Out-of-Distribution Detection with Attention Head Masking for Multimodal Document Classification

Anonymous ACL submission

Abstract

Detecting out-of-distribution (OOD) data is crucial in machine learning applications to mitigate the risk of model overconfidence, thereby enhancing the reliability and safety of deployed systems. The majority of existing OOD detection methods predominantly address uni-modal inputs, such as images or texts. In the context of multi-modal documents, there is a notable lack of extensive research on the performance of these methods, which have primarily been developed with a focus on computer vision tasks. We propose a novel methodology termed as attention head masking (AHM) for multi-modal OOD tasks in document classification systems. Our empirical results demonstrate that the proposed AHM method outperforms all state-of-the-art approaches and significantly decreases the false positive rate (FPR) compared to existing solutions up to 7.5%. This methodology generalizes well to multi-modal data, such as documents, where visual and textual information are modeled under the same Transformer architecture. To address the scarcity of high-quality publicly available document datasets and encourage further research on OOD detection for documents, we introduce FinanceDocs, a new document AI dataset. Our code[1] and dataset[2] are publicly available.

1 Introduction

Out-of-distribution (OOD) detection presents a significant challenge in the field of document classification. When a classifier is deployed, it may encounter types of documents that were not included in the training dataset. This can lead to mishandling of such documents, causing additional complications in a production environment. Effective OOD detection facilitates the identification of unfamiliar documents, enabling the system to manage them appropriately which allows the classifier to maintain its reliability and accuracy in real-world applications. This has heightened the focus on OOD detection, where the primary objective is to determine if a new document belongs to a known in-distribution (ID) class or an OOD class. A significant challenge lies in the lack of supervisory signals from the unknown OOD data, which can encompass any content outside the ID classes. The complexity of this problem increases with the semantic similarity between the OOD and ID data (Fort et al., 2021).

A number of approaches have been developed to differentiate OOD data from ID data, broadly classified into three categories: (i) confidence-based methods, which focus on softmax confidence scores (Liu et al., 2020; Hendrycks and Gimpel, 2016;

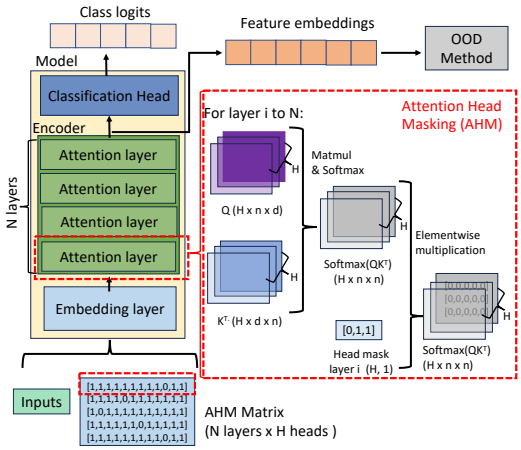


Figure 1: Visual demonstration of AHM on a transformer-based model: For each attention layer, we utilize the corresponding attention head mask from the AHM matrix. Following query-key multiplication and the subsequent softmax operation, the resulting attention scores undergo element-wise multiplication with the relevant attention head mask. This process effectively reduces the attention scores of certain heads to zero, thereby inhibiting the propagation of their respective information through the value matrix.

[1] https://anonymous.4open.science/r/OOD-AHM-FE25/README.md

[2] https://drive.google.com/drive/folders/1dV9obe_3hTsDoWJyYuNLBAXEiwOPwCw7

Hendrycks et al., 2019; Huang et al., 2021; Liang et al., 2017), (ii) features/logits-based methods, which emphasize logit outputs Sun and Li (2021); Sun et al. (2021); Wang et al. (2022); Djurisic et al. (2023), and (iii) distance/density-based methods, which concentrate on dense embeddings from the final layers (Ming et al., 2023; Lee et al., 2018; Sun et al., 2022). Recent research also investigates domain-invariant representations, such as HYPO (Ming et al., 2024), and introduces new OOD metrics like NECO (Ammar et al., 2024), which leverage neural collapse properties (Papyan et al., 2020). Confidence-based methods can be unreliable as they often yield overconfident scores for OOD data. Features/logits-based methods attempt to combine class-agnostic scores from the feature space with the ID class-dependent logits. Our approach focuses on identifying more robust class-agnostic scores from the feature space, and as such, we conduct our experiments using distance/density-based methods.

Many OOD detection techniques have been developed, but most have been evaluated only in unimodal systems, such as text or images, and not extensively tested in the document domain (Gu et al., 2023). This may be due to the lack of high-quality public document datasets, mostly based on IIT-CDIP (et al., 2006). To address the lack of comprehensive research in the document domain, we introduce a new document AI dataset, FinanceDocs. Additionally, we propose a novel technique called attention head masking (AHM) to effectively improve feature representations for distinguishing between ID and OOD data. Our method is illustrated in Figure 1. Our contributions can be summarized as follows: **(1) FinanceDocs Dataset**: We introduce FinanceDocs, the first high-quality digital document dataset for OOD detection with multimodal documents, offering digital PDFs instead of low-quality scans. **(2) AHM**: We propose a multi-head attention masking mechanism for transformer-based models applied post-fine-tuning. By identifying masks that enhance similarity between ID training and evaluation features, we generate robust representations that improve the separation of ID and OOD data using distance/density-based OOD techniques. Our AHM method surpasses existing OOD solutions on key metrics.

2 Related Work

Learning embedding representations that generalize effectively and facilitate better differentiation between ID and OOD data is a well-recognized challenge in the field of machine learning (Zhou et al., 2023). To tackle this challenge, various studies have focused on specialized learning frameworks aimed at optimizing intra-class compactness and inter-class separation (Ye et al., 2021). Building on the principles of contrastive representation learning, researchers such as Chen et al. (2020) and Li et al. (2021) introduced prototypical learning (PL). This approach leverages prototypes derived from offline clustering algorithms to enhance unsupervised representation learning. Furthermore, Ming et al. (2024) integrated PL into their OOD learning framework, HYPO, achieving effective separation between ID and OOD data. This line of research was further advanced by Lu et al. (2024), who introduced the concept of multiple prototypes per cluster and employed a maximum likelihood estimation (MLE) loss to ensure that sample embeddings closely align with their corresponding prototypes. Additionally, approaches such as VOS (Du et al., 2022) and NPOS (Tao et al., 2023) have focused on regularizing the decision boundary between ID and OOD data by generating synthetic OOD samples, while Lin and Gu (2023) utilized open-source data as an OOD signal.

In our proposed methodology, we similarly aim to enhance the distinction between ID and OOD data through improved embedding representations. However, unlike previous studies that explore customized learning frameworks diverging from the standard cross-entropy loss, we concentrate on feature regularization during inference using our proposed attention head masking methodology. Our approach deliberately avoids altering the network's training procedure, thereby mitigating potential negative impacts on performance and preventing increased training costs. By focusing on inference rather than training modifications, our method ensures robust and cost-effective OOD detection.

Other inference-based methods, such as Avg-Avg (Chen et al., 2022) and Gnome (Chen et al., 2023), have also sought to enhance OOD detection through innovative techniques. Avg-Avg operates by averaging embeddings across both sequence length and different layers of a fine-tuned model, while Gnome combines embeddings from both a pre-trained and a fine-tuned model. These approaches, like our own, emphasize the importance of embedding manipulation during inference to achieve improved OOD detection without modifying the underlying training framework.

3 Method

The proposed AHM method, focuses on the feature extraction mechanisms inherent in transformer models, specifically the self-attention mechanism (Vaswani et al., 2017). Based on the premise that OOD data exhibit less semantic similarity to ID data, our goal is to generate embedding features that enhance the separation between ID and OOD data. The embeddings are then used in distance or density-based OOD detection methods, such as the Mahalanobis (Lee et al., 2018) or kNN+ (Sun et al., 2022). Our method is provided in Algorithm 1 (cf. Appendix A.1 for the theoretical framework) and the masking step is summarised in Figure 1.

Algorithm 1 Optimization of Transformer-based Model using Attention Head Masking for OOD Detection – cf. Appendix A.2 for more details

Input: Budget T, model weights $W_{\text{pretrained}}$, percentage masking p, neighbors K, layers N, attention heads H, top attention head matrices to select F
Output: Optimal ensemble embedding
1. Fine-tune Model $W_{\text{pretrained}} \rightarrow W_{\text{finetuned}}$
for trial = 1 to T **do**
  2. InitializeAttention Head Matrix
  Create $N \times H$ matrix A, $A[i, j] = 1$
  3. Mask Attention Heads
  Randomly set elements of $A[i, j]$ to 0
  4. Extract Embeddings
  Extract $\text{embed}_{\text{train}} \in \mathbb{R}^{O \times Hid}$ and $\text{embed}_{\text{eval}} \in \mathbb{R}^{Q \times Hid}$
  5. Compute Similarity Scores
  For $e_i \in \text{embed}_{\text{eval}}$, get K nearest neighbors in $\text{embed}_{\text{train}}$ and compute mean score S_i
  6. Assign and Collect Scores
  Average similarity score: $\frac{1}{Q}\sum_{i=1}^{Q} S_i$. Collect scores S_i and their respective $A[i, j]$
end for
7. Select Top Scores
Sort scores S_i , select top F masks $A[i, j]$
8. Ensemble Embedding Generation
Use top F masks $A[i, j]$ to generate and average embeddings for OOD detection

4 Results and Discussion

4.1 Datasets

We utilized two datasets in our experiments: Tobacco3482 and FinanceDocs. The Tobacco3482 dataset (Kumar et al., 2014) comprises 10 classes: Memo (619), Email (593), Letter (565), Form (372), Report (261), Scientific (255), Note (189), News (169), Advertisement (162), and Resume (120). As a subset of IIT-CDIP (et al., 2006), it was further processed to remove blank and rotated pages, preserving the rich textual and image modalities essential for a multi-modal system. Despite these efforts, some instances exhibit poor OCR quality due to the low-quality scans.

We present FinanceDocs (cf. Appendix A.5 for per-category details and A.6 for dataset samples), a newly created dataset comprising 10 classes derived from open-source financial documents, including SEC Form 13 (663), Financial Information (360), Resumes (287), Scientific AI Papers (267), Shareholder Letters (256), List of Directors (188), Company 10-K Forms (181), Articles of Association (176), SEC Letters (141), and SEC Forms (121). Unlike Tobacco3482, FinanceDocs consists of high-quality digital PDFs (Annual Reports; SEC EDGAR Database; Companies House Service; ACL Anthology; Resume Dataset). The FinanceDocs dataset was labeled through the following process: a PDF parsing package (`PyPDF2`) was used to extract content from the original PDF documents. Each page was then visualized individually by a human annotator, who determined the relevance of the page to the collected classes and assigned the appropriate class label (cf. Appendix A.4 for annotator training and validation).

4.2 Experimental Setup

We employ two widely recognized OOD metrics to assess the performance of our proposed AHM method in comparison to other OOD benchmarks (Yang et al., 2024): AUROC, which measures the area under the ROC curve (higher values indicate better performance), and FPR, the false positive rate at a 95% true positive rate. A higher AUROC signifies better discrimination, while a lower FPR indicates greater robustness in rejecting OOD data.

For our experiments, we utilize `LayoutLMv3` (Huang et al., 2022), a transformer-based multi-modal model with 125.92 million parameters. We conduct both cross-dataset and intra-dataset OOD experiments. In cross-dataset OOD, the model is trained on the classes of one dataset and evaluated on the entirety of the other dataset as OOD. In intra-dataset OOD, one of the 10 classes is designated as OOD, and the model is trained on the remaining 9 classes, with the ID data split into training and

Method	Tobacco3482 (ADVE OOD) AUROC	Tobacco3482 (ADVE OOD) FPR	Tobacco3482 (Cross-dataset OOD) AUROC	Tobacco3482 (Cross-dataset OOD) FPR	FinanceDocs (Resume OOD) AUROC	FinanceDocs (Resume OOD) FPR	FinanceDocs (Cross-dataset OOD) AUROC	FinanceDocs (Cross-dataset OOD) FPR
energy	0.951 ± 0.012	0.267 ± 0.057	0.944 ± 0.014	0.157 ± 0.042	0.848 ± 0.093	0.413 ± 0.218	0.846 ± 0.016	0.567 ± 0.039
gradNorm	0.940 ± 0.025	0.330 ± 0.116	0.824 ± 0.040	0.410 ± 0.094	0.742 ± 0.153	0.664 ± 0.251	0.724 ± 0.128	0.817 ± 0.145
kl	0.914 ± 0.016	0.448 ± 0.099	0.970 ± 0.014	0.071 ± 0.035	0.902 ± 0.040	0.295 ± 0.106	0.840 ± 0.025	0.630 ± 0.047
knn	0.958 ± 0.011	0.269 ± 0.074	0.991 ± 0.004	0.030 ± 0.018	0.965 ± 0.023	0.172 ± 0.127	0.891 ± 0.017	0.589 ± 0.067
Mahalanobis	0.976 ± 0.009	0.155 ± 0.053	0.996 ± 0.002	0.010 ± 0.009	0.977 ± 0.013	0.122 ± 0.100	0.898 ± 0.017	0.541 ± 0.090
$\text{mah}_{\text{AvgAvg}}$	0.942 ± 0.008	0.375 ± 0.054	0.997 ± 0.001	0.0004 ± 0.0005	0.996 ± 0.003	0.006 ± 0.005	0.949 ± 0.015	0.353 ± 0.196
$\text{mah}_{\text{Gnome}}$	0.971 ± 0.009	0.155 ± 0.054	0.992 ± 0.003	0.037 ± 0.016	0.938 ± 0.035	0.314 ± 0.165	0.822 ± 0.024	0.646 ± 0.114
maxLogit	0.946 ± 0.012	0.311 ± 0.063	0.945 ± 0.013	0.151 ± 0.033	0.851 ± 0.086	0.410 ± 0.203	0.846 ± 0.017	0.584 ± 0.037
msp	0.929 ± 0.009	0.471 ± 0.103	0.952 ± 0.016	0.140 ± 0.050	0.883 ± 0.041	0.400 ± 0.142	0.846 ± 0.032	0.612 ± 0.048
neco	0.971 ± 0.012	0.164 ± 0.046	0.995 ± 0.002	0.013 ± 0.011	0.975 ± 0.012	0.132 ± 0.096	0.888 ± 0.020	0.546 ± 0.114
residual	0.976 ± 0.008	0.149 ± 0.051	0.996 ± 0.002	0.011 ± 0.009	0.976 ± 0.014	0.130 ± 0.106	0.896 ± 0.016	0.541 ± 0.089
vim	0.976 ± 0.008	0.147 ± 0.044	0.996 ± 0.002	0.011 ± 0.009	0.976 ± 0.014	0.125 ± 0.101	0.899 ± 0.015	0.537 ± 0.086
knn_{AHM}	$\mathbf{0.969 \pm 0.009}$	$\mathbf{0.182 \pm 0.039}$	$\mathbf{0.991 \pm 0.003}$	$\mathbf{0.024 \pm 0.013}$	$\mathbf{0.975 \pm 0.014}$	$\mathbf{0.114 \pm 0.088}$	$\mathbf{0.885 \pm 0.011}$	$\mathbf{0.562 \pm 0.096}$
mah_{AHM}	$\mathbf{0.985 \pm 0.005}$	$\mathbf{0.071 \pm 0.041}$	$\mathbf{0.997 \pm 0.002}$	$\mathbf{0.006 \pm 0.006}$	$\mathbf{0.978 \pm 0.012}$	$\mathbf{0.099 \pm 0.086}$	$\mathbf{0.892 \pm 0.013}$	$\mathbf{0.522 \pm 0.126}$
$\text{mah}_{\text{AvgAvg\_AHM}}$	$\mathbf{0.956 \pm 0.007}$	$\mathbf{0.267 \pm 0.007}$	$\mathbf{0.998 \pm 0.001}$	$\mathbf{0.0001 \pm 0.0009}$	$\mathbf{0.996 \pm 0.003}$	$\mathbf{0.004 \pm 0.003}$	$\mathbf{0.951 \pm 0.012}$	$\mathbf{0.302 \pm 0.012}$

Table 1: Performance metrics (arithmetic mean and standard deviation) for different methods across two datasets with intra-dataset and cross-dataset experiments configurations per dataset using AUROC (higher is better) and FPR (lower is better) – (cf. Appendix A.3 for hyperparameter tuning details).

evaluation sets. We select Advertisement (ADVE) and Resumes as the OOD classes for Tobacco3482 and FinanceDocs, respectively.

The models are trained over 5 random runs, with checkpoints saved at high ID classification metrics. Checkpoints with low silhouette scores $s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$ are filtered out to optimize intra-class similarity and inter-class separation. Our experiments were conducted using a single NVIDIA A100 GPU (80GB) for 72 GPU compute hours. We trained the models for a maximum of 15 epochs with an initial learning rate of 5×10^{-5}.

4.3 Current Benchmarks

We evaluated the peformance of various OOD detection methods, comparing them with our proposed methods, **knn_{AHM}**, **mah_{AHM}**, and **$\text{mah}_{\text{AvgAvg\_AHM}}$**, which apply k-Nearest Neighbor (kNN) and Mahalanobis methods to dense embeddings generated by AHM. **$\text{mah}_{\text{AvgAvg\_AHM}}$** is similar to **$\text{mah}_{\text{AHM}}$** but uses the AvgAvg embedding aggregation method (Chen et al., 2022).

As shown in Table 1, for the Tobacco3482 dataset with ADVE as the OOD class, our proposed **mah_{AHM}** outperformed other methods, achieving an AUROC of 0.985 and an FPR of 0.071. The high AUROC indicates that our method significantly enhances the Mahalanobis distance-based approach in distinguishing between ID and OOD samples. The notably lower FPR compared to previous methods like *vim* and *residual* (FPRs of 0.147 and 0.149, respectively) demonstrates the robustness of **mah_{AHM}** in correctly rejecting OOD samples.

For the FinanceDocs dataset, with Resumes as the OOD class, both **knn_{AHM}** and **mah_{AHM}** achieved superior performance, with AUROCs of 0.975 and 0.978, and FPRs of 0.114 and 0.099, respectively. Our **$\text{mah}_{\text{AvgAvg\_AHM}}$** method also improved performance over $\text{mah}_{\text{AvgAvg}}$, highlighting the effectiveness of our approach in creating more separable embeddings between ID and OOD data. This is further evidenced by cross-dataset results in Table 1, where **$\text{mah}_{\text{AvgAvg\_AHM}}$** consistently outperformed $\text{mah}_{\text{AvgAvg}}$, notably reducing the FPR by 5% on FinanceDocs and achieving an AUROC of 0.99 with an FPR of 0.0001 on Tobacco3482. This performance surpasses the respective method $\text{mah}_{\text{AvgAvg}}$ without AHM applied. In fact, across all methods tested $\text{mah}_{\text{AvgAvg}}$, Mahalanobis and knn, the application of our AHM technique consistently resulted in improved performance.

Overall, the AHM technique significantly enhances the performance of kNN, Mahalanobis, and $\text{mah}_{\text{AvgAvg}}$, resulting in superior outcomes for **knn_{AHM}**, **mah_{AHM}**, and **$\text{mah}_{\text{AvgAvg\_AHM}}$**, as evidenced by higher AUROCs and lower FPRs across intra-dataset and cross-dataset experiments, demonstrating strong generalizability across diverse datasets and methods.

5 Conclusion

In this study, we present the AHM technique for OOD detection in transformer-based document classification. Our methods, **knn_{AHM}**, **mah_{AHM}** and **$\text{mah}_{\text{AvgAvg\_AHM}}$**, demonstrated significant improvements in AUROC and FPR metrics across various datasets. These results underscore the effectiveness of optimizing attention mechanisms to enhance feature separation between ID and OOD data. Additionally, we introduce the FinanceDocs dataset, contributing valuable resources to OOD detection research. Our findings highlight AHM as a promising approach for achieving robust and accurate OOD detection in document classification.

6 Limitations

While AHM techniques significantly reduced FPR in most cases, the improvements were marginal in cross-dataset scenarios where the Tobacco dataset served as the OOD data. This suggests a potential dependency on specific datasets. Additionally, AHM is a technique limited to attention-based DNN architectures that employ multi-head self-attention. Future research should aim to broaden the range of datasets explored.

References

ACL Anthology. ACL Anthology. https://aclanthology.org/. Accessed: 15 June 2024.

Mouïn Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. 2024. Neco: Neural collapse based out-of-distribution detection. *Preprint*, arXiv:2310.06823.

Annual Reports. Annual Reports. https://www.annualreports.com/Browse/Industry. Accessed: 15 June 2024.

Sishuo Chen, Xiaohan Bi, Rundong Gao, and Xu Sun. 2022. Holistic sentence embeddings for better out-of-distribution detection. *Preprint*, arXiv:2210.07485.

Sishuo Chen, Wenkai Yang, Xiaohan Bi, and Xu Sun. 2023. Fine-tuning deteriorates general textual out-of-distribution detection by distorting task-agnostic features. *Preprint*, arXiv:2301.12715.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Companies House Service. Companies House Service. https://find-and-update.company-information.service.gov.uk/. Accessed: 15 June 2024.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. 2023. Extremely simple activation shaping for out-of-distribution detection.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. 2022. Vos: Learning what you don't know by virtual outlier synthesis. *Preprint*, arXiv:2202.01197.

D. Lewis et al. 2006. Building a test collection for complex document information processing.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. 2021. Exploring the limits of out-of-distribution detection. *Preprint*, arXiv:2106.03004.

Jiuxiang Gu, Yifei Ming, Yi Zhou, Jason Kuen, Vlad Morariu, Handong Zhao, Ruiyi Zhang, Nikolaos Barmpalios, Anqi Liu, Yixuan Li, Tong Sun, and Ani Nenkova. 2023. A critical analysis of document out-of-distribution detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4973–4999, Singapore. Association for Computational Linguistics.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. 2019. A benchmark for anomaly segmentation. *CoRR*, abs/1911.11132.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *CoRR*, abs/1610.02136.

Rui Huang, Andrew Geng, and Yixuan Li. 2021. On the importance of gradients for detecting distributional shifts in the wild. *CoRR*, abs/2110.00218.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. *Preprint*, arXiv:2204.08387.

Jayant Kumar, Peng Ye, and David Doermann. 2014. Structural similarity for document image classification and retrieval. *Pattern Recognition Letters*, 43:119–126. ICPR2012 Awarded Papers.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Preprint*, arXiv:1807.03888.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. 2021. Prototypical contrastive learning of unsupervised representations. *Preprint*, arXiv:2005.04966.

Shiyu Liang, Yixuan Li, and R. Srikant. 2017. Principled detection of out-of-distribution examples in neural networks. *CoRR*, abs/1706.02690.

Haowei Lin and Yuntian Gu. 2023. Flats: Principled out-of-distribution detection with feature-based likelihood ratio score. *Preprint*, arXiv:2310.05083.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *CoRR*, abs/2010.03759.

Haodong Lu, Dong Gong, Shuo Wang, Jason Xue, Lina Yao, and Kristen Moore. 2024. Learning with mixture of prototypes for out-of-distribution detection. *Preprint*, arXiv:2402.02653.

Yifei Ming, Haoyue Bai, Julian Katz-Samuels, and Yixuan Li. 2024. Hypo: Hyperspherical out-of-distribution generalization. *Preprint*, arXiv:2402.07785.

Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. 2023. How to exploit hyperspherical embeddings for out-of-distribution detection?

Vardan Papyan, X. Y. Han, and David L. Donoho. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663.

Resume Dataset. Resume Dataset. https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset. Accessed: 15 June 2024.

SEC EDGAR Database. SEC EDGAR Database. https://www.sec.gov/edgar/search/. Accessed: 15 June 2024.

Yiyou Sun, Chuan Guo, and Yixuan Li. 2021. React: Out-of-distribution detection with rectified activations. *CoRR*, abs/2111.12797.

Yiyou Sun and Yixuan Li. 2021. On the effectiveness of sparsification for detecting the deep unknowns. *CoRR*, abs/2111.09805.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. 2022. Out-of-distribution detection with deep nearest neighbors. *Preprint*, arXiv:2204.06507.

Leitian Tao, Xuefeng Du, Xiaojin Zhu, and Yixuan Li. 2023. Non-parametric outlier synthesis. *Preprint*, arXiv:2303.02966.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. 2022. Vim: Out-of-distribution with virtual-logit matching.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. 2024. Generalized out-of-distribution detection: A survey. *Preprint*, arXiv:2110.11334.

Haotian Ye, Chuanlong Xie, Tianle Cai, Ruichen Li, Zhenguo Li, and Liwei Wang. 2021. Towards a theoretical framework of out-of-distribution generalization. *Preprint*, arXiv:2106.04496.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2023. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415.

A Appendix

This section provides supplementary material in the form of dataset examples, implementation details, etc. to bolster the reader's understanding of the concepts presented in this work.

A.1 Proposed Methodology and Theoretical Framework

The central hypothesis underlying the proposed solution is predicated on the assumption that ID data should exhibit greater similarity in their feature representations when compared to OOD data. Consequently, we posit that when considering a pair of data points from two similar ID classes (denoted as Pair A) and a pair consisting of one ID and one OOD data point (denoted as Pair B), the application of a masking procedure on input features (whether textual or visual) would result in a more pronounced divergence in the feature space for Pair B as compared to Pair A. Initial experiments were conducted with random masking of input features. For textual data, this involved replacing tokens randomly with the '`[MASK]`' token. For visual data, random image patches were set to zero, effectively splitting the image into patches and nullifying selected segments. These preliminary experiments revealed two critical factors influencing the final feature embeddings used in distance-based OOD detection methods, such as the Mahalanobis distance: (a) the input tensors provided to the model, and (b) the feature extraction mechanism employed by the model, specifically the attention mechanism.

Although the early experiments primarily focused on input masking, achieving a consistent masking strategy proved challenging. While a consistent mask could be established for visual data by dividing images into uniformly sized chunks and consistently masking specific segments, such consistency was elusive for textual features. The variability in sequence length across different tokens complicated the masking process, often leading to strategies that involved masking padding tokens rather than meaningful data.

In light of these challenges, our focus shifted from input masking to the feature extraction process itself, particularly the attention mechanism within the model. We discovered that consistent masking could be achieved by selectively masking attention heads within different layers of the encoder. These heads are responsible for learning different representations and capturing different aspects of the input sequence. Hence by shutting down heads we are effectively deactivating certain pattern-extracting mechanisms within the attention architecture.

A.2 Description of Algorithm 1

As detailed in Algorithm 1, we begin with a fine-tuned model and proceed by randomly initializing various attention head masks based on a masking hyperparameter p. This hyperparameter represents the percentage of attention heads H set to zero within each attention layer N of the model. For each random mask, we extract dense hidden representations from both the training and evaluation datasets. The objective is to identify which of these randomly generated attention head masks minimizes the divergence between the representations of the evaluation and training data in the feature space. This is accomplished by calculating the average similarity score among the top K nearest neighbors for each evaluation data point.

The attention head masks are then ranked based on these aggregated similarity scores. Finally, we select the top F masks with the highest similarity scores between the evaluation and training data and use them to generate new feature representations. These features are then ensembled (i.e., averaged) and subsequently utilized in a distance-based OOD detection method, such as the Mahalanobis distance.

A.3 Hyperparameter Tuning

Table 2 summarizes the hyperparameters for model training. The model was trained using a carefully selected set of hyperparameters to optimize its performance. The training batch size per device was set to 32, while the evaluation batch size was configured at 8, ensuring efficient computation throughout the process. To stabilize updates, gradient accumulation was performed over 8 steps. The learning rate was set at 5×10^{-5}, with no weight decay applied, to prevent the risk of overfitting.

The Adam optimizer was configured with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and an epsilon value of 1×10^{-8} to ensure effective convergence. To maintain stability during training, the maximum gradient norm was capped at 1.0. The model underwent training for 65 epochs, with evaluations delayed by 5 steps to monitor progress at appropriate intervals, allowing for a well-tuned and stable learning process.

The hyperparameters chosen for the proposed

AHM method are presented in Table 3. Following the procedure outlined in Algorithm 1, an exploration budget of 25 was allocated for potential AHM configurations. To assess the effectiveness of different configurations, masking percentages of 0.1 and 0.2 were applied during the process.

To ensure robust performance, similarity scores between ID validation data and ID training data were computed. These scores were determined by averaging the similarity of the top 10 nearest neighbors for each validation data point. Using these similarity scores, the top five AHM heads were selected to generate the final representation embeddings, which were then combined through an ensemble approach to enhance the overall model performance.

Table 2: Hyperparameters for model training.

Hyperparameter	Value
per_device_train_batch_size	32
per_device_eval_batch_size	8
gradient_accumulation_steps	8
eval_delay	5
learning_rate	5e-05
weight_decay	0.0
adam_beta1	0.9
adam_beta2	0.999
adam_epsilon	1e-08
max_grad_norm	1.0
num_train_epochs	65

Table 3: Hyperparameters for AHM.

Hyperparameter	Value
Exploration budget (T)	25
Percentage masking (p)	[0.1, 0.2]
Neighbors (K)	10
Top AHM matrices select (F)	5

A.4 Annotator Training and Validation

To maintain high-quality annotation in line with ethical standards, we enlisted three postgraduate students fluent in English. They received instruction and participated in sessions with finance professionals to address any task-related questions. The annotation process spanned about four months, involving 90 training sessions, with breaks scheduled every 45 minutes. The students were compensated through gift vouchers and honorariums per minimum wage requirements[3].

A.5 Dataset description of FinanceDocs

The FinanceDocs dataset comprises a diverse collection of financial and legal documents sourced from various reliable platforms, offering a comprehensive view of corporate disclosures, shareholder communications, and regulatory filings. Each document type serves a distinct purpose, providing insights into different aspects of corporate governance, financial performance, and regulatory compliance, as detailed below:

- **SEC form documents:** These documents were collected from the Securities Exchange Commission (SEC) website. These forms are statements of changes in beneficial ownership.
- **Shareholder letter documents**: These documents were collected from annual reports. A shareholder letter in an annual report provides a summary of the company's financial performance, highlighting key achievements, strategic initiatives, and market conditions over the past year. It offers leadership's perspective on successes and challenges while outlining future goals and potential risks. The letter also emphasizes the company's commitment to corporate governance, social responsibility, and long-term growth.
- **SEC letter documents:** These documents were collected from the SEC website. These are letters from companies to the SEC about various company disclosures.
- **SEC-13 form documents:** These documents were collected from the SEC website. These forms disclose significant information about an entity's ownership or control over securities, typically required for investors with large holdings.
- **10k form documents:** These documents were collected from annual reports. These represent the 10k forms of an annual report
- **Financial info documents:** These documents were collected from annualreports (Annual Reports). They consist of various financial information, including the income statement,

[3] https://www.minimum-wage.org/international/united-states

balance sheet, and cash flow statement, which detail the company's revenue, expenses, assets, liabilities, and cash movements. It also includes financial ratios and metrics to assess profitability, liquidity, and leverage.

- **Articles of scientific paper documents:** These documents were collected from ACL Anthology[4]. It is a comprehensive digital archive of research papers in computational linguistics and natural language processing, published by the Association for Computational Linguistics.

- **Articles of resume documents:** These documents were collected from Kaggle. They represent resumes from different occupations.

- **Articles of Association documents:** These documents were collected from Companies House Services UK. They represent documents relating to articles of association of a company. These involve information such as directors powers and responsibilities, interpretation and limitation of liability as well as distribution of shares.

- **Director documents:** These documents were collected from annual reports and Companies House Services UK[5]. It involves information about the directors of a company.

[4] https://aclanthology.org/

[5] https://www.gov.uk/government/organisations/companies-house

A.6 Dataset examples of FinanceDocs

Presented below are examples from each document category included in FinanceDocs, providing the reader with a comprehensive visual overview of the dataset.

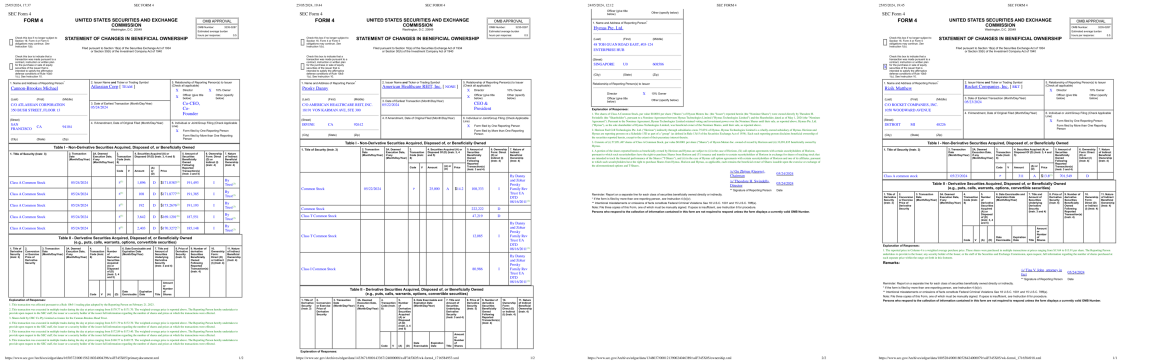


Figure 2: Examples of SEC form documents.

Figure 3: Examples of shareholder letter documents.

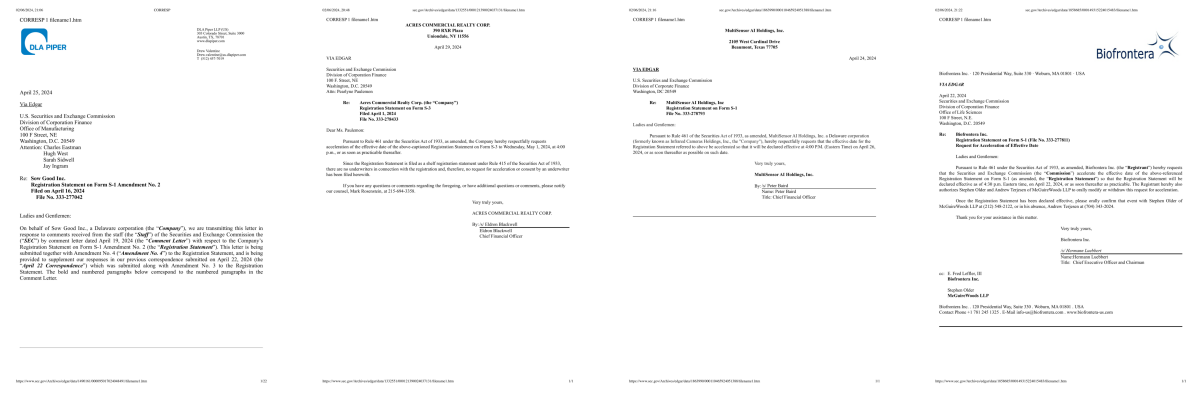


Figure 4: Examples of SEC letter documents.

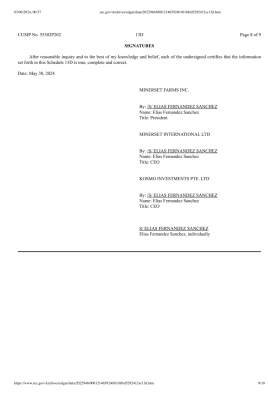
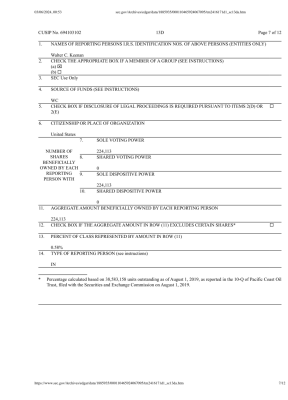
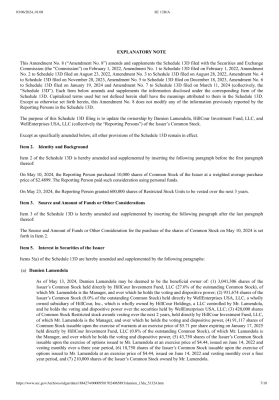
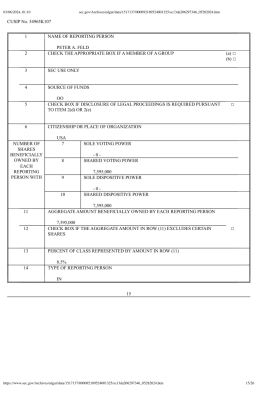
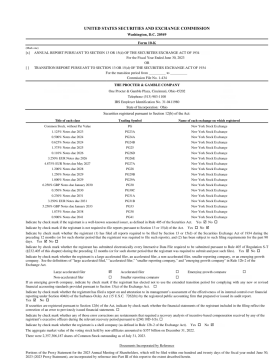
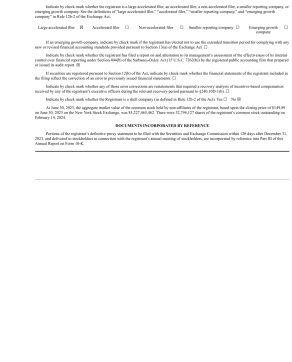





Figure 5: Examples of SEC-13 form documents.

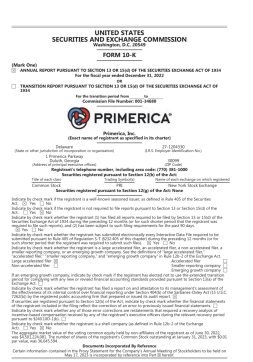


Figure 6: Examples of 10k form documents.

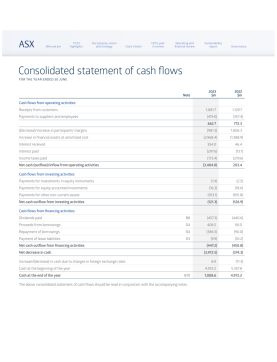
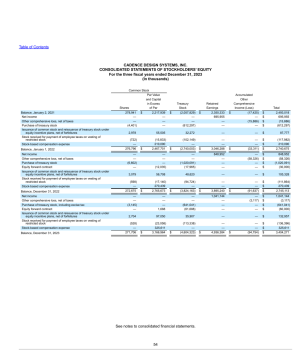
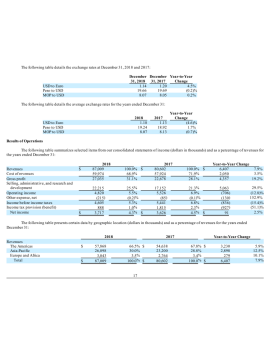
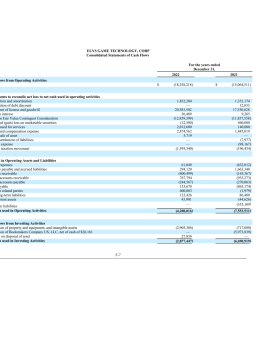





Figure 7: Examples of financial info documents.

Figure 8: Examples of scientific paper documents.

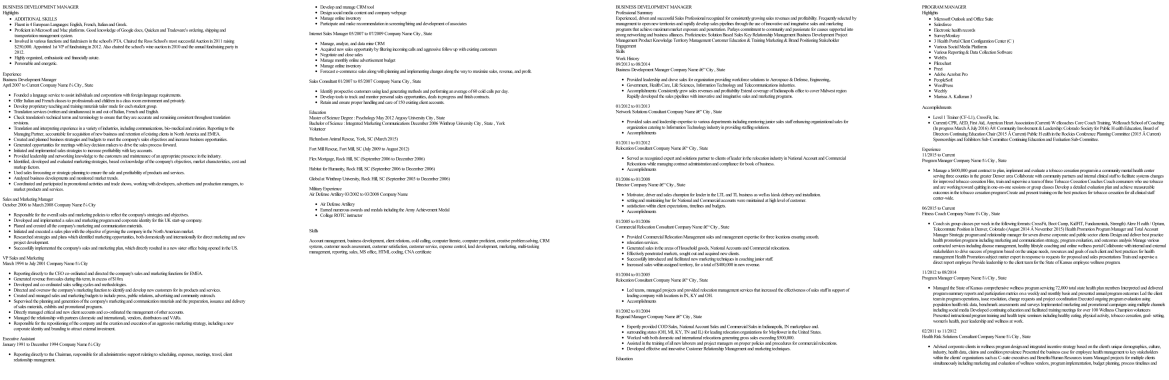


Figure 9: Examples of resume documents.

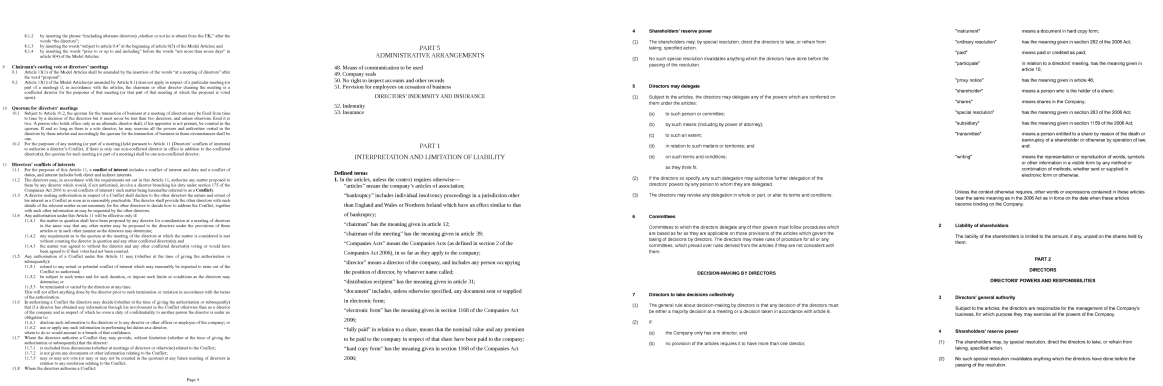
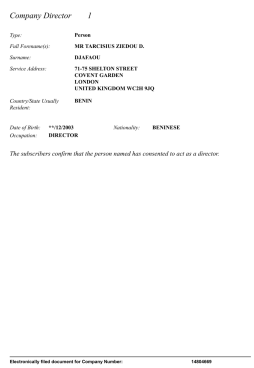


Figure 10: Examples of Articles of Association documents.



Company Director 1

Type: **Person**

Full Forename(s): **MR TARCISIUS ZIEDOU D.**

Surname: **DJAFAOU**

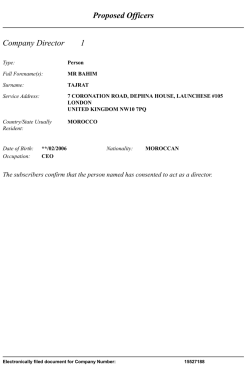
Service Address: **71-75 SHELTON STREET COVENT GARDEN LONDON UNITED KINGDOM WC2H 9JQ**

Country/State Usually Resident: **BENIN**

Date of Birth: ****/12/2003** *Nationality:* **BENINESE**
Occupation: **DIRECTOR**

The subscribers confirm that the person named has consented to act as a director.

Electronically filed document for Company Number: 14884669



Proposed Officers

Company Director 1

Type: **Person**

Full Forename(s): **MR BAHIM**

Surname: **TAJRAT**

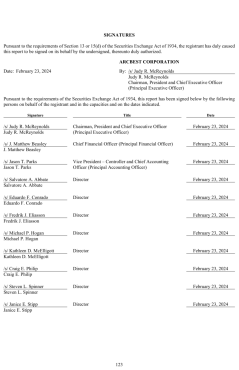
Service Address: **7 CORONATION ROAD, DEPHNA HOUSE, LAUNCHESE #105 LONDON UNITED KINGDOM NW10 7PQ**

Country/State Usually Resident: **MOROCCO**

Date of Birth: ****/02/2006** *Nationality:* **MOROCCAN**
Occupation: **CEO**

The subscribers confirm that the person named has consented to act as a director.

Electronically filed document for Company Number: 15527185



SIGNATURES

Pursuant to the requirements of Section 13 or 15(d) of the Securities Exchange Act of 1934, the registrant has duly caused this report to be signed on its behalf by the undersigned, thereunto duly authorized.

ARCREST CORPORATION

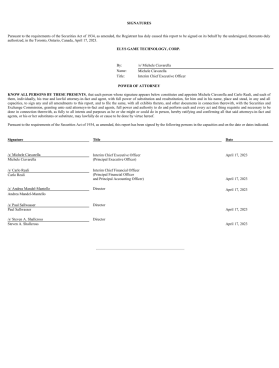
Date: February 23, 2024

By: /s/ Judy R. McReynolds
Judy R. McReynolds
Chairman, President and Chief Executive Officer
(Principal Executive Officer)

Pursuant to the requirements of the Securities Exchange Act of 1934, this report has been signed below by the following persons on behalf of the registrant and in the capacities and on the dates indicated.

Signature	Title	Date
/s/ Judy R. McReynolds Judy R. McReynolds	Chairman, President and Chief Executive Officer (Principal Executive Officer)	February 23, 2024
/s/ J. Matthew Beasley J. Matthew Beasley	Chief Financial Officer (Principal Financial Officer)	February 23, 2024
/s/ Jason T. Parks Jason T. Parks	Vice President - Controller and Chief Accounting Officer (Principal Accounting Officer)	February 23, 2024
/s/ Salvatore A. Abbate Salvatore A. Abbate	Director	February 23, 2024
/s/ Eduardo F. Conrado Eduardo F. Conrado	Director	February 23, 2024
/s/ Fredrik J. Eliasson Fredrik J. Eliasson	Director	February 23, 2024
/s/ Michael P. Hogan Michael P. Hogan	Director	February 23, 2024
/s/ Kathleen D. McElligott Kathleen D. McElligott	Director	February 23, 2024
/s/ Craig E. Philip Craig E. Philip	Director	February 23, 2024
/s/ Steven L. Spinner Steven L. Spinner	Director	February 23, 2024
/s/ Janice E. Stipp Janice E. Stipp	Director	February 23, 2024

123



SIGNATURES

Pursuant to the requirements of the Securities Act of 1934, as amended, the Registrant has duly caused this report to be signed on its behalf by the undersigned, thereunto duly authorized, in the Toronto, Ontario, Canada, April 17, 2023.

ELYS GAME TECHNOLOGY, CORP.

By: /s/ Michele Ciavarella
Name: Michele Ciavarella
Title: Interim Chief Executive Officer

POWER OF ATTORNEY

KNOW ALL PERSONS BY THESE PRESENTS, that each person whose signature appears below constitutes and appoints Michele Ciavarella and Carlo Rindi, and each of them, individually, his true and lawful attorneys-in-fact and agents, with full power of substitution and resubstitution, for him and in his name, place and stead, in any and all capacities, to sign any and all amendments to this report, and to file the same, with all exhibits thereto, and other documents in connection therewith, with the Securities and Exchange Commission, granting unto said attorneys-in-fact and agents, and each of them, full power and authority to do and perform each and every act and thing requisite and necessary to be done in connection therewith, as fully to all intents and purposes as he or she might or could do in person, hereby ratifying and confirming all that said attorneys-in-fact and agents or any of them or their substitute or substitutes may lawfully do or cause to be done by virtue hereof.

Pursuant to the requirements of the Securities Act of 1934, as amended, this report has been signed by the following persons in the capacities and on the date or dates indicated.

Signature	Title	Date
/s/ Michele Ciavarella Michele Ciavarella	Interim Chief Executive Officer (Principal Executive Officer)	April 17, 2023
/s/ Carlo Rindi Carlo Rindi	Interim Chief Financial Officer (Principal Financial Officer and Principal Accounting Officer)	April 17, 2023
/s/ Andrea Mandel-Mantello Andrea Mandel-Mantello	Director	April 17, 2023
/s/ Paul Salfenauer Paul Salfenauer	Director	April 17, 2023
/s/ Steven A. Mathyer Steven A. Mathyer	Director	April 17, 2023

Figure 11: Examples of list of director documents.