right panel of Figure 5. Averaging over the number of "novel" cocktails and all benchmarks, GMVAE leads to a 5.0% increase in the F1 score. With at least one "novel" cocktail, this increases to 12.2%. Again, we attribute this increased accuracy and robustness to GMVAE's reconstruction learning and "uncertainty." From [9], we surmise the latter's effect is substantial for more homogeneous, difficult-to-discriminate samples. This is certainly true of healthcare data. Note, Figure 5 is in table format in Appendix B.

While F1 scores paint a broad picture of classification accuracy, the confusion matrices in Table 7 more closely inspect classification ability on an individual cocktail basis. We clearly see that the benchmarks ii-loss, OpenMax, and SVM are more accurate in the closed-set regime. In particular, cocktails 3 and 4 are classified very accurately. However, this comes at the expense of severely under-recognizing "novel" cocktails. The ii-loss and OpenMax models rarely classify "novel" leading to a dramatic decrease in the overall open-set classification accuracy. No doubt this is due, in part, to them solely optimizing known cocktail discrimination with no regard for capturing underlying feature information. SVM does capture some via LOF which is the reason it performs better in the open-set regime. Conversely, GMVAE is less accurate in the closed-set regime but more readily recognizes "novel" cocktails. The main point these confusion matrices demonstrate is that there is a tradeoff between accurately classifying the known classes and robustly identifying novel or unknown classes. This is evidenced by GMVAE's over-classifying the known cocktails as "novel."

Finally, we wish to further address the threshold selection. To alleviate concerns that open-set accuracies are more sensitive to threshold selection and that we've not unknowingly picked advantageous thresholds, we plot the test F1 scores versus thresholds in neighborhoods of GMVAE's τ^*, ii-loss's α, OpenMax's ϵ_{OM}, and SVM's ϵ_{LOF} in Figure 6. We consider a "10 percent" neighborhood so that $\alpha \in [0, 0.1]$, $\tau \in [0.64, 0.74]$, $\epsilon_{\mathrm{OM}} \in [0.15, 0.25]$, and $\epsilon_{\mathrm{LOF}} \in [1.42, 1.58]$. We clearly see that GMVAE's F1 scores are greater and more robust with respect to τ than those of our benchmarks. This translates to GMVAE being more robust to threshold selection (error) and, relatedly, its underlying embedding having a stronger ability to distinguish "novel" cocktails.

4.2. Three existing and four "novel" cocktails

For this second experiment, we divide the cocktails according to Table 8. While "known" and "novel" splits again have a similar number of cocktails as in the previous experiment, here we increase the number of "novel" cocktails by one and consider the different scenario in which there are more "novel" samples. There are approximately 1.5 times as many samples in the "novel" cocktails as "known" cocktails.

The goal of this second experiment is to reiterate GMVAE's success while varying the number of "known" cocktails and having a qualitatively different "known"-to-"novel" samples ratio. Parallel Figures 7, 8, and 9 and Table 9 from the first experiment show just this. Figure 7 again plots the known validation F1 scores versus τ for GMVAE's U quantity with the saturation point and corresponding picked threshold τ^* plotted in red.

Again, we incrementally increase the number of "novel" cocktails (according to the order in Table 8) and plot the F1 scores in Figure 8. The behavior is qualitatively the same as the first experiment. The benchmark methods are more accurate with just the "known" cocktails as it directly optimizes class separation. However, GMVAE yields much higher open-set classification accuracies for increasing "novel" cocktails and samples. In fact, averaging over the number of "novel" cocktails in Figure 8's right panel, GMVAE leads to an average F1 increase of 22.9%. Again, this is formatted in a table in Appendix B. We stress that

Table 7
Single test set confusion matrices with true cocktail class as rows and predicted cocktail as columns for (top left) GMVAE, (top right) ii-loss, (bottom left) OpenMax, and (bottom right) SVM. Double horizontal line divides known and novel cocktails.

GMVAE

Cocktails	1	2	3	4	Novel
1	163	117	33	20	120
2	101	110	21	10	70
3	0	0	393	60	140
4	3	2	49	704	151
5	3	2	98	78	148
6	7	2	131	86	176
7	3	2	97	98	236

ii-loss

Cocktails	1	2	3	4	Novel
1	213	117	54	64	5
2	114	115	37	31	15
3	16	14	468	95	0
4	21	0	92	796	0
5	40	15	156	118	0
6	45	8	221	128	0
7	43	41	212	140	0

OpenMax

Cocktails	1	2	3	4	Novel
1	294	85	47	27	0
2	153	105	32	22	0
3	16	15	507	55	0
4	20	7	29	853	0
5	31	20	174	104	0
6	46	18	239	99	0
7	31	45	182	178	0

SVM

Cocktails	1	2	3	4	Novel
1	284	78	41	33	17
2	154	97	29	20	12
3	9	5	519	47	13
4	10	5	15	842	37
5	31	15	167	100	16
6	45	9	242	95	11
7	47	20	185	138	46

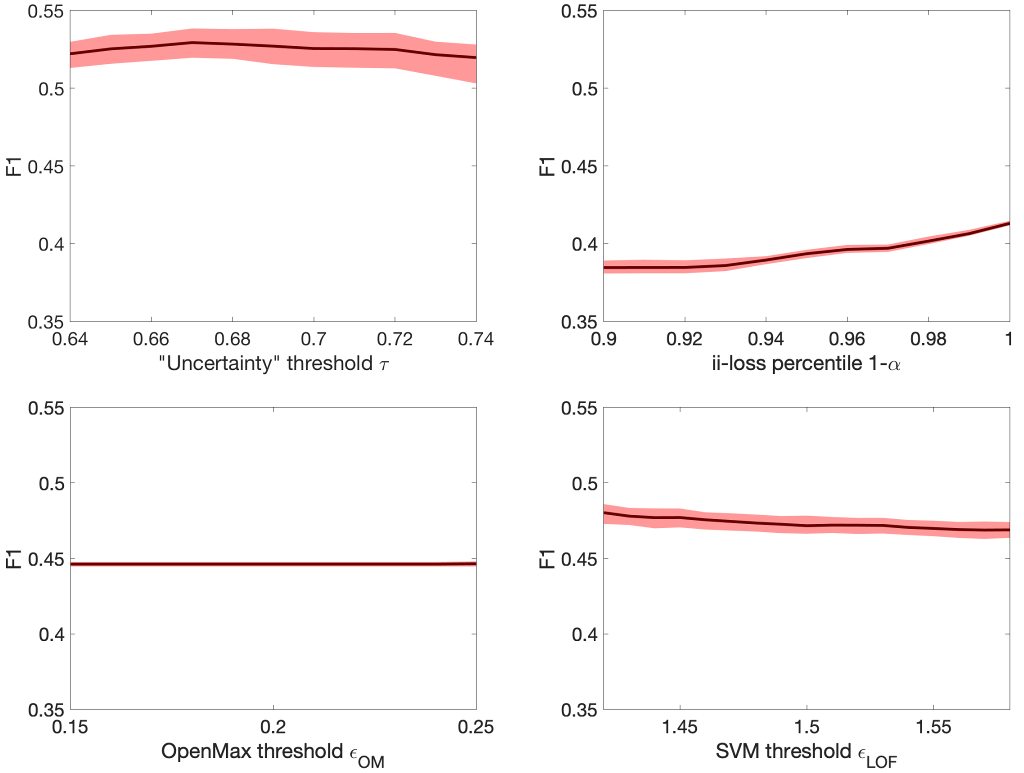

Figure 6: Open-set test F1 score intervals, calculated using all "novel" cocktails, versus nearby neighborhood of (top left) GMVAE's τ^*, (top right) ii-loss's α, (bottom left) OpenMax's ϵ_{OM}, and (bottom right) SVM's ϵ_{LOF}. Red shaded bands show maximums and minimums for the respective points. Note that the y-axis scale is zoomed for detail.

this increased open-set recognition is from GMVAE's reconstruction learning and "uncertainty" leading to better discernment of "novel" cocktails. This is made clear in the confusion matrices in Table 9. The benchmarks' embeddings only captures class information and therefore it is difficult to tease "novel" cocktail information from them.

To again show robustness to threshold selection, we plot the test F1 scores versus thresholds in neighborhoods in Figure 9. We consider a "10 percent" neighborhood so that

Table 8
Breast cancer drug cocktails split into three "known" classes and one "novel" class according to FDA approval year.

"Knowns"/"novel" split	Cocktail numbers	Number of encounters	FDA approval year
"Knowns"	1	2,714	1991
	2	1,868	1991
	3	3,553	1992
"Novel"	4	5,454	1998
	5	1,977	2002
	6	2,416	2003
	7	2,616	2010

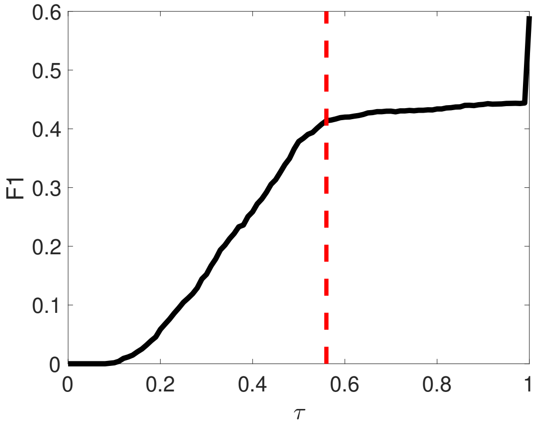

Figure 7: Second experiment's known cocktails validation F1 scores versus GMVAE's "uncertainty" threshold τ and picked threshold $\tau^* = 0.56$ (dashed red line).

$\alpha \in [0, 0.1]$, $\tau \in [0.51, 0.61]$, $\epsilon_{\text{OM}} \in [0.15, 0.25]$, and $\epsilon_{\text{LOF}} \in [1.42, 1.58]$. Here we clearly see that all F1 scores are relatively constant with respect to thresholds However, the difference in F1 is stark with GMVAE dominating.

5. Limitations and future work

While the experimental results highlight GMVAE's capability, we do wish to stress again the importance of methodically selecting a single threshold for rejecting "novel" samples in testing. Previous open-set experiments [2, 8, 10] escape this consideration by comparing high-level metrics like AUC. While this may give an indication of the overall behavior, it does little to inform actual model usage in a practical setting. GMVAE's validation F1 curve saturation procedure begins to address this decision boundary optimization without "novel" samples. However, it is still an ad-hoc heuristic worthy of further development. A satisfying

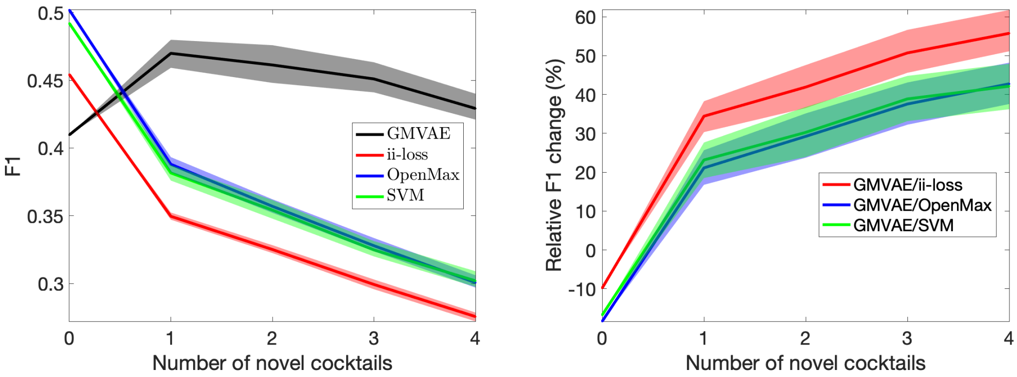

Figure 8: (Left) Second experiment's open-set test F1 score intervals versus number of "novel" cocktails. Shaded bands show maximums and minimums for the respective points. Note that the y-axis scale is zoomed for detail. (Right) GMVAE's relative changes in F1 scores compared to benchmarks.

Table 9
Second experiment's single test set confusion matrices with true cocktail class as rows and predicted cocktail as columns for (top left) GMVAE, (top right) ii-loss, (bottom left) OpenMax, and (bottom right) SVM. Double horizontal line divides known and novel cocktails.

GMVAE

Cocktails	1	2	3	Novel
1	231	55	48	119
2	129	57	42	84
3	2	1	459	131
4	37	24	446	402
5	8	7	172	142
6	3	1	191	207
7	9	3	302	122

ii-loss

Cocktails	1	2	3	Novel
1	288	125	33	7
2	133	122	52	5
3	17	107	456	13
4	187	427	292	3
5	35	133	153	8
6	41	156	202	3
7	79	185	167	5

OpenMax

Cocktails	1	2	3	Novel
1	306	100	46	1
2	147	122	43	0
3	18	14	558	3
4	209	141	540	19
5	52	42	233	2
6	54	32	313	3
7	89	64	279	4

SVM

Cocktails	1	2	3	Novel
1	314	79	54	6
2	160	99	38	15
3	18	3	563	9
4	281	101	507	20
5	64	30	224	11
6	66	22	309	5
7	104	53	261	18

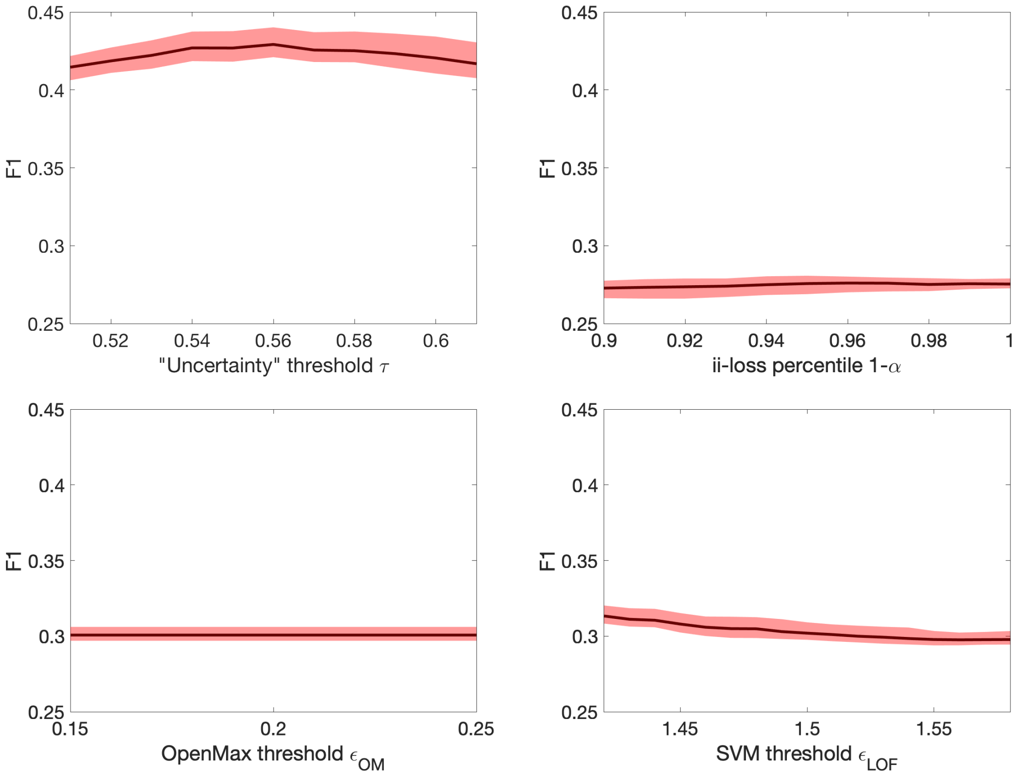

Figure 9: Second experiment's open-set test F1 score intervals, calculated using all "novel" cocktails, versus nearby neighborhood of (top left) GMVAE's τ^*, (top right) ii-loss's α, (bottom left) OpenMax's ϵ_{OM}, and (bottom right) SVM's ϵ_{LOF}. Red shaded bands show maximums and minimums for the respective points. Note that the y-axis scale is zoomed for detail.

solution to this subproblem is critical for real applications to a current patient's treatment plan.

Additional discussions of our current application of open-set recognition to drug treatment classifications are more subtle. From the experiments above, while we achieve more accurate results, an F1 score below 0.5 has room for improvement. Prior work and our own experience suggest there exists a tradeoff between closed-set classification and open-set recognition. In other words, it is natural to expect a compromise between accurately classifying the "known" classes and robustly identifying “novel” or “unknown” classes. Herein lies this issue, as it is generally more difficult to discriminate and distinguish real healthcare data (as opposed to academic image datasets used in most open-set studies), we start off at a severe disadvantage with separating the known cocktails. We visualize this with a t-Distributed Stochastic Neighbor Embedding (t-SNE) [28] plot of GMVAE’s latent embedding from the first experiment in Figure 10. It indicates there is high degree of feature “overlap” among the cocktails and thus it is difficult to distinguish patient encounters. This is likely expected from only utilizing the phenotypic features available in the dataset, but perhaps also from the nature of patients potentially benefiting from multiple cocktails. In that respect, open-set recognition within a multi-label setting may be a natural extension we may pursue in future work.

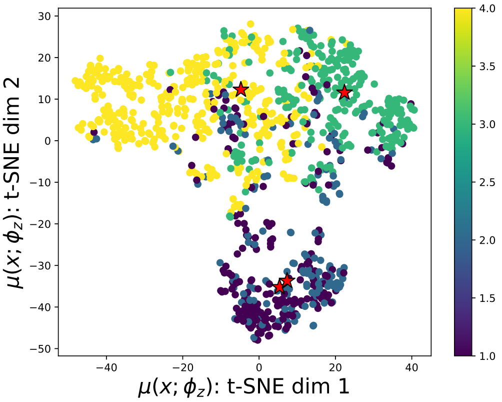

Figure 10: t-SNE plot of $\mu(x;\phi_z)$ from the first experiment's GMVAE training latent representations for the known cocktails. Colors distinguish cocktails and red stars are the class centroids.

Lastly, perhaps drug treatment classifications, at least in our cancer context, necessitate a longitudinal study. One would like to track multiple encounters for a given patient and be able to make breast cancer drug treatment recommendations along the way. With this framework, the efficacy feedback information from past treatments could potentially provide significant guidance. While this would require a complete overhaul from data collection to modifying the model to accept a variable-length time series, it certainly represents an obvious and exciting expansion. Relatedly, this immediately suggests the addition of multi-modal data such as mammogram images and physician notes’ text in a recurrent neural network model structure.

6. Conclusion

We formulate breast cancer drug cocktail treatment classsifications in terms of open-set recognition, focusing on a methodology conducive to a practical clinical implementation, and accordingly apply the GMVAE and “uncertainty” model framework. Together, these achieve more accurate and robust classification results for our patient-encounter healthcare dataset, compared to multiple benchmarks. In doing so, we also resolve obstacles in prior work concerning open-set recognition applications to healthcare. First, we emphasize formally addressing a methodical selection of a specific threshold for rejecting “novel” or “unknown” samples as we believe it is more meaningful in deployment to compare and use models with a single testing instance. Second, many other works on this subject take advantage of fabricated or auxiliary data to model “novel” or “unknown” samples. We dismiss the implicit assumption that this step is always feasible and instead call for methods like GMVAE which only learn from known, available data. Finally, we spotlight the inherent limitations to solely classification-based models in open-set recognition. Whether it is reconstruction or not, embeddings must encapsulate structural information of the data to be more effective. To be sure, this particular application to healthcare opens interesting avenues of further research to the expanding scope of open-set recognition. Likewise, this study hopefully represents a stride towards these techniques benefiting actual patients’ treatments in the future.

Acknowledgments

The work of the first author is supported by the Predoctoral Training Program in Biomedical Data Driven Discovery (BD3) at Northwestern University (National Library of Medicine Grant 5T32LM012203). The work of the last author is supported in part by NIH Grant R01LM013337.

Conflicts of interest statement

The authors declare there are no conflicts of interest.

A. Background for GMVAE and “uncertainty” algorithm

Here in this Appendix section, we briefly overview GMVAE, which extends standard unsupervised VAEs by assuming a Gaussian mixture prior for each class. To accommodate this, the basic VAE architecture is modified with additional latent variables. [9] starts with C known classes with each class composed of K_c mixture components where $c = 1, 2, ..., C$. The features $x \in \mathbb{R}^d$ and labels $y \in \mathbb{R}^C$, represented as one-hot vectors, comprise the labeled, known dataset. GMVAE’s decoder model $p_{\beta,\theta}(x, v, w, z|y) = p_\theta(x|z)p_\beta(z|w, y, v)p(w)p(v|y)$ conditions on class and factors as

$$
\begin{aligned}
&w \sim \mathcal{N}(0, I)\\
&(v|y) \in \mathbb{R}^{K_c} \sim \mathrm{Mult}(\pi(y))\\
&(z|w, y, v)\\
&\quad \sim \prod_{c=1}^{C}\prod_{k=1}^{K_c} \mathcal{N}\left(\mu_{ck}(w;\beta), \mathrm{diag}\left(\sigma^2_{ck}(w;\beta)\right)\right)^{y_c \cdot v_k}\\
&(x|z) \sim \mathcal{B}(\mu(z;\theta))
\end{aligned}
$$

Table 10
Mean F1 scores corresponding with Figure 5. All benchmark means are highly significant with p-values less than 10^{-24} from a two-sided t-test against corresponding GMVAE means.

Number of novel cocktails	GMVAE	ii-loss	OpenMax	SVM
0	0.482	0.505	0.696	0.558
1	0.529	0.472	0.519	0.535
2	0.529	0.438	0.481	0.496
3	0.527	0.407	0.446	0.472

where $\mu_{ck}(\cdot;\beta)$, $\sigma^2_{ck}(\cdot;\beta)$, and $\mu(\cdot;\theta)$ are neural networks parametrized by β and θ, respectively. It is common to assume a uniform prior $\pi(y)$. The encoder process is factorized as $q_\phi(v, w, z|x, y) = p_\beta(v|z, w, y) q_{\phi_w}(w|x, y) q_{\phi_z}(z|x)$ where $\phi = (\phi_x, \phi_w)$. Factors ϕ are parametrized with networks that output mean and diagonal covariance for Gaussian posteriors:

$$\begin{aligned}(z|x) &\sim \mathcal{N}\left(\mu(x;\phi_z), \operatorname{diag}\left(\sigma^2(x;\phi_z)\right)\right)\\ (w|x,y) &\sim \mathcal{N}\left(\mu(x,y;\phi_w), \operatorname{diag}\left(\sigma^2(x,y;\phi_w)\right)\right).\end{aligned}$$

There is a p_β factor in the q_ϕ factorization because it is derived from the generative factors (see [9]). GMVAE's objective is to maximize the log-evidence lower bound (ELBO) given by

$$\begin{aligned}\mathcal{L}(K) &= \mathbb{E}_{q_\phi(v,w,z|x,y)}\left[\log \frac{p_{\beta,\theta}(x,v,w,z|y)}{q_\phi(v,w,z|x,y)}\right]\\ &= \mathbb{E}_{q_{\phi_z}(z|x)}\left[\log p_\theta(x|z)\right] \quad \text{(reconstruction)}\\ &- \mathbb{E}_{q_{\phi_w}(w|x,y)q_{\phi_z}(z|x)}\Big[\log q_{\phi_z}(z|x)\\ &\qquad - \sum_{j=1}^{K_c} p_\beta(v=j|z,w,y)\log p_\beta(z|w,y,v=j)\Big]\\ &\qquad\qquad\qquad \text{(latent covering)}\\ &- KL(q_{\phi_w}(w|x,y)||p(w)) \quad (w\text{-prior})\\ &- \mathbb{E}_{q_{\phi_w}(w|x,y)q_{\phi_z}(z|x)}\left[KL(p_\beta(v|z,w,y)||p(v|y))\right]\\ &\qquad\qquad\qquad \text{(component } v\text{-prior).}\end{aligned}$$

Vector $K = (K_1, K_2, ..., K_C)$ is decided by the user and so the ELBO dependence on K is made explicit. The reconstruction term endeavors to group samples with similar features together in latent space z. Simultaneously, the latent covering term attempts to cluster the latent representations z based on classes. The w-prior and component v-prior terms aim for the respective posteriors and priors to coincide. This mirrors the standard ELBO with reconstruction and regularization terms.

B. Additional results tables

Tables 10 and 11 explicitly print mean F1 scores plotted in Figures 5 and 8.

Table 11
Mean F1 scores corresponding with Figure 8. All benchmark means are highly significant with p-values less than 10^{-152} from a two-sided t-test against corresponding GMVAE means.

Number of novel cocktails	GMVAE	ii-loss	OpenMax	SVM
0	0.410	0.454	0.502	0.492
1	0.470	0.350	0.388	0.382
2	0.461	0.325	0.357	0.354
3	0.451	0.299	0.328	0.325
4	0.429	0.276	0.301	0.302

References

[1] C. Geng, S.-j. Huang, S. Chen, Recent advances in open set recognition: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[2] Y. Xiao, S. Gao, Z. Chai, K. Zhou, T. Zhang, Y. Zhao, J. Cheng, J. Liu, Open-set oct image recognition with synthetic learning, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), IEEE, 2020, pp. 1788–1792.

[3] K. F. Chung, Defining phenotypes in asthma: a step towards personalized medicine, Drugs 74 (2014) 719–728.

[4] S. R. N. Kalhori, M. Tanhapour, M. Gholamzadeh, Enhanced childhood diseases treatment using computational models: Systematic review of intelligent experiments heading to precision medicine, Journal of Biomedical Informatics 115 (2021) 103687.

[5] J. Boshuizen, D. S. Peeper, Rational cancer treatment combinations: An urgent clinical need, Molecular Cell 78 (2020) 1002 – 1018.

[6] J. Yang, Z. Xu, W. K. K. Wu, Q. Chu, Q. Zhang, Graphsynergy: a network-inspired deep learning model for anticancer drug combination prediction, Journal of the American Medical Informatics Association 28 (2021) 2336–2345.

[7] X. Li, Y. Xu, H. Cui, T. Huang, D. Wang, B. Lian, W. Li, G. Qin, L. Chen, L. Xie, Prediction of synergistic anti-cancer drug combinations based on drug target network and drug induced gene expression profiles, Artificial intelligence in medicine 83 (2017) 35–43.

[8] V. Prabhu, A. Kannan, G. J. Tso, N. Katariya, M. Chablani, D. Sontag, X. Amatriain, Open set medical diagnosis, arXiv preprint arXiv:1910.02830 (2019).

[9] A. Cao, Y. Luo, D. Klabjan, Open-set recognition with Gaussian mixture variational autoencoders, Proceedings of the AAAI Conference on Artificial Intelligence (2021).

[10] M. Hassen, P. K. Chan, Learning a neural-network-based representation for open set recognition, in: Proceedings of the 2020 SIAM International Conference on Data Mining, SIAM, 2020, pp. 154–162.

[11] A. Bendale, T. E. Boult, Towards open set deep networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1563–1572.

[12] T. Walkowiak, S. Datko, H. Maciejewski, Algorithm based on modified angle-based outlier factor for open-set classification of text documents, Applied Stochastic Models in Business and Industry 34 (2018) 718–729.

[13] C. Zhou, R. C. Paffenroth, Anomaly detection with robust deep autoencoders, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 665–674.

[14] D. Hendrycks, M. Mazeika, T. Dietterich, Deep anomaly detection with outlier exposure, in: International Conference on Learning Representations, 2019.

[15] L. Gao, L. Zhang, C. Liu, S. Wu, Handling imbalanced medical image data: A deep-learning-based one-class classification approach, Artificial intelligence in medicine 108 (2020) 101935.

[16] S. Hela, B. Amel, R. Badran, Early anomaly detection in smart home: A causal association rule-based approach, Artificial intelligence in

medicine 91 (2018) 57–71.

[17] J. Thongkam, G. Xu, Y. Zhang, F. Huang, Support vector machine for outlier detection in breast cancer survivability prediction, in: Y. Ishikawa, J. He, G. Xu, Y. Shi, G. Huang, C. Pang, Q. Zhang, G. Wang (Eds.), Advanced Web and Network Technologies, and Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 99–109.

[18] A. J. Boddy, W. Hurst, M. Mackay, A. e. Rhalibi, Density-based outlier detection for safeguarding electronic patient record systems, IEEE Access 7 (2019) 40285–40294.

[19] W. J. Scheirer, A. Rocha, A. Sapkota, T. E. Boult, Towards open set recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (2013).

[20] L. P. Jain, W. J. Scheirer, T. E. Boult, Multi-class open set recognition using probability of inclusion, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 393–409.

[21] H. Zhang, V. M. Patel, Sparse representation-based open set recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017) 1690–1696.

[22] R. Yoshihashi, W. Shao, R. Kawakami, S. You, M. Iida, T. Naemura, Classification-reconstruction learning for open-set recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[23] P. Oza, V. M. Patel, C2AE: Class conditioned auto-encoder for open-set recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2307–2316.

[24] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5967–5976. doi:10.1109/CVPR.2017.632.

[25] X. Sun, Z. Yang, C. Zhang, K.-V. Ling, G. Peng, Conditional gaussian distribution learning for open set recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13480–13489.

[26] M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander, Lof: identifying density-based local outliers, in: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 2000, pp. 93–104.

[27] S. v. Buuren, K. Groothuis-Oudshoorn, MICE: Multivariate imputation by chained equations in R, Journal of Statistical Software (2010) 1–68.

[28] L. v. d. Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579–2605.