Combining BioClinicalBERT and ResNet-18 for Multimodal Anomaly Detection in Radiology: An Unsupervised IF–LOF Approach

Satish Wagle[1], Saroj Baral[2], Jorge Vargas [3], and Khem Poudel[2]

[1]Computational and Data Science, Middle Tennessee State University, Murfreesboro, TN, USA

[2]Department of Computer Science, Middle Tennessee State University, Murfreesboro, TN, USA

[3]Department of Engineering Technology, Middle Tennessee State University, Murfreesboro, TN, USA

sw8k@mtmail.mtsu.edu, sb2ek@mtmail.mtsu.edu, jorge.vargas@mtsu.edu, khem.poudel@mtsu.edu

Abstract. Timely detection of anomalies is important in clinical data for identifying rare conditions, diagnostic errors, and healthcare efficiencies. This study proposed an unsupervised hybrid anomaly detection approach to combine clinical reports and chest X-rays using BioClinicalBERT and ResNet-18 embeddings. It integrates outlier filtering with Isolation Forest, applies a pruning step to retain the most anomalous candidates, and then performs local refinement using Local Outlier Factor. The pipeline was evaluated on 3,851 samples from the Open-I dataset. Model performance was explored with various outlier detection seeds, different pruned percentiles, and different LOF settings. The results indicated consistent and strong anomaly detection of outliers indices (121, 1632, 2265, 2673) that were identified as anomalies for most of the different configuration setting. The Jaccard similarity heatmap showed moderate to high overlap between results (scores between 0.4 and 0.8), and the t-SNE plots visually confirmed that different types of samples were well separated in space. IF-LOF modeling approach demonstrated evidence in capturing the multimodal physical patterns inherent in clinical data; images and medical text representations yet required no labelled data, and leveraged the benefits of global partitioning and local density-based refinement, complemented by globally semantic embeddings from BioClinicalBERT for reliable and valid anomaly detection in clinical practice.

Keywords: Multimodal Anomaly Detection, Open-I Dataset, BioClinicalBERT, ResNet-18, Unsupervised Learning, Isolation Forest, Local Outlier Factor, Pruning, t-SNE, Jaccard Similarity

1 Introduction

Anomalies are referred as exceptions, outliers or the variation from the expected pattern of the system's actual behavior. Anomaly detection in health care is a vital analytical process that assists in identifying infrequent but clinically significant divergence from normal clinical observations [8]. Effective and timely anomaly identification considerably contributes toward improved patient outcomes by enabling prompt medical treatment and avoiding diagnostic error. Anomaly detection techniques further improve the allocation of healthcare resources by pointing towards inefficiencies and scope for improvement in clinical work flows [1][19].

Traditional medical anomaly detection mainly uses unimodal analyses that considers clinical text and medical images separately. Multimodal analysis is important for a more

complete detection process because the anomalies in healthcare most commonly appear as complex multimodal relationships [19]. The combination of different modalities of the data provides more concrete evidence to make the detection more feasible. Clinical text used in combination with images offers complementary views which enable more context-sensitive and accurate anomaly identification [4]. Most of the existing literature have confined their focus on topics like classification and image-text retrieval rather than anomaly detection, hence, leaving the studies on multimodal fusion unexplored despite the improvement in the performance of unimodal analysis [26]. Embedding methods are the cornerstone of multimodal analysis. It helps the complex, high-dimensional multimodal data to be represented into a low-dimensional feature spaces. BioClinicalBERT, a clinical form of BERT was able to encapsulate the semantic nuances of clinical narratives with deep contextual embedding, effectively outpacing more traditional NLP methods in clinical contexts including diagnostic coding, sentiment analysis, and anomaly detection of clinical text [2][18][23]. It not only helps convert natural language into vector embedding, but also captures the deeper contextual meaning between the relationship between the words in the clinical setting, contributing to a robust feature for anomaly detection.

Similarly, convolutional neural networks (CNNs) have gained much popular for feature extraction when applying machine learning techniques to medical imaging. As they are capable of learning complex hierarchical features, they are considered to be a favorable choice. More specifically ResNet-18, a pre-trained CNN architecture that includes residual connections, which significantly improved gradient flow, allowed for deeper networks, and could be used accurately in medical imaging contexts like classification, segmentation, and detection [14][25][24]. Moreover, the extensive practical use of ResNet-18 in medical imaging illustrates its superior performance and use of computational resources in processing images compared to traditional CNNs [11][16].

The proposed study aims to leverage the contextual understanding capabilities of BioClinicalBERT to represent textual data in a low-dimensional feature space, combined with the feature extraction strengths of the pre-trained CNN model ResNet-18. These representations are utilized by anomaly detection algorithms, specifically Isolation Forest (IF) and Local Outlier Factor (LOF), to perform multimodal anomaly detection effectively. Isolation Forest (IF) and Local Outlier Factor (LOF) are two of the most popular methods used to perform anomaly detection because of their complimentary strengths. Isolation Forest is a global anomaly detection method that can efficiently isolate anomalies by randomly splitting the feature space and is known to be efficient when working with datasets containing many observations [20]. LOF, on the other hand, is based on neighborhood density for detecting local anomalies allowing it to pinpoint more subtle deviations among dense clusters [5]. The use of IF and LOF within multimodal methods provides both global and local perspectives to improve detection [21].

Hybrid approaches with conjunctions of Isolation forest and Local outlier factor have also received major interest in the studies. These hybrid approaches include the use of global detection methods such as IF or other global approaches for broader filtration which is further followed by localized density approaches such as LOF for further refinement. They have shown to create a balance between computation and precision using two approaches, reducing false positives, which helped with recall in high dimensionally medical datasets [13][1][6]. Furthermore, the proposed study also explains the efficiency of these hybrid methods in a variety of domains, therefore, meaningful in the domain of healthcare anomaly in recognition of its contribution to the area of their research interest [7][22].

Despite increasing interest in combining Isolation Forest with Local Outlier Factor or deep

Table 1. Comparison of the proposed study with existing literature

Study	Domain	Modalities	Feature Source	Hybrid Variant	Clinical	Key Finding
Alsini et al. (2021)	Concrete-mix quality	Tabular	Shallow hand-crafted	IF + LOF (sliding window)	No	Outperformed standalone LOF on 1,030 mixes
Cheng et al. (2019)	8 public benchmarks	Tabular	Shallow hand-crafted	Progressive IF→LOF	No	Higher detection rate with lower cost
Zhou et al. (2021)	Welding inspection	Multichannel sensors	Shallow hand-crafted	IF + LOF	No	IF achieved better class separation than LOF
John & Naaz (2019)	Credit-card fraud	Tabular	Shallow hand-crafted	IF + LOF	No	LOF AUROC ≈ 0.97
Fadul (2023)	Credit-card fraud	Tabular	Shallow hand-crafted	IF + LOF	No	LOF AUROC ≈ 0.97
Heigl et al. (2021)	Streaming outliers	Tabular	Shallow hand-crafted	Extended IF (no LOF)	No	Resource-aware framework improved detection
Xu et al. (2023)	Tabular, graph, time-series	Mixed	Deep random reps.	Deep IF (no LOF)	No	Outperformed classical IF on hard anomalies
Proposed Framework	Chest X-ray anomaly	Image + Text	ResNet-18 + BioClinicalBERT	IF + LOF	Yes	Reliable detection of clinical anomalies using fused text and image embeddings (Jaccard ≥ 0.4)

models, most prior work remains unimodal and non-clinical. Alsini et al. applied IF and LOF in a sliding-window to detect flawed concrete-mix recipes, outperforming standalone LOF on 1,030 samples, though inputs were limited to tabular variables [3]. Cheng et al. proposed a two-layer ensemble: first evaluating inlier with IF, then LOF to refine the candidate set. They reported an increase in the detection rate with a more computationally efficient approach in eight public datasets [9]. Zhou et al. combined IF and LOF to monitor 308,274 butt-weld joints using multichannel sensors and found IF yielded better class separation than LOF, but like prior studies, omitted textual data [28]. In finance, John Naaz and Fadul used IF–LOF for credit-card fraud detection, reporting AUROC 0.97 using only shallow features [17, 12]. Heigl et al. integrated IF into a resource-aware streaming pipeline but excluded a density-based second stage [15]. Xu et al. introduced Deep IF using random representations, improving performance on tabular, graph, and time-series data, yet remained single-modal and single-stage [27]. None of these efforts incorporate pre-trained clinical text embeddings or test a global–local hybrid on multimodal medical data; revealing a clear gap in the literature.

- Most existing studies are unimodal, overlooking the complementary value of combining radiology images with associated clinical reports.
- Hybrid IF-LOF methods are often built on shallow or hand-crafted features, lacking the representation power of domain-specific, pre-trained embeddings.
- Current IF-LOF applications are predominantly non-clinical, with limited exploration in real-world medical imaging contexts such as radiology.

The proposed study aims to address this gap by:

– Proposing a multimodal anomaly detection framework that integrates BioClinicalBERT-based text embeddings with ResNet-18 image features, specifically for chest X-ray data.
– Demonstrating that simple, interpretable global-local anomaly detectors (IF–LOF) can uncover outliers in widely used datasets, potentially contributing to improved dataset quality and reliability.
– Highlighting the value of incorporating pre-trained clinical language models to enrich the diagnostic context in anomaly detection tasks, supporting better understanding of multimodal clinical data.

Table 1 provides a visual comparison between the proposed approach and the existing literature.

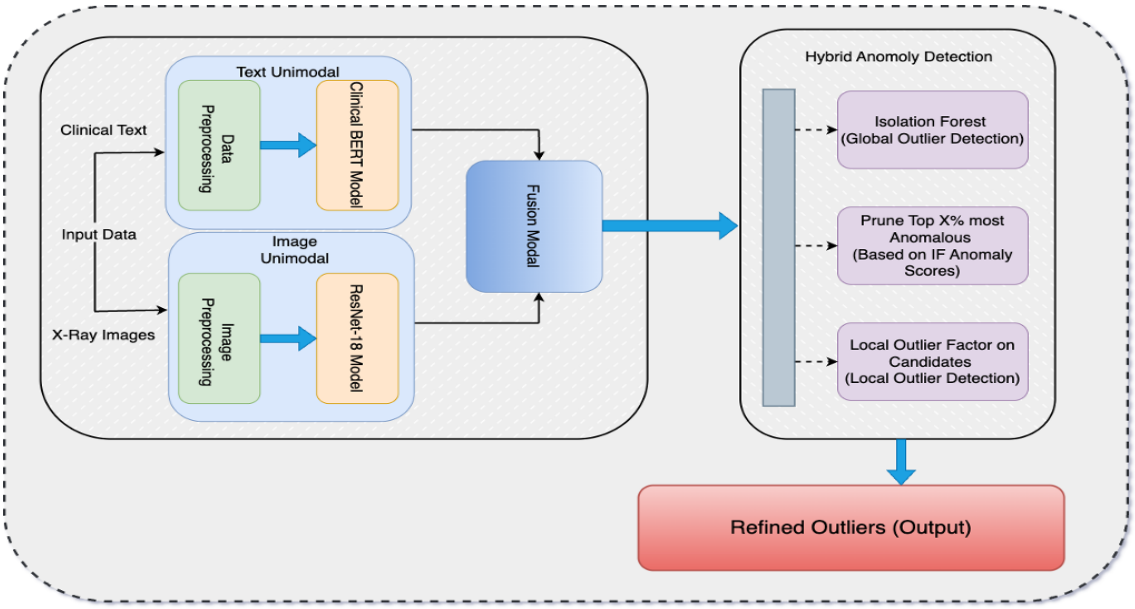

Fig. 1. Overall architecture of the proposed hybrid IF–LOF anomaly detection framework using multimodal embeddings from BioClinicalBERT and ResNet-18.

2 Materials & Methods

This study proposes an multimodal unsupervised hybrid anomaly detection framework based on Isolation Forest and Local Outlier Factor. The text data were represented using BioClinicalBERT and image features were extracted using ResNet-18. The text and image embeddings were concatenated to represent information and supplement semantics that exists in clinical notes and visualization presented in the X-ray scans. Overall framework consists of important steps that included data pre-processing, unimodal feature extraction, embedding fusion and anomaly detection using a hybrid IF-LOF approach. The overall workflow is demonstrated in the architectural diagram in Figure 1.

2.1 Dataset

The study uses the Open-I chest X-ray dataset provided via Hugging Face which includes paired frontal and lateral chest X-ray images and corresponding clinical reports [10]. Each

entry includes MeSH terms, problem descriptions, indications, comparison, findings, impressions, img_frontal and img_lateral. The total number of entries is the dataset was found to be 3851.

2.2 Data Pre-processing and Unimodal Embedding generation

The dataset contains a significant number of missing values, which are visually illustrated in the missingness pattern shown in Figure 2. Table 2 shows how missing text features were handled while generating structured text input. Clinical text fields (MeSH, Problems, Indications, Findings, and Impressions) were concatenated into structured narratives. Text pre-processing includes stop words removal and punctuation removal, while medical stop words were not removed. The cleaned text was tokenized and embedded using the pretrained BioClinicalBERT model. The [CLS] token representation was extracted to represent each report.

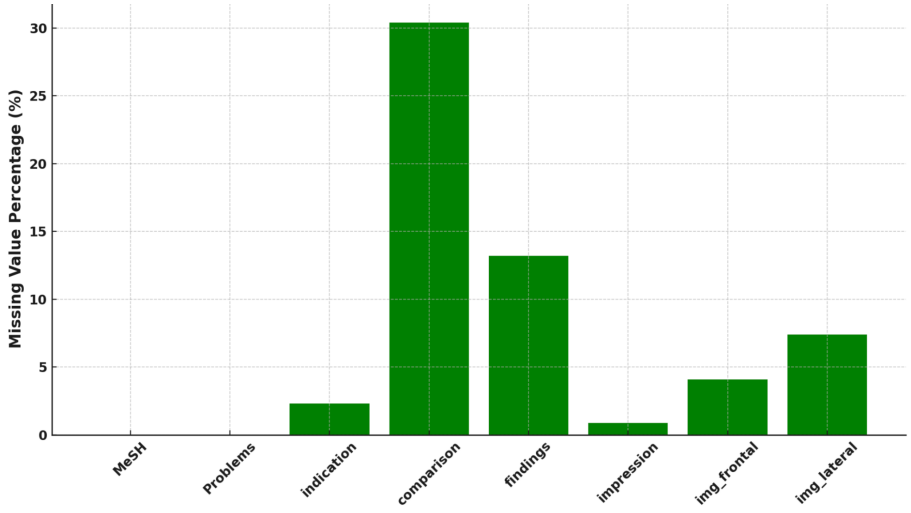

Fig. 2. Missing values per feature in percentage.

Table 2. Handling of Missing Fields in Clinical Text Data

Missing Field	**Solution**
Indication	Replaced with common values from the dataset
Findings, Impression	Replaced with most common values from the dataset
Comparison	Not included in structured text input

Similarly, for the image processing; frontal and lateral images were resized to 224×224 pixels and normalized. The features in the images were extracted using a pre-trained ResNet-18 model, with the final classification layer removed. The resulting 512-dimensional embeddings for each image were concatenated to form a 1024-dimensional vector. For the missing images, it was handled as represented in the Table 3.

Table 3. Handling of Missing Image Data

Missing Field	**Solution**
Frontal Image	Replaced with mean frontal image embeddings
Lateral Image	Replaced with mean lateral image embeddings
Both Images	Substituted with zero vectors

2.3 BioClinicalBERT for Text Embedding Generation

Each clinical sample i contains four primary textual fields; Indication, Findings, Comparison and Impression. These components are aggregated to form a structured clinical input document:

$$T_i = T_i^{(\text{ind})} \circ T_i^{(\text{find})} \circ T_i^{(\text{comp})} \circ T_i^{(\text{imp})} \tag{1}$$

The aggregated text T_i is tokenized and fed into the BioClinicalBERT encoder Φ_{BERT}, a domain-adapted language model pretrained on MIMIC-III clinical notes [2]. The contextual representation $\mathbf{e}_i \in \mathbb{R}^d$ is computed via mean pooling over the final hidden states:

$$\mathbf{e}_i = \text{MeanPool}(\Phi_{\text{BERT}}(T_i)) \tag{2}$$

The resulting vector $\mathbf{e}_i$ captures rich semantic information from structured radiology narratives. This embedding is later fused with the visual representation $\mathbf{v}_i$ extracted via ResNet-18 to form the final multimodal feature vector.

2.4 RESNET-18 for Image Embedding Generation

Each clinical sample i includes a frontal image $I_i^{(f)}$ and a lateral image $I_i^{(l)}$ which are passed through a pre-trained ResNet-18 encoder to generate feature embeddings:

$$\mathbf{v}_i^{(f)} = \Phi_{\text{ResNet}}(I_i^{(f)}), \quad \mathbf{v}_i^{(l)} = \Phi_{\text{ResNet}}(I_i^{(l)}) \tag{3}$$

These embeddings are concatenated to form the final image representation:

$$\mathbf{v}_i = [\mathbf{v}_i^{(f)} \,\|\, \mathbf{v}_i^{(l)}] \tag{4}$$

The image vector $\mathbf{v}_i$ is then combined with BioClinicalBERT text embedding $\mathbf{e}_i$ to yield the multimodal feature vector. The final representation $\mathbf{z}_i$ is used in a hybrid isolation Forest–Local Outlier Factor for multi modal anomaly detection.

$$\mathbf{z}_i = [\mathbf{e}_i \,\|\, \mathbf{v}_i] \tag{5}$$

The proposed study used the encoder from the "torchxrayvision" library, which provides domain-specific feature representations well-suited to radiographs. This clinically relevant pretraining improves embedding quality and downstream anomaly detection.

2.5 Hybrid IF-LOF Model

In order to identify anomalies in the multi modal feature space, the study develops a two-stage hybrid pipeline that takes advantage of both global and local methods of outlier detection. In the first stage, Isolation Forest $\mathcal{F}_s$, initialized with a given random seed s, is trained on the entire standardized feature matrix $\mathbf{Z} \in \mathbb{R}^{n \times d}$, where each row $\mathbf{z}_i$ represents a multimodal feature vector (text + image) for sample i. The anomaly score for each point is defined as the average depth at which it gets isolated in the trees, inverted to reflect anomaly strength:

$$a_i^{(s)} = -\frac{1}{|\mathcal{F}_s|} \sum_{T \in \mathcal{F}_s} \ell_T(\mathbf{z}_i) \tag{6}$$

where, $\ell_T(\mathbf{z}_i)$ is the path length in tree T for point i. A higher score indicates a more anomalous point. To focus LOF on the most promising candidates, a pruning step is applied. A threshold $\tau_p^{(s)}$ is computed at the pth percentile (90, 92, 95) of all IF scores, and only samples with scores above this threshold are retained for the next stage:

$$\mathcal{C}_p^{(s)} = \left\{ i \mid a_i^{(s)} \geq \tau_p^{(s)} \right\} \tag{7}$$

Percentiles 90–95 were selected to incrementally assess how increasing selectivity in IF affects the precision of downstream LOF refinement and were chosen based on exploratory runs ensuring sufficient anomaly candidates for LOF analysis. In the second stage, Local Outlier Factor is applied to just these top-scoring samples $\mathbf{Z}_{\mathcal{C}_p^{(s)}}$. LOF computes how much a sample deviates from the density of its neighbors. This is quantified by comparing its local reachability density (LRD) with that of its neighbors:

$$\mathrm{LOF}_k(\mathbf{z}_i) = \frac{1}{|N_k(i)|} \sum_{j \in N_k(i)} \frac{\mathrm{lrd}_k(j)}{\mathrm{lrd}_k(i)} \tag{8}$$

Points that receive a LOF score significantly greater than 1 are considered local outliers. Finally, the set of confirmed outliers under this hybrid scheme is defined as:

$$\mathcal{O}_{s,p,c} = \left\{ i \in \mathcal{C}_p^{(s)} \mid \mathrm{LOF}_k(\mathbf{z}_i) > \theta_c \right\} \tag{9}$$

where, θ_c is a threshold determined by the contamination setting $c \in \{0.01, 0.02, \text{auto}\}$. This approach balances global partitioning of the feature space (via IF) with local density-based refinement (via LOF). Importantly, while IF provides a global structural mechanism, the true semantic generality is introduced via BioClinicalBERT embeddings, which encode full clinical narratives. This combination enables interpretable and semantically informed anomaly detection in clinical data.

3 Results and Discussion

The proposed hybrid IF–LOF pipeline was evaluated on 3,851 multimodal samples from the Open-I chest X-ray dataset. To ensure robustness, the model was run with five random seeds across different parameter settings. Since both IF and LOF incorporate randomness in tree construction and neighborhood density [20, 5], running multiple seeds helped

Table 4. Summary of Outliers under selected Hybrid Configurations

Configuration	Seed	Pruning Percentile	LOF Setting	Outlier Indices
1	42	90	0.01	[121, 1453, 1632, 2673]
2	42	90	0.02	[121, 1453, 1473, 1632, 1919, 2200, 2265, 2673]
3	42	92	0.01	[121, 1632, 2265, 2673]
4	42	92	0.02	[121, 1473, 1632, 1919, 2200, 2265, 2673]
5	42	95	0.01	[1632, 2673]
6	42	95	0.02	[121, 1632, 2265, 2673]
7	72	90	0.01	[655, 1632, 2265, 2673]
8	72	90	0.02	[121, 479, 655, 1453, 1632, 2200, 2265, 2673]
9	72	92	0.01	[655, 1632, 2265, 2673]
10	72	92	0.02	[121, 479, 655, 1453, 1632, 2265, 2673]
11	72	95	0.01	[2265, 2673]
12	72	95	0.02	[655, 1632, 2265, 2673]
13	92	90	0.01	[121, 1632, 2673, 2790]
14	92	90	0.02	[121, 655, 1453, 1632, 1919, 2265, 2673, 2790]
15	92	92	0.01	[121, 1632, 2265, 2790]
16	92	92	0.02	[121, 655, 1473, 1632, 1919, 2265, 2790]
17	92	95	0.01	[1632, 1919]
18	92	95	0.02	[121, 655, 1632, 1919]
19	112	90	0.01	[655, 1632, 2265, 2673]
20	112	90	0.02	[121, 655, 1453, 1473, 1632, 1919, 2265, 2673]
21	112	92	0.01	[121, 655, 1632, 2673]
22	112	92	0.02	[121, 655, 1473, 1632, 2265, 2673, 3569]
23	112	95	0.01	[1632, 2673]
24	112	95	0.02	[655, 1632, 2265, 2673]
25	212	90	0.01	[121, 655, 2265, 2673]
26	212	90	0.02	[121, 655, 1453, 1473, 1632, 1919, 2265, 2673]
27	212	92	0.01	[121, 655, 2265, 2673]
28	212	92	0.02	[121, 655, 1453, 1632, 1919, 2265, 2673]
29	212	95	0.01	[2265, 2673]
30	212	95	0.02	[655, 1632, 2265, 2673]

confirm consistency in detected outliers. Table 4 shows that increasing the IF pruning percentile reduced LOF-confirmed anomalies, reflecting a more selective candidate set.

In conjunction, to assess the consistency of anomaly detection across multiple experimental configurations, the Jaccard similarity was computed between sets of outlier indices computed by the hybrid IF-LOF pipeline. In the context of each combination of random seed, pruning percentile, and LOF contamination setting, an subset of anomalies were generated and retained. The Jaccard indices were then computed pair-wise across all sets of results using the following equation.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{10}$$

where, A and B are two sets of detected outlier indices from different configurations. This allowed for the creation of a Jaccard similarity matrix with all combinations of outlier detection found across seeds and parameter choices. The Jaccard similarity matrix provided a measure of the overlap of outlier detection and was used to measure stability of the model's output across experimental conditions, with higher scores indicating model output that retained anomaly detection behavior across experiments as seen in Figure 3.

To visualize the learned multimodal embeddings in a 2D space, t-SNE was employed with perplexity tuned to its default value (=30) to qualitatively evaluate the anomaly detection pipeline by seeing the degree to which the identified outliers are spatially separated

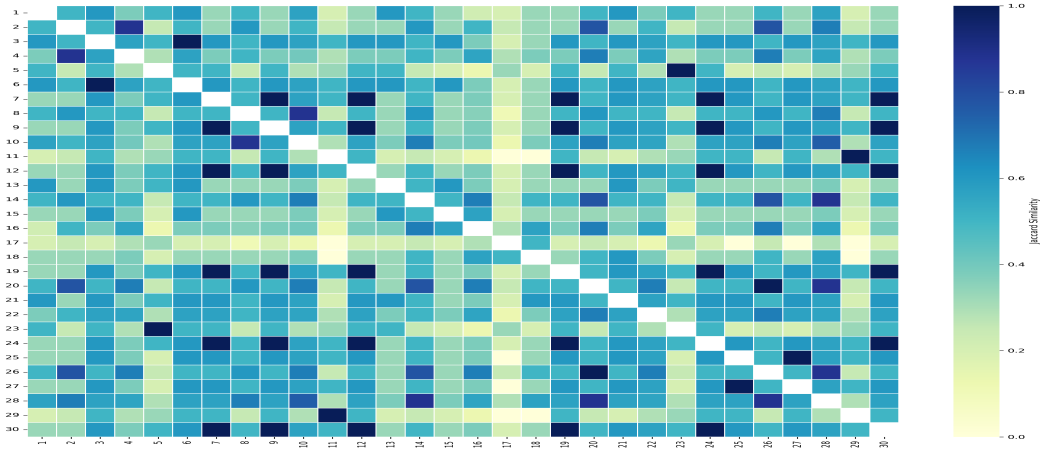

Fig. 3. Jaccard Similarity Heatmap Between Experiment Configurations.

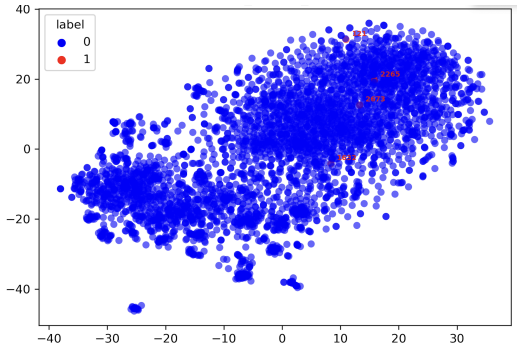

Fig. 4. t-SNE visualization for 42, pruning percentile=90, LOF=0.01

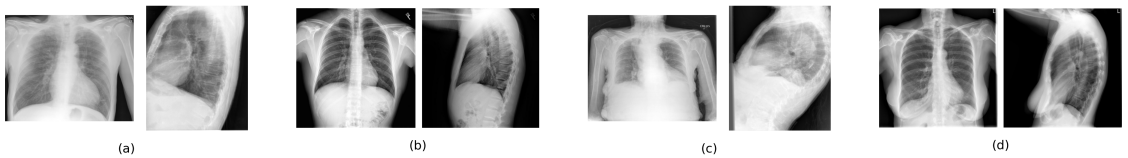

Fig. 5. Examples of normal chest X-ray image pairs (frontal and lateral views). These scans appear clear with no visible tubes, devices, or unusual opacities.

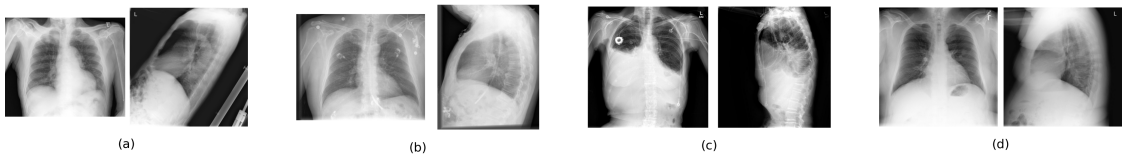

Fig. 6. Chest X-ray image pairs (frontal and lateral views) corresponding to the following indices: **(a)** index 121 – shows upper lung asymmetry; **(b)** index 1632 – visible feeding tube placement; **(c)** index 2265 – medical device and diffuse opacities; **(d)** index 2673 – abnormal density in the left midlung. These cases include structural irregularities or artificial elements that visually deviate from typical patterns.

from the normal data. t-SNE visualizations were generated for all experimental configurations; representative plots for Configurations 1 (as defined in Table 4) are presented in Figure 4. Red points in a plot denotes anomalies detected by the hybrid IF–LOF pipeline, while blue points represent inliers. The visualizations highlights the effectiveness of the

two-stage detection strategy in identifying structurally distinct and contextually sparse samples.

Across all configurations, outlier indices such as 121, 1632, 2265, and 2673 were detected repeatedly, highlighting their robust anomalous characteristics independent of specific random seed or LOF contamination level. This consistency is a desirable trait in clinical anomaly detection, as it suggests the method's reliability in identifying rare or irregular samples. Across all configuration setup, specifically outlier indices 121, 1632, 2265, and 2673, actively tracked their outlier course and consistently showed correlating variation regardless of random seed or LOF contamination. This consistency is important in clinical anomaly detection because it shows that the method reliably identifies unusual or rare cases, regardless of the specific settings used. With the increase in the pruning percentile in Isolation Forest, fewer candidate samples were passed to the LOF stage, resulting in a smaller but more focused group of anomalies which was logical; since strict filtering tends to highlight only the most unusual cases. The t-SNE plots also showed instances of the outliers being clearly separated from the rest of the anomalies, supporting that the combined text and image features effectively capture meaningful differences in the data.

To qualitatively evaluate the model's ability to distinguish normal from abnormal cases, we present representative chest X-ray image pairs in Figures 5 and 6. Each pair includes frontal and lateral views to reflect the multi-view nature of the input. Figure 5 showcases examples of normal scans, which appear clear and unremarkable, with no visible medical devices, structural abnormalities, or unusual opacities. In contrast, Figure 6 highlights cases that were flagged as anomalous by our model. These include: (a) index 121, which shows upper lung asymmetry; (b) index 1632, where a feeding tube is visibly placed; (c) index 2265, which contains a medical device and diffuse opacities; and (d) index 2673, demonstrating an abnormal density in the left midlung region. These examples illustrate the model's effectiveness in identifying visual deviations that correspond to clinically relevant anomalies or artificial elements. The alignment between model predictions and visible abnormal features supports the utility of our approach for interpretable anomaly detection in clinical imaging workflows.

The Jaccard similarity heatmap shows that many configurations have a reasonably strong overlap in the outliers all configurations detect, with similarity scores largely falling between 0.4 to 0.8. This level of consistency suggests the model is identifying some meaningful aspects of the data and does not just react to random noise. There were some configurations that were lower on the similarity measure, but these tended to have it due to random nature of IF and LOF. Result highlights the importance of testing across different seeds and settings; especially in sensitive areas like healthcare allowing to trust the model's performance. Overall, the experiments presented indicate that the hybrid IF-LOF pipeline is effective in combining both global and local views to identify anomalies, specifically in the context of complex, multimodal clinical data. Both BioClinicalBERT and ResNet-18 were effective for generating text and images embeddings respectively, and provided a rich and complementary representation to more reliably capture anomalous values.

4 Conclusion and Limitations

The proposed study presents a hybrid anomaly detection model that integrates clinical language and imaging data via BioClinicalBERT and ResNet-18. The approach combines Isolation Forest for global partitioning over the multimodal feature space with Local Outlier Factor for refined local analysis. While IF operates across the entire embedding space,

it can still reflect local patterns due to shallow tree splits. In contrast, BioClinicalBERT provides true global semantic abstraction by encoding full clinical narratives. This distinction enhances interpretability and reflects the complementary nature of both components. The dual detection methodology consistently identified clinically significant outliers, as evidenced by the recurrence of certain data points and their separation in t-SNE plots. Furthermore, the Jaccard similarity model output analysis indicated reliability across varying hyperparameter and random seed selections. The framework was consistently reproducible across model checking experiments, which is advantageous in clinical applications where reproducing results is key. The proposed framework is applicable to other 2D medical imaging modalities with associated text, such as ultrasound, dermatology, and pathology beyond chest X-rays. This adds up the eligibility of the framework to work with 2D image modalities for clinical generalizability.

However, the study does have some limitations and needs further analysis to be included as a part of future work.

- No expert validation, so clinical relevance of anomalies is uncertain. This limits the ability to draw strong real-world conclusions.
- Manually chosen pruning percentile for Isolation Forest and LOF contamination thresholds may further benefit from Adaptive thresholding.
- The proposed framework is not directly applicable to 3D volumetric data like MRI or CT without architectural modifications, such as integrating 3D CNNs or transformers and aligning multi-slice volumes with sectioned reports.

References

1. Charu C Aggarwal and Charu C Aggarwal. *An introduction to outlier analysis*. Springer, 2017.
2. Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.
3. Raed Alsini, Abdullah Almakrab, Ahmed Ibrahim, and Xiaogang Ma. Improving the outlier detection method in concrete mix design by combining the isolation forest and local outlier factor. *Construction and Building Materials*, 270:121396, 2021.
4. Ivo M Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific reports*, 9(1):6381, 2019.
5. Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.
6. Guilherme O Campos, Arthur Zimek, Jörg Sander, Ricardo JGB Campello, Barbora Micenková, Erich Schubert, Ira Assent, and Michael E Houle. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data mining and knowledge discovery*, 30:891–927, 2016.
7. Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.
8. Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.
9. Zhangyu Cheng, Chengming Zou, and Jianwei Dong. Outlier detection using isolation forest and local outlier factor. In *Proceedings of the conference on research in adaptive and convergent systems*, pages 161–168, 2019.
10. Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.
11. Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.
12. Abubakar Mastour Adam Fadul. Anomaly detection based on isolation forest and local outlier factor. *Africa University*, 2023.

13. Markus Goldstein and Seiichi Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one*, 11(4):e0152173, 2016.
14. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
15. Michael Heigl, Kumar Ashutosh Anand, Andreas Urmann, Dalibor Fiala, Martin Schramm, and Robert Hable. On the improvement of the isolation forest algorithm for outlier detection with streaming data. *Electronics*, 10(13):1534, 2021.
16. Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.
17. Hyder John and Sameena Naaz. Credit card fraud detection using local outlier factor and isolation forest. *Int. J. Comput. Sci. Eng*, 7(4):1060–1064, 2019.
18. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
19. Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.
20. Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.
21. Mohsin Munir, Shoaib Ahmed Siddiqui, Andreas Dengel, and Sheraz Ahmed. Deepant: A deep learning approach for unsupervised anomaly detection in time series. *Ieee Access*, 7:1991–2005, 2018.
22. Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021.
23. Dhavalkumar Patel, Prem Timsina, Larisa Gorenstein, Benjamin S Glicksberg, Ganesh Raut, Satya Narayan Cheetirala, Fabio Santana, Jules Tamegue, Arash Kia, Eyal Zimlichman, et al. Traditional machine learning, deep learning, and bert (large language model) approaches for predicting hospitalizations from nurse triage notes: Comparative evaluation of resource management. *JMIR AI*, 3(1):e52190, 2024.
24. Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019.
25. Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.
26. Xing Wu, Jingwen Li, Jianjia Wang, and Quan Qian. Multimodal contrastive learning for radiology report generation. *Journal of Ambient Intelligence and Humanized Computing*, 14, 09 2022.
27. Hongzuo Xu, Guansong Pang, Yijie Wang, and Yongjun Wang. Deep isolation forest for anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12591–12604, 2023.
28. Lei Zhou, Tianyi Zhang, Zhongdian Zhang, Zhenglong Lei, and Shiliang Zhu. A new online quality monitoring method of chain resistance upset butt welding based on isolation forest and local outlier factor. *Journal of Manufacturing Processes*, 68:843–851, 2021.

Authors

Satish Wagle is a PhD student in Computer Science at Middle Tennessee State University (MTSU). He received his MS in Computer Science and holds a Bachelor of Science in Information Technology. His research interests include machine learning, deep learning, big data algorithms, and medical imaging.

Saroj Baral is a Master's student in Computer Science at Middle Tennessee State University (MTSU). He holds a Bachelor degree in Computer Engineering. His research interests

include computer vision, machine learning, large language models, deep learning methods, and geometric learning.

Dr. Jorge M. Vargas is an Associate Professor in the Department of Engineering Technology at Middle Tennessee State University (MTSU), Murfreesboro, TN, USA. He received his PhD in Electrical Engineering. His research interests include microelectronics, MEMS, RF/microwave engineering, radar sensors for autonomous vehicles, sensor fusion, and machine learning.

Dr. Khem Poudel is an Assistant Professor in the Department of Computer Science at Middle Tennessee State University (MTSU), Murfreesboro, TN, USA. He received his PhD in Computer Science. His research interests include machine learning, high-performance computing, big data, healthcare informatics.