

1 **Uncovering the effects of model initialization on deep model**
2 **generalization: A study with adult and pediatric chest X-ray**
3 **images**

4

5

6 Sivaramakrishnan Rajaraman, Ghada Zamzmi#, Feng Yang, Zhaohui Liang, Zhiyun Xue, Sameer Antani*

7

8

9 Computational Health Research Branch, National Library of Medicine, National Institutes of Health,

10 Maryland, USA

11

12

13 * Corresponding author

14 E-mail: sameer.antani@nih.gov

15

16

17 # Note: Dr. Ghada Zamzmi is currently a Staff Fellow at the Center for Devices and Radiological Health

18 at the Food and Drug Administration.

19

20

21

22 **Abstract**

23 Model initialization techniques are vital for improving the performance and reliability of deep
24 learning models in medical computer vision applications. While much literature exists on non-medical
25 images, the impacts on medical images, particularly chest X-rays (CXRs) are less understood. Addressing
26 this gap, our study explores three deep model initialization techniques: Cold-start, Warm-start, and Shrink
27 and Perturb start, focusing on adult and pediatric populations. We specifically focus on scenarios with
28 periodically arriving data for training, thereby embracing the real-world scenarios of ongoing data influx
29 and the need for model updates. We evaluate these models for generalizability against external adult and
30 pediatric CXR datasets. We also propose novel ensemble methods: F-score-weighted Sequential Least-
31 Squares Quadratic Programming (F-SLSQP) and Attention-Guided Ensembles with Learnable Fuzzy
32 Softmax to aggregate weight parameters from multiple models to capitalize on their collective knowledge
33 and complementary representations. We perform statistical significance tests with 95% confidence intervals
34 and p -values to analyze model performance. Our evaluations indicate models initialized with ImageNet-
35 pretrained weights demonstrate superior generalizability over randomly-initialized counterparts,
36 contradicting some findings for non-medical images. Notably, ImageNet-pretrained models exhibit
37 consistent performance during internal and external testing across different training scenarios. Weight-level
38 ensembles of these models show significantly higher recall ($p < 0.05$) during testing compared to individual
39 models. Thus, our study accentuates the benefits of ImageNet-pretrained weight initialization, especially
40 when used with weight-level ensembles, for creating robust and generalizable deep learning solutions.

41

42 **Author Summary**

43 In this research, we actively explore various techniques for optimal initialization of deep learning
44 models for analyzing medical images such that the resulting models are generalizable and also demonstrate
45 high performance. Generalizability is an area of significant importance. It is often ignored in favor of the

46 model achieving high performance at the cost of maintaining on previously unseen, i.e. external data, that
47 may also be out-of-distribution. This may result in the classifier performing inadequately thereby reducing
48 its value. We demonstrate that unlike general-purpose images, such as those found in ImageNet collection,
49 medical images, such as chest X-rays (CXRs), are different in their visual characteristics. We show that
50 counter to previously reported results using non-medical images, ImageNet pre-trained models trained on
51 medical images, in fact, converge sooner and generalize better than randomly initialized models. We
52 compare three distinct model initialization methods using internal adult CXR data to train the models which
53 are subsequently tested on external CXR images for both adult and pediatric populations. Additionally, we
54 consolidate several of these models into an 'ensemble' to demonstrate that they achieve a more accurate
55 identification of relevant cases during both internal and external testing. Therefore, our work underscores
56 the promising potential of employing ImageNet-pretrained models for medical images and merging them
57 into ensembles, aiming to enhance the reliability of AI in medical image analysis.

58

59 **Introduction**

60 The prowess of Deep learning (DL) has been well established for medical imaging artificial
61 intelligence (AI) applications with automation making way for improved and efficient image acquisition,
62 quality assessment, object detection and tracking, disease screening, diagnostics, and prediction [1]. As a
63 subset of machine learning (ML), DL comprises multilayered neural networks for automated feature
64 extraction and predictions, outperforming traditional techniques in accuracy and robustness.

65 Chest X-rays (CXRs) are a routinely used diagnostic imaging modality. Despite lower sensitivity
66 compared to computed tomography (CT) scans, CXRs offer several advantages, including cost-
67 effectiveness, reduced radiation exposure, and accessibility, making them practical in resource-limited
68 settings [2] [3]. Several CXR datasets are available to the ML community which has resulted in significant
69 advances in disease detection [4–8]. This dataset listing is not intended to be exhaustive as new datasets are
70 being made available with higher frequency.

71 A key step in developing high-performing DL solutions is determining appropriate model
72 initialization strategies [9]. Model initialization refers to the method of assigning initial values to neural
73 network weights and biases. Optimal selection of the initialization strategy depends on various factors such
74 as data characteristics including dimensionality, variability due to differences in patient anatomy, disease
75 states, image acquisition procedures, and requirement for expert interpretation among others, activation
76 functions, and optimization algorithms selected in the design [10]. Understanding the intricacies of model
77 initialization and its impact on performance is essential for devising effective training methodologies, and
78 addressing various issues in the training process, including vanishing or exploding gradients, achieving
79 faster convergence, and stable training dynamics. An appropriately selected initialization strategy can also
80 result in reliable and enhanced medical AI performance which is crucial for precision medicine applications.

81 The significance of model initialization is amplified when we consider challenges in model
82 generalization which are primarily due to feature distribution shifts between training datasets and real-world
83 use. For example, a model trained and tested on adult CXR data from the same source (*internal testing*)
84 may result in significantly higher performance compared to testing it on adult CXR data from another
85 source (*external testing*) [11]. Additional performance degradation may be observed when pediatric images
86 exhibiting the same disease(s) are included in the testing. The inherent high-dimensional complexity and
87 variability of medical images exacerbate this problem, causing models to overfit the training data. In this
88 work, we present findings from our investigations on the impact of different model initialization techniques
89 on DL models and propose mechanisms to improve generalizability.

90 A review of DL literature on model initialization reveals two main techniques, namely, *cold-start*
91 and *warm-start*, each with distinct implications for model training dynamics, generalizability, and
92 performance [10]. The cold-start method initializes new weights and biases with small random values which
93 results in training a new model from scratch. This technique offers an unbiased foundation but deprives the
94 model of initialization guidance thereby resulting in slower convergence. Conversely, the warm-start
95 strategy leverages weights and biases from a model that has been previously trained on data from similar
96 content. Initialization guidance offered using this approach enables faster model convergence and also

97 provides potentially enhanced performance. However, a previous study [10] has conversely reported that
98 warm-start consistently underperforms with non-medical images, yielding models with poorer
99 generalization and lower prediction accuracy compared to cold-start models. The *Shrink and Perturb*
100 method proposed in [10] shrinks existing model weights towards zero and adds noise, resulting in faster
101 training than cold-start and improved generalization over warm-start models. However, that and other
102 studies focused on non-medical images [9,10,12–15] left a gap in understanding the impact of model
103 initialization techniques on medical computer vision. Unlike non-medical images, medical images have
104 unique characteristics including (i) variations in imaging modalities, e.g., CT, MRI, ultrasound, X-ray,
105 pathology, endoscopy, where each modality captures different aspects of the human body at varying levels
106 of resolution, contrast, and noise levels; (ii) image acquisition conditions including patient positioning,
107 imaging protocols, and the expertise of medical professionals during acquisition that impacts the quality
108 and appearance; (iii) varying anatomical structures that depict internal organs, tissues, and systems, and
109 physiological processes that provide vital information for the diagnosis, treatment planning, and monitoring
110 of diseases, (iv) limited and imbalanced data where instances of specific diseases or conditions with varying
111 levels of progression are significantly smaller compared to healthy cases, and (v) ethical and regulatory
112 considerations in handling medical data since they involve sensitive patient information, thereby ensuring
113 the confidentiality and other critical factors [16,17].

114 Model generalizability is defined as the ability of a trained model to capture generalized patterns
115 and perform well on unseen data. Medical computer vision relies on model generalizability for several
116 reasons [11] including accommodating patient diversity, adapting to various data sources and quality,
117 addressing ethical considerations, and enhancing clinical utility. A general model is robust to different data
118 sources and population distributions, considering factors such as the patient/study subject’s ethnicity, sex,
119 and severity of the disease(s) expressed on the image. Further, in many ML applications, data continuously
120 flows into the system which may require regular model updates and may be unreasonable or difficult to
121 implement. Therefore, developing reliable and generalizable models mandates both internal and
122 external/out-of-distribution testing [18].

123 Most of the literature has focused on assessing internal generalization due to the lack of widely
124 available data sets [5,19–21] and the findings, though significant, may not guarantee optimal model
125 performance with external data. Federated learning methods have been proposed that use decentralized
126 training to address challenges in achieving external generalization by incorporating diverse data
127 distributions [22]. This approach could mitigate the risk of performance degradation when the model
128 encounters unseen data distributions. However, this approach has its limitations, such as requiring
129 consistent communication and synchronization between data sources, which can be challenging in real-
130 world settings with privacy concerns or network instability [23]. Further, there could be data
131 interoperability and completeness issues that limit generalization gains. Therefore, while federated learning
132 provides a path toward achieving external generalization, it also introduces new challenges. This presents
133 us with an opportunity for considering and evaluating other novel and efficient methods for achieving
134 external generalization.

135 For this work, we use adult and pediatric CXRs to evaluate model generalizability as they
136 simultaneously exhibit significant similarities and differences due to anatomy and disease presentation
137 across age groups [24]. These include: (i) Developmental stages: Evolving thoracic anatomy in pediatric
138 patients is distinct in appearance from adults. There are thinner chest walls and more compliant rib cages
139 in children. (ii) Unique abnormalities: Pediatric disease can present differently than adults or similar
140 presentations could indicate different diseases. (iii) Imaging technique: Distinct protocols for pediatric
141 CXRs can result in variations in intensity and contrast. Further, inspiration may be inconsistent across
142 patients. (iv) Patient pose: Pediatric patients may need to be held down resulting in the presence of other
143 hands in the image and unusual and variable pose of the patient. These discrepancies present challenges for
144 DL models trained on adult data when directly applied to pediatric cases, potentially leading to sub-optimal
145 generalizability, and reduced clinical utility. Prior work in pediatric CXR image analysis includes the
146 development and evaluation of a ResNet-50 model trained to classify pediatric CXRs as showing
147 pneumonia-consistent manifestations or normal lungs [11]. The model demonstrated comparatively
148 improved performance on the internal test set (area under the curve (AUC): 0.95) compared to the external

149 NIH-CXR test set (AUC: 0.54), highlighting potential limitations in model generalizability. There is limited
150 literature analyzing the generalizability of deep models trained on adult CXRs to the pediatric population.

151 Our study presents key contributions to address the knowledge gap in the current literature
152 regarding the impact of model initialization methods on the generalizability of DL models when we apply
153 them to external adult and pediatric populations after training on internal adult CXR data. We specifically
154 focus on scenarios with periodically arriving data for training, which is a common challenge faced by
155 medical computer vision algorithms. Our investigation delves into the performance of widely-used model
156 initialization methods, providing insights into their adaptability and their implications on generalizability.
157 Furthermore, we propose novel weight-level ensemble methods to improve model generalizability. This
158 crucial understanding will pave the way for the successful deployment of DL models in medical imaging
159 applications, ultimately improving clinical decision-making and patient outcomes.

160

161 **Materials and methods**

162 **Datasets**

163 This retrospective study utilizes the following datasets:

164 (i) RSNA-CXR dataset: This publicly available CXR collection results from a collaboration
165 between the RSNA, the Society of Thoracic Radiology (STR), and the National Institutes of Health (NIH)
166 for the Kaggle pneumonia detection challenge [25]. The objective was to help support the design and
167 development of image analysis and ML algorithms through a challenge targeting automatic classification
168 of CXRs as normal, containing non-pneumonia-related, or pneumonia-related opacities. The collection
169 comprises 26,684 deidentified anterior-posterior (AP) and posterior-anterior (PA) CXRs in DICOM format,
170 featuring 8,851 normal lungs and 17,833 other abnormal radiographic patterns, of which 6,012 manifest
171 pneumonia-related opacities. We use this dataset to train, validate, and internally test the DL model.

172 (ii) Indiana-CXR dataset: The Indiana CXR dataset contains 7,470 frontal and lateral CXR
173 projections [26] in DICOM format, accompanied by multiple annotations, including indications, findings,
174 and impressions in textual form. These images are sourced from hospitals affiliated with the Indiana
175 University School of Medicine. Among these, 2,378 PA CXRs exhibit abnormal pulmonary manifestations,
176 and 1,726 CXRs have normal lung appearances. This de-identified dataset is stored at the National Library
177 of Medicine (NLM) and has been exempted from Institutional Review Board review (OHSRP # 5357). We
178 use this dataset as the external adult test set.

179 (iii) VINDR-PCXR dataset: The VINDR-PCXR dataset is a publicly available pediatric CXR
180 collection [27] developed to support computer-aided diagnosis algorithm development for pediatric CXR
181 interpretation. It consists of 9,125 CXR scans, in DICOM format, collected from three major Vietnamese
182 hospitals between 2020 and 2021. The pediatric dataset includes deidentified images of 5,354 males, 3,709
183 females, and 62 patients with unknown gender. Among the 8,755 pediatric CXRs, 5,876 show normal lungs,
184 and 2,879 exhibit other cardiopulmonary abnormalities, with age distributions as follows: 5,335 CXRs for
185 ages 1 day to under 24 months, 3,351 CXRs for ages 24 months to under 11 years, and 69 CXRs for ages
186 11 to under 18 years. We use this dataset as an external pediatric test.

187 (iv) NIH-CXR dataset: The NIH-CXR dataset is a publicly accessible, large-scale collection of
188 deidentified CXRs [28] compiled by the NIH Clinical Center. It contains 112,120 frontal-view CXR images
189 in PNG format, from 30,805 unique patients. The dataset includes 14 cardiopulmonary disease labels, text-
190 mined from radiological reports using a Natural Language Processing (NLP) labeler. Among these, 5,257
191 pediatric CXRs represent normal lungs ($n = 3,066$) and other cardiopulmonary abnormalities ($n = 2,191$),
192 divided into three age groups: 34 CXRs captured from pediatric patients of ages 1 day to under 24 months,
193 1,787 CXRs of ages 24 months to under 11 years, and 3,486 CXRs of ages 11 to under 18 years. The
194 pediatric group consists of 3,018 males and 2,239 females, while 106,863 CXRs belong to patients older
195 than 18 years. We use this dataset as the external pediatric test.

196 We further partition the RSNA-CXR dataset at the patient level into 70% for training, 10% for
197 validation, and 20% for internal testing. The training and validation sets are additionally divided into two

198 equal-sized subsets to simulate periodic data arrival for training and validation and facilitate the simplest
 199 case of warm-start. The DL model trains to converge on the first half of the data and then trains on the full
 200 collection, which represents 100% of the data. We name the first half *RSNA-Partial (P)* and the full
 201 collection *RSNA-Full (F)*. The internal test remains the same for both the RSNA-P and RSNA-F datasets.
 202 Table 1 provides details of this partition.

203

204 **Table 1. Training, validation, and internal test split using the RSNA-CXR dataset.**

Dataset	Train		Val		Internal test	
	No finding	Abnormal	No finding	Abnormal	No finding	Abnormal
RSNA-P	3098	6242	442	891	1770	3566
RSNA-F	6196	12484	885	1783	1770	3566

205

206 The external test sets consist of adult CXRs from the Indiana-CXR collection and pediatric CXRs from the
 207 NIH-CXR and VINDR-PCXR collections. We categorize the pediatric CXRs into three groups: Ped-2 (1
 208 day to under 24 months), Ped-11 (24 months to under 11 years), and Ped-18 (11 years to under 18 years),
 209 based on the lung developmental stages from infancy to adulthood as discussed in [29]. Table 2 shows the
 210 categorization of test CXRs according to various age groups.

211

212 **Table 2. External test set categorization across various age groups.**

Dataset	1 day to < 24 months		24 months to < 11 years		11 years to < 18 years		> 18 years	
	No finding	Abnormal	No finding	Abnormal	No finding	Abnormal	No finding	Abnormal
Indiana-CXR	-	-	-	-	-	-	1726	2378
NIH-CXR	29	5	1059	728	1978	1458	-	-
VINDR-PCXR	3341	1994	2475	876	60	9	-	-
Total	3370	1999	3534	1604	2038	1467	1726	2378

213

214 Lung region delineation and cropping

215 We utilize a UNet [30] model with an ImageNet-pretrained Inception-V3 encoder backbone from
 216 our previous study [31] to delineate the lung regions and crop them to the size of a bounding box. The
 217 purpose of lung cropping is to prevent the DL model from learning irrelevant features for cardiopulmonary
 218 disease detection. We resize the cropped lung bounding boxes to 256×256 -pixel dimensions and normalize
 219 them to the range $[0, 1]$ to reduce computational complexity.

220

221 **Model architecture and training**

222 For the model architecture and training scenario, we employ a VGG-16 model [32] architecture.
 223 We truncate it at its deepest pooling layer and append a global average pooling (GAP) layer and a final
 224 dense layer with two nodes and Softmax activation. This modified model, referred to as *VGG-16-M*,
 225 predicts whether the CXRs show normal lungs or other cardiopulmonary abnormalities. We choose the
 226 VGG-16 model due to its simplicity, effectiveness, and well-known performance in medical image
 227 classification tasks, particularly using CXRs [33–35]. Selecting an optimal model falls beyond the scope of
 228 this research, as our study aims to analyze the impact of model initialization strategies on deep model
 229 generalization. The proposed technique can be applied to any model suitable for the characteristics of the
 230 data under study. Table 3 provides a list of the data and model terminologies used in this study.

231

232 **Table 3. Data and model terminologies.**

Terminologies	Explanation
R, I	Model initialization: random weights (R) or ImageNet-pretrained weights (I)
P, F	VGG-16-M model dataset usage: RSNA-P (P) or RSNA-F (F)
Cold-RP	Random initialization, trained on RSNA-P
Cold-IP	ImageNet-pretrained initialization, trained on RSNA-P
Cold-RF	Random initialization, trained on RSNA-F
Warm-RF	Cold-RP model fine-tuned on RSNA-F
Shrink-RF	Cold-RP model with weights shrunk by factor αl , found via Bayesian search
Cold-IF	ImageNet-pretrained initialization, trained on RSNA-F

Warm-IF	Cold-IP model fine-tuned on RSNA-F
Shrink-IF	Cold-IP model with weights shrunk by factor α , found via Bayesian search

233

234

235

236

237

238

239

240

241 **Optimizing the weight-scaling factor**

242

243

244

245

Each model undergoes training and validation using the RSNA-CXR dataset with a mini-batch size of 64. We utilize the Adam optimizer with an initial learning rate of 0.001 to minimize the categorical cross-entropy loss. Model checkpoints are stored via callbacks when a decrease in validation loss is observed. The checkpoint exhibiting the lowest validation loss is used to generate predictions for both the internal and external test datasets. Test performance evaluation occurs at the ideal classification threshold, determined by maximizing the F-score for the validation dataset.

The method proposed in the Shrink and Perturb technique [10] involves shrinking the existing model weights by multiplying with a factor α and incorporating a small noise β to accelerate DL model convergence and enhance generalization compared to standard cold-start and warm-start methods. Let W be the set of model weights. We calculate the updated weights W' using Equation (1):

$$W' = \alpha W + \beta. \tag{1}$$

246

247

248

249

250

251

252

253

254

Here, α denotes the weight-scaling factor. Previous experiments [10] used discrete α values and fixed β at 0.01. In contrast, while we continue to use a fixed value for β as 0.01, we apply Bayesian optimization via Gaussian Process (GP) minimization [36] to identify the optimal α for shrinking the weights of the Cold-RP and Cold-IP models. These are subsequently used to initialize the weights in the Shrink-RF and Shrink-IF models, respectively. Bayesian optimization using GP minimization reduces susceptibility to local minima, enabling more effective identification of the optimal α within a continuous interval compared to the grid or random search methods at discrete intervals. GP minimization explores the search space more thoroughly and converges efficiently by modeling the objective function as a Gaussian process sample. We define the continuous search space for α within the range [0.1, 0.9]. We create a function that accepts α as

255 input and performs the following steps: (i) instantiate and compile the model with the current weights, (ii)
256 train and validate the model, storing the best model weights, validation loss, α , and training history
257 whenever the validation loss decreases, and (iii) perform GP minimization for 100 function calls and 30
258 random starts to converge to the optimal α with minimal validation loss. The hyperparameters for GP
259 minimization follow the default settings in the scikit-optimize Python library.

260

261 **Weight-level ensembles**

262 We are also proposing ensemble methods that merge the weights of multiple models. Our approach
263 is different from traditional techniques that aggregate model predictions [37–39]. Our proposed *weight-*
264 *level ensembles* harness the power of diverse weight initializations, capitalizing on complementary learning
265 dynamics to foster robust generalization in complex, high-dimensional medical data landscapes.

266 We perform Equal Weight Averaging (EWA), which combines the weights of multiple trained
267 models to create an average model. This technique aims to enhance classification performance by
268 leveraging the complementary strengths of individual models in capturing data patterns. We achieve this
269 by iterating through each model's layers, retrieving and averaging the layer weights with equal weight
270 factors, resulting in a new model with a similar architecture for prediction.

271 We introduce a novel F-score-weighted Sequential Least-Squares Quadratic Programming (F-
272 SLSQP)-based weighted ensemble method to determine the optimal multiplication factors for combining
273 the weights of multiple models in the ensemble. We identify these optimal factors by minimizing the error,
274 as defined in Equation (2), through SLSQP-based constrained minimization [40].

$$Error = 1 - (F - score_{validation}). \quad (2)$$

275 The process of determining the optimal multiplication factors involves the following steps: (i) defining a
276 function to compute the weighted average of weights for the ensemble models, (ii) defining a function to
277 create a new model with the same architecture as the models in the ensemble, (iii) creating a global variable
278 for the best multiplication factors, (iv) defining a function to calculate the error from the weighted average

279 of the models, (v) setting the optimization parameters, including the constraints and bounds, where the
280 constraint ensures the sum of scaling factors equals 1.0 and the bounds ensure each scaling factor is within
281 the range [0, 1], (vi) executing the SLSQP algorithm multiple times ($n = 100$) to minimize the error and
282 find the optimal multiplication factors, (vii) performing weighted averaging with the optimal multiplication
283 factors to create the weighted ensemble model, and (viii) compiling and saving the weighted ensemble
284 model for prediction.

285 Additionally, we present a novel method for developing an attention-guided ensemble
286 incorporating a learnable Fuzzy Softmax layer (AGELFS). This technique utilizes attention mechanisms
287 [41] to emphasize relevant features of each model while mitigating less significant ones. The ensemble
288 construction involves the following steps: (i) instantiating and freezing the constituent models with their
289 respective weights, (ii) processing training input through these models and appending a GAP layer to each
290 model's output, (iii) concatenating the outputs of the GAP layers, (iv) introducing an attention layer to
291 derive attention-based weights for the concatenated outputs, (v) appending a dense layer with conventional
292 Softmax activation to learn attention-based weights for the concatenated outputs, (vi) applying a learnable
293 Fuzzy Softmax (LFS) layer to the dense layer output, and (vii) training the ensemble. The Fuzzy Softmax
294 layer [42] enhances the conventional Softmax function by introducing a learnable Fuzziness parameter that
295 controls the uncertainty level in output probabilities, as described in Equation (3), where x_i and x_j
296 represent the input logits, and *Fuzziness* is the learnable parameter.

$$(Learnable\ Fuzzy\ Softmax(x_i) = \exp(fuzziness * x_i) / \sum(\exp(fuzziness * x_j))). \quad (3)$$

297

298 **Performance evaluation and statistical significance analysis**

299 We examine model performance using key metrics, including balanced accuracy, precision, recall,
300 the area under the precision-recall curve (AUPRC), F-score, and Matthews Correlation Coefficient (MCC).
301 Each metric provides valuable insights into the model's effectiveness in various aspects of classification
302 tasks. We present the statistical significance of the MCC by utilizing 95% binomial confidence intervals

303 (CIs) and ascertain them through the Clopper-Pearson Exact methodology to distinguish model efficacy.
 304 We determine the p -values based on the CI-based Z-test [43]. We obtain the MCC values and their
 305 corresponding 95% CIs for the compared models. For each model, we compute the standard error (SE)
 306 using Equation (4):

$$SE = (CI_{upper} - CI_{lower}) / (2 * 1.96). \quad (4)$$

307 Here, CI_{upper} and CI_{lower} represent the upper and lower bounds of the CIs, respectively. We compute the
 308 difference in the MCC (ΔMCC) and SE (ΔSE) values using Equations (5) and (6) respectively:

$$\Delta MCC = MCC2 - MCC1. \quad (5)$$

$$\Delta SE = \text{sqrt}(SE1^2 + SE2^2). \quad (6)$$

309 Here, $MCC1$, $MCC2$, $SE1$, and $SE2$ are the MCC and SE metrics of the compared models. We compute the
 310 Z-score from this difference using Equation (7):

$$Z = \Delta MCC / \Delta SE. \quad (7)$$

311 We calculate the corresponding p -value for the Z-score using an online Z-table. A threshold of 0.05 is
 312 utilized to establish statistical significance using the 95% CIs. If the p -value is less than 0.05, we observe
 313 that the difference in performance, as gauged by MCC, is statistically significant. We repeat this process to
 314 present the statistical significance of the recall values for the proposed weight-level ensembles.

315

316 **Results and discussion**

317 We first present a comparative analysis between the performances of the Cold-RP and Cold-IP
 318 models. Recall that the Cold-RP model initializes the VGG-16 backbone of the VGG-16-M model with
 319 random weights and trains it on the RSNA-P dataset. Conversely, the Cold-IP model initializes the VGG-
 320 16 backbone of the VGG-16-M model with ImageNet-pretrained weights and also trains it on the RSNA-P
 321 dataset.

322 Table 4 displays performance metrics when predicting the RSNA-P test set (internal adult test set),
 323 while Fig 1 illustrates the AUPRC, confusion matrices, and a comparison of MCC values. Based on the
 324 information in Table 4, we deduce the following: (i) The Cold-IP model converges considerably faster than
 325 the Cold-RP model, and (ii) the Cold-IP model exhibits a significantly higher MCC ($p < 0.00001$) and
 326 notably higher values for other performance metrics compared to the Cold-RP model.

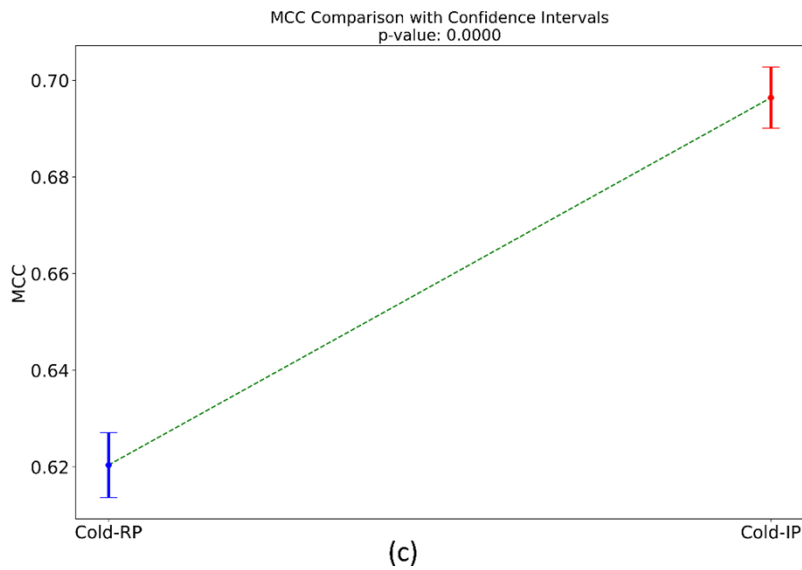
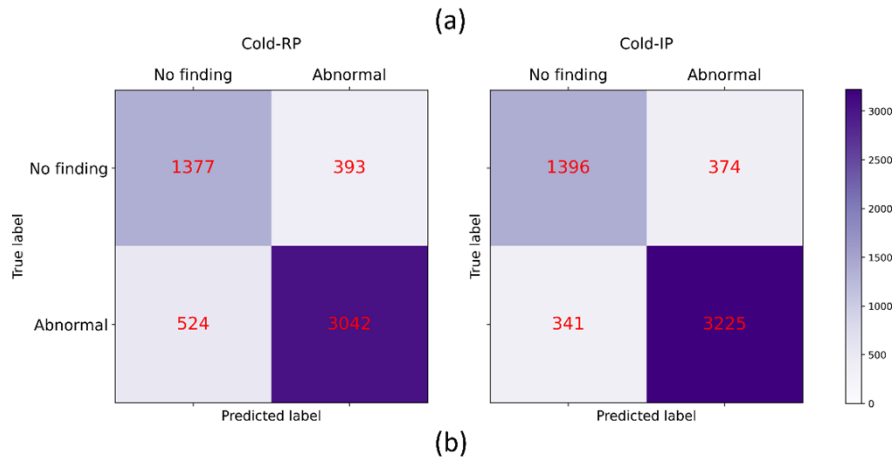
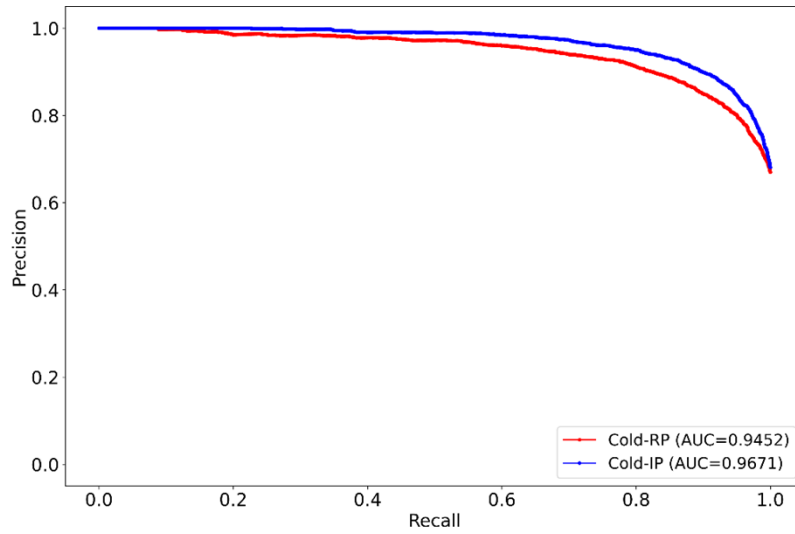
327

328 **Table 4. Performance of models initialized with random and ImageNet-pretrained weights on the**
 329 **internal adult test set.** The terms B. Acc., P, R, and F denote balanced accuracy, precision, recall, and F-
 330 score, respectively. Bold numerical values denote superior performance in respective columns. Values in
 331 parentheses represent the 95% CIs for the MCC metric. The * denotes statistically significant MCC
 332 ($p < 0.00001$).

Model	AUPRC	B. Acc.	P	R	F	MCC	Training time (in sec.)	p -MCC
Cold-RP	0.9452	0.8156	0.8856	0.8531	0.8690	0.6204 (0.6073, 0.6335)	2052.83	<0.00001
Cold-IP	0.9671	0.8466	0.8961	0.9044	0.9002	0.6964 (0.6840, 0.7088)*	812.52	

333

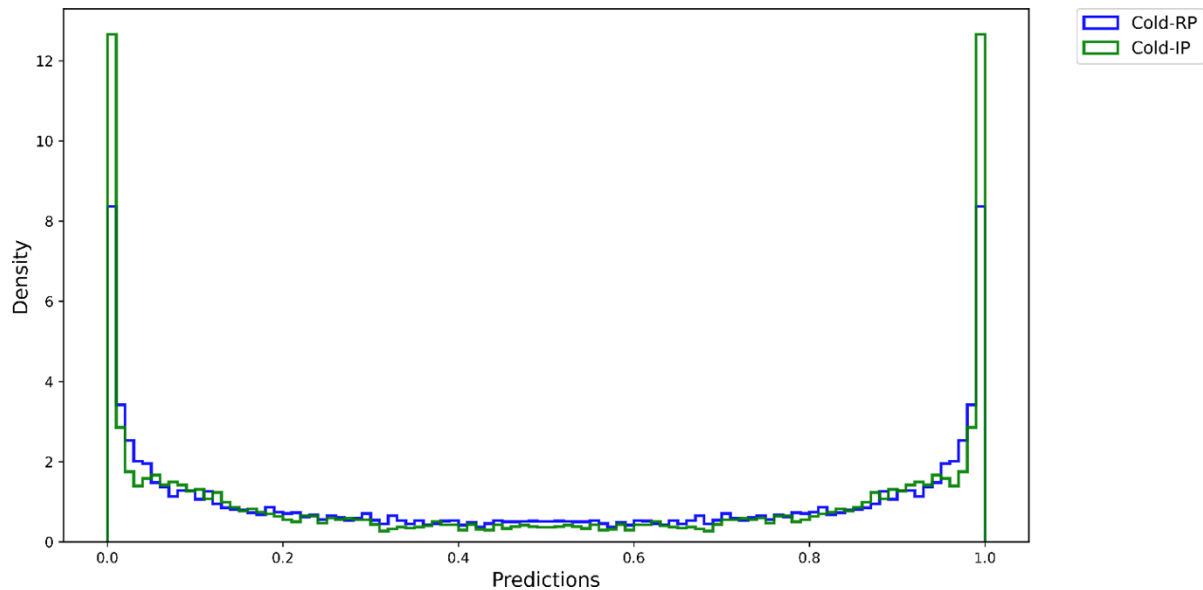
334 Fig 2 depicts histograms that illustrate the distribution of Softmax activations for the positive (1 -
 335 *Abnormal*) and negative (0 - *No Finding*) classes when predicting the RSNA-P test set using the Cold-RP
 336 and Cold-IP models. The Softmax histograms provide insight into the correctness and confidence of each
 337 model's predictions, as well as differences in Softmax predictions and overall performance. The x-axis
 338 represents Softmax activations, and the y-axis indicates the density of these activations. The histograms'
 339 shape and density reveal a more distinct separation between the two classes in the Cold-IP model,
 340 characterized by two clear peaks near 0 and 1. This distinction may result from the Cold-IP model's
 341 initialization with ImageNet-pretrained weights, allowing it to leverage useful features learned from a large-
 342 scale dataset.



343

344 **Fig 1. Internal adult test performance comparison between the Cold-RP and Cold-IP models. (a)**

345 AUPRC, (b) Confusion matrices, and (c) MCC comparison with the p -value.



346

347 **Fig 2. Histograms of the Softmax activations of Cold-RP and Cold-IP models.**

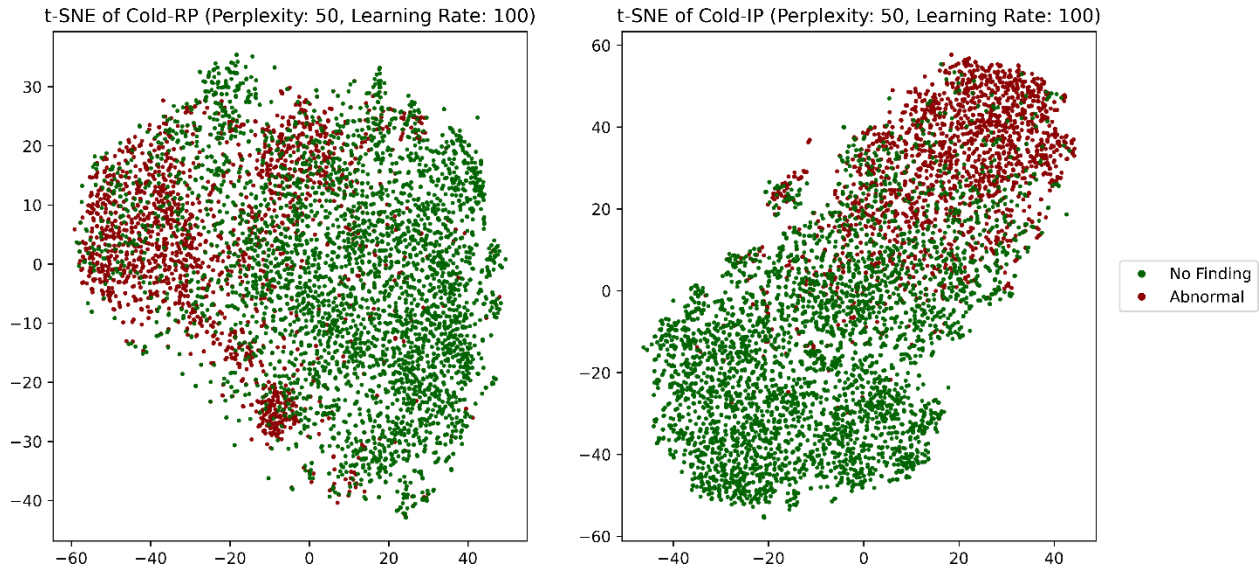
348

349 Consequently, the model converges more effectively and generates more accurate and confident predictions
 350 for both classes. In contrast, the Cold-RP model, initialized with random weights, exhibits a less distinct
 351 separation and a wider distribution of predictions around 0.5, suggesting lower confidence and correctness
 352 in its predictions. These findings underscore the superior performance of the Cold-IP model relative to the
 353 Cold-RP model.

354 We also use t-SNE visualizations [44] to assess the feature representations learned by the Cold-RP
 355 and Cold-IP models in the 2D space (Fig 3). The t-SNE visualization allows us to effectively evaluate each
 356 model's ability to capture the data's underlying structure and its generalizability. We determine the optimal
 357 perplexity and learning rate parameters through rigorous empirical analysis. The t-SNE plot highlights
 358 distinct visual disparities in the models' learned features. Although both models acquire meaningful data
 359 representations, the Cold-IP model's t-SNE presents two well-defined clusters for the *No Finding* and
 360 *Abnormal* classes, indicating that the Cold-IP model more effectively captures the data's essential features
 361 and generalizes to the internal adult test set. This representation can potentially enhance classification
 362 performance on unseen data. Conversely, the Cold-RP model's t-SNE displays greater class overlap.
 363 Despite some separation, the clusters are less defined compared to the Cold-IP model. This diminished

364 separation implies that the Cold-RP model struggles to accurately classify test set instances, particularly
365 near the decision boundary. The t-SNE visualizations underscore that the Cold-IP model exhibits superior
366 generalization capabilities relative to the Cold-RP model.

367



368

369 **Fig 3. The t-SNE visualization of the features learned by the Cold-RP (left) and Cold-IP (right)**
370 **models.**

371

372 We proceed to train and evaluate models on 100% of the data, i.e., the RSNA-F dataset, with the
373 aforementioned configurations for Cold-RF, Warm-RF, Shrink-RF, Cold-IF, Warm-IF, and Shrink-IF
374 models (Table 3). The weights of the Cold-RP and Cold-IP models that are used to initialize the weights
375 for the Shrink-RF model and the Shrink-IF models, respectively, are shrunk by an optimal scaling factor of
376 0.7209 (α_1) and 0.9 (α_2), respectively, as determined by Bayesian optimization through GP minimization
377 in the constrained continuous interval of [0.1, 0.9].

378

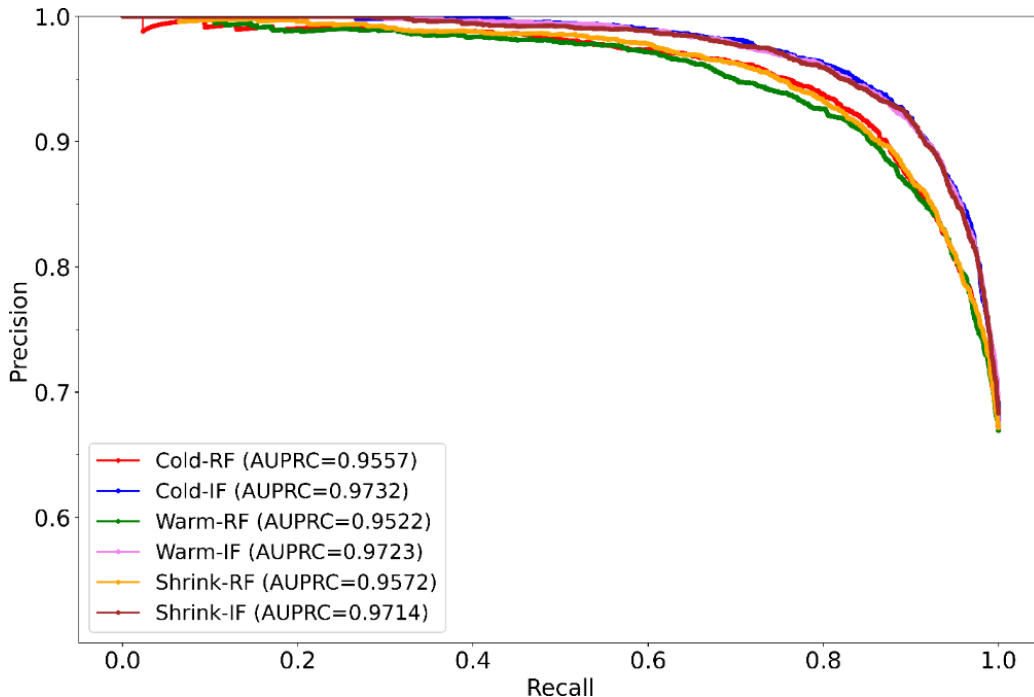
379 Table 5 displays performance metrics, while Fig 4 illustrates the AUPRC achieved by each model
380 when predicting the RSNA-F test (i.e., the internal adult test set). We observe that the models initialized
with ImageNet-pretrained weights (Cold-IF, Warm-IF, Shrink-IF) converge considerably faster and also

381 significantly outperform their randomly initialized counterparts (Cold-RF, Warm-RF, Shrink-RF) in terms
 382 of MCC ($p < 0.00001$) and other metrics.

383
 384 **Table 5: Performance of models on the internal adult test set.** Bold numerical values denote superior
 385 performance in their respective columns. The * denotes statistically significant MCC among each model
 386 pair, i.e., (Cold-RF, Cold-IF), (Warm-RF, Warm-IF), and (Shrink-RF, Shrink-IF) ($p < 0.00001$).

Model	AUPRC	B. Acc.	P	R	F	MCC	Training time (in sec.)	p -MCC
Cold-RF	0.9557	0.8383	0.9015	0.8676	0.8842	0.6650 (0.6523,0.6777)	3098.99	<0.00001
Cold-IF	0.9732	0.8534	0.8958	0.9232	0.9093	0.7187 (0.7066,0.7308)*	1458.36	
Warm-RF	0.9522	0.8036	0.8593	0.9061	0.8821	0.6267 (0.6137,0.6397)	1453.58	<0.00001
Warm-IF	0.9723	0.8686	0.9214	0.8904	0.9056	0.7258 (0.7138,0.7378)*	1067.58	
Shrink-RF	0.9572	0.8158	0.8711	0.8999	0.8853	0.6431 (0.6302,0.6560)	1982.63	<0.00001
Shrink-IF	0.9714	0.8508	0.8934	0.9237	0.9083	0.7150 (0.7028,0.7272)*	1205.31	

387



388

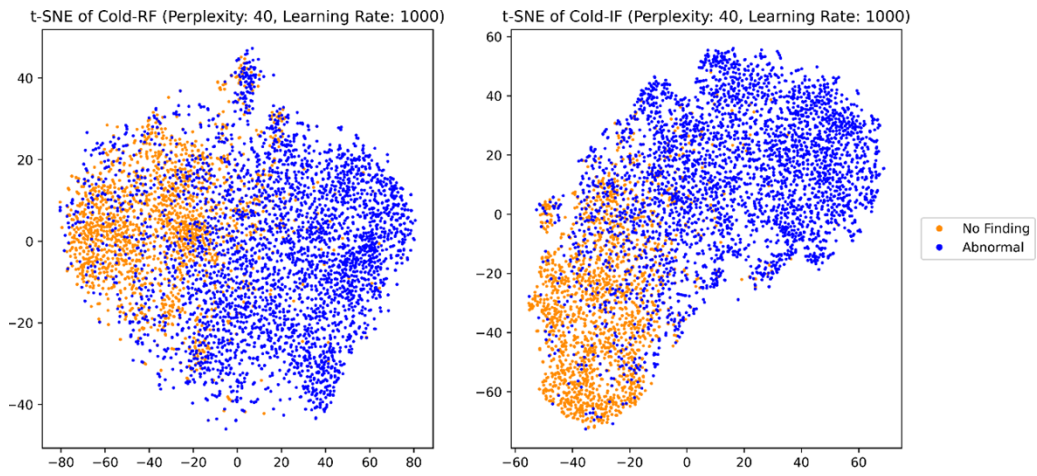
389 **Fig 4. AUPRC of the models while predicting the internal adult test set.**

390 Among the ImageNet-initialized models, the Cold-IF model takes slightly longer to converge (1458.36
391 seconds) compared to the Warm-IF (1067.58 seconds) and Shrink-IF models (1205.31 seconds). The
392 confusion matrices of these models are shown in S1 Figure. S2 Figure shows a comparison of MCC values
393 for each model pair, i.e., (Cold-RF, Cold-IF), (Warm-RF, Warm-IF), and (Shrink-RF, Shrink-IF).

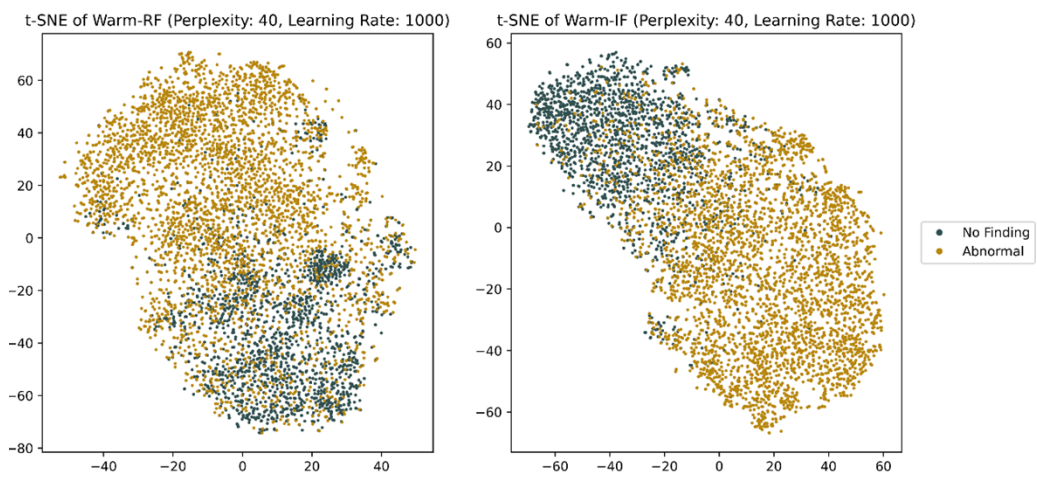
394 Analyzing the t-SNE visualizations in Fig 5 allows us to glean insights into the generalization
395 abilities of each model within their respective pairs: (Cold-RF, Cold-IF), (Warm-RF, Warm-IF), and
396 (Shrink-RF, Shrink-IF). We determine the ideal perplexity and learning rate values for each model through
397 extensive empirical evaluations. In the Cold-RF versus Cold-IF comparison, the Cold-IF model, initialized
398 with ImageNet-pretrained weights, showcases more distinct clustering and superior class separation than
399 its randomly initialized counterpart. Similarly, the Warm-IF model demonstrates clearer data point
400 separation into distinct clusters compared to Warm-RF in their respective comparison. Lastly, the Shrink-
401 IF model presents more well-defined clusters and class separations than Shrink-RF. Based on the t-SNE
402 visualizations, we observe that ImageNet-initialized models (Cold-IF, Warm-IF, Shrink-IF) exhibit
403 enhanced generalization capabilities compared to their randomly initialized counterparts. These
404 observations underscore the significance of employing ImageNet-pretrained weights to boost performance
405 and generalizability in such models.

406 The Softmax activation histograms (S3 Figure) help visualize performance disparities. In the Cold-
407 RF versus Cold-IF comparison, the Cold-IF model exhibits a bimodal distribution with peaks near 0 and 1,
408 suggesting confident, accurate predictions for both classes. Conversely, the Cold-RF model displays a
409 uniform distribution without a preference for either class, indicating less confident, less accurate
410 predictions. In the Warm-RF versus Warm-IF comparison, we observe that the Warm-IF model's histogram
411 displays a distinct bimodal distribution, indicative of confident, accurate predictions. The Warm-RF model
412 exhibits a less pronounced bimodal distribution, signaling lower prediction confidence. The Warm-IF
413 model's superior performance corresponds with its histogram's well-defined bimodal distribution.
414 Similarly, the Shrink-RF and Shrink-IF model pair reveal performance differences. The Shrink-IF model's
415 histogram presents a prominent bimodal distribution, implying greater confidence and accuracy in

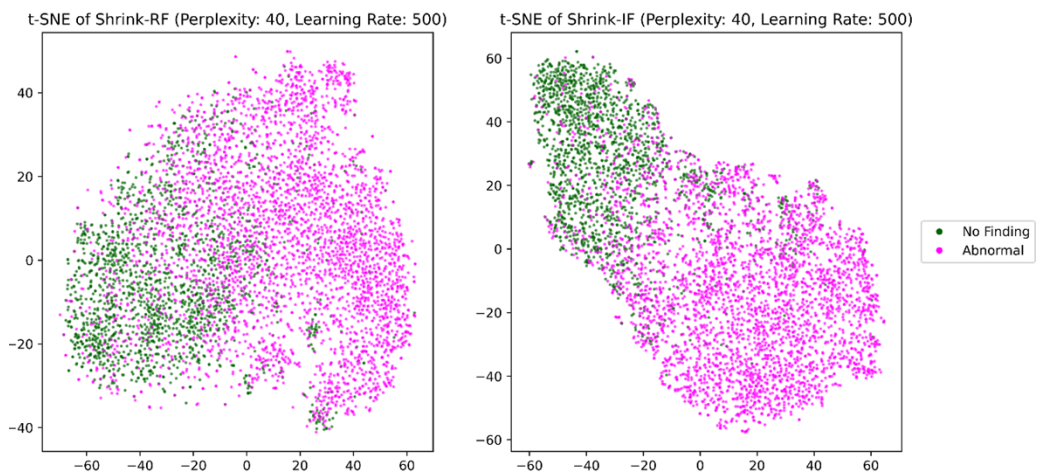
416 predictions, whereas the Shrink-RF model shows a less distinct distribution, reflecting weaker prediction
417 capabilities.



(a)



(b)



(c)

418

419 **Fig 5. t-SNE visualization of the features learned by the model pairs.** (a) Cold-RF and Cold-IF; (b)
420 Warm-RF and Warm-IF, and (c) Shrink-RF and Shrink-IF.

421

422 As we examine Table 5, we make an intriguing observation. When predicting the internal adult
423 test, the Cold-IF model only marginally outperforms the Warm-IF and Shrink-IF models in terms of
424 AUPRC, F-score, and MCC. The Warm-IF model shows slightly higher balanced accuracy and precision,
425 while the Shrink-IF model exhibits marginally better recall. However, there are no significant performance
426 differences observed for the MCC metric ($p>0.05$). Other metrics also demonstrate similar values across
427 the models. Nevertheless, considering their superior performance compared to their randomly-initialized
428 counterparts, the Cold-IF, Warm-IF, and Shrink-IF models did not demonstrate significant differences in
429 their generalizability to the internal test set.

430 Similar trends are observed when assessing external generalization in Table 6. For the external
431 adult test, the Cold-IF model only marginally, but not significantly ($p>0.05$), outperforms the Warm-IF and
432 Shrink-IF models in terms of MCC. This observation holds for balanced accuracy, recall, and F-score. The
433 Warm-IF model slightly outperforms the other models in terms of AUPRC and precision. When predicting
434 the external Ped-2 test, the Cold-IF model slightly outperforms the Warm-IF and Shrink-IF models in terms
435 of all metrics. However, no significant difference in performance is observed for the MCC metric ($p>0.05$).
436 Similar performance trends are observed for the Ped-11 and Ped-18 test sets. With the Ped-11 test, the
437 Warm-IF model marginally outperforms ($p>0.05$) the other models in terms of MCC. The Cold-IF model
438 demonstrates slightly superior values for AUPRC, recall, and F-score. The Shrink-IF model performs the
439 worst among all models. With the Ped-18 test, the Warm-IF model achieves marginally superior balanced
440 accuracy, precision, F-score, and MCC. The Shrink-IF model shows slightly better recall and AUPRC,
441 while the Cold-IF model exhibits the lowest performance. These observations suggest that, despite
442 differences in training scenarios, the ImageNet-initialized models, namely Cold-IF, Warm-IF, and Shrink-
443 IF, might have converged to distinct local optima that enable comparable generalization performance across
444 the external test sets.

445 **Table 6. Comparing the model performances when predicting the external adult and pediatric test**
 446 **sets. Bold numerical values denote superior performance in their respective columns.**

Test	Models	AUPRC	B. Acc.	P	R	F	MCC	p -MCC
Adult	Cold-IF	0.8490	0.7170	0.8452	0.5807	0.6884	0.4378 (0.4226, 0.4530)	>0.05
	Warm-IF	0.8573	0.6942	0.8939	0.4643	0.6112	0.4180 (0.4029, 0.4331)	
	Shrink-IF	0.8472	0.7073	0.8599	0.5345	0.6592	0.4263 (0.4111, 0.4415)	
Ped-2	Cold-IF	0.4685	0.5480	0.3997	0.8794	0.5496	0.1206 (0.1118, 0.1294)	>0.05
	Warm-IF	0.4289	0.5394	0.3953	0.8514	0.5399	0.0955 (0.0876, 0.1034)	
	Shrink-IF	0.4589	0.5362	0.3927	0.8769	0.5425	0.0936 (0.0858, 0.1014)	
Ped-11	Cold-IF	0.5936	0.6235	0.4861	0.7519	0.5905	0.2458 (0.2340, 0.2576)	>0.05
	Warm-IF	0.5876	0.6327	0.5063	0.6976	0.5868	0.2595 (0.2475, 0.2715)	
	Shrink-IF	0.5887	0.6243	0.4881	0.7444	0.5896	0.2465 (0.2347, 0.2583)	
Ped-18	Cold-IF	0.6726	0.7116	0.5871	0.8569	0.6968	0.4281 (0.4117, 0.4281)	>0.05
	Warm-IF	0.682	0.7324	0.6229	0.8241	0.7095	0.4614 (0.4448, 0.4780)	
	Shrink-IF	0.6822	0.7113	0.5849	0.8643	0.6977	0.4293 (0.4129, 0.4457)	

447

448 To assess weight distribution similarity, we generated a heatmap of Earth Mover Distance (EMD)

449 values in Fig 6. Lower EMD values indicate higher weight similarity due to shared ImageNet-pretrained

450 weight initialization for the Cold-IF, Warm-IF, and Shrink-IF models. This similarity, supported by the low

451 EMD values, aligns with the observation that the models' performance differences are not pronounced.

452 We further analyze the weight distribution similarity of the Cold-IF, Warm-IF, and Shrink-IF models using

453 scatter plots (Fig 7). The plots visually depict the relationship between the weight distributions of each

454 model pair, namely (Cold-IF, Warm-IF), (Cold-IF, Shrink-IF), and (Warm-IF, Shrink-IF). Each point in the

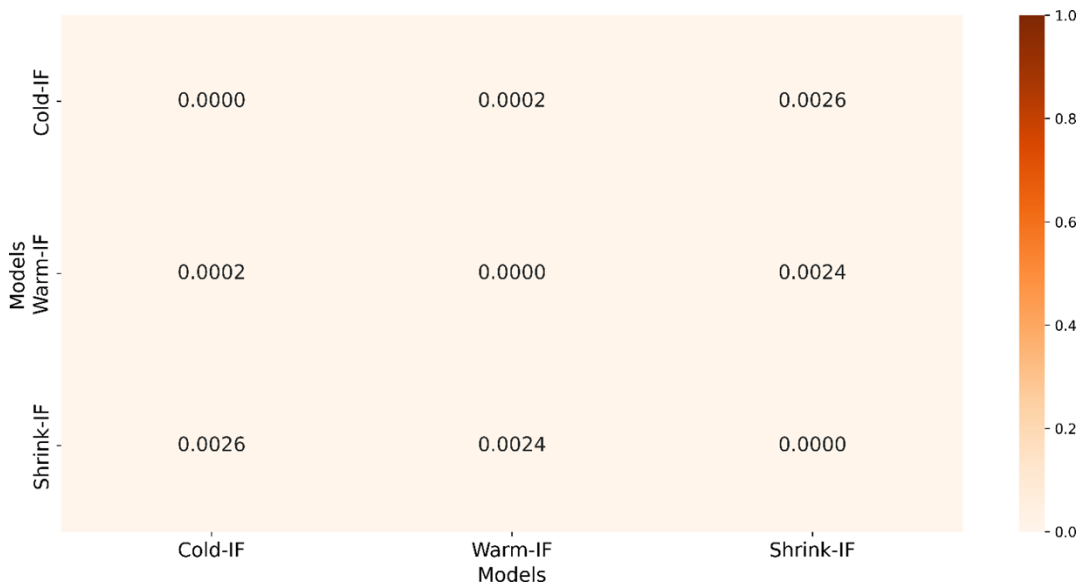
455 scatter plot represents a pair of weights from the compared models, with the x-axis and y-axis representing

456 the weights of the respective models. Dense point distributions along the diagonal indicate higher weight

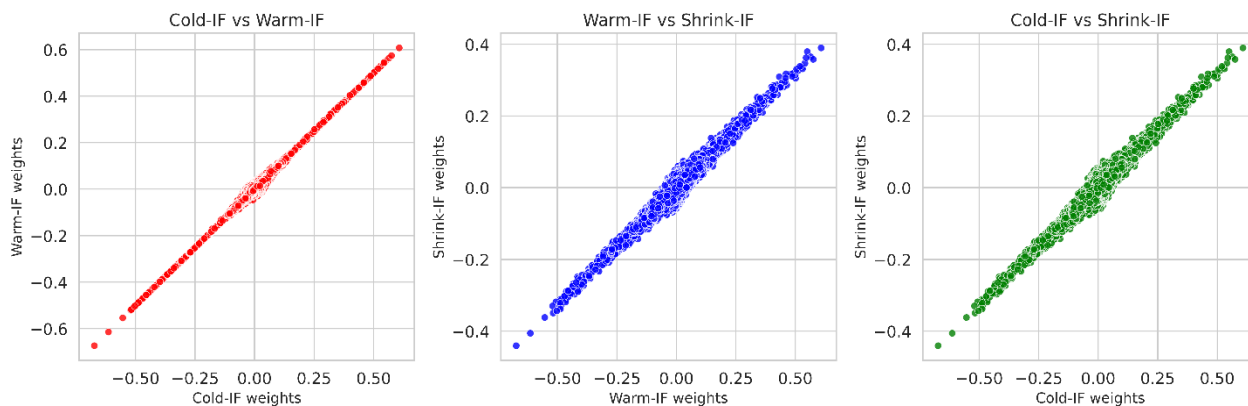
457 similarity, while more dispersed distributions suggest less similarity. We observe a strong positive

458 correlation between weight distributions as evident from the scatter plot patterns. The scatter plots

459 demonstrate a dense diagonal distribution, indicating highly similar weight distributions for the compared
 460 models. This similarity implies that the models learned similar features and representations during training,
 461 resulting in comparable Softmax predictions for the positive and negative classes, as supported by their
 462 performance metrics.



463
 464 **Fig 6. Heatmap showing EMD values between each model pair for the Cold-IF, Warm-IF, and**
 465 **Shrink-IF models.**



467
 468 **Fig 7. Scatter plots show the correlation in weights between pairs of models.**

469 We also apply ensemble methods to evaluate if the generalization performance with internal and
 470 external test sets can surpass that of the individual models. Table 7 presents the ensemble performances
 471 when predicting the internal adult test.

472

473 **Table 7. Model performances achieved with the internal adult test.** Bold numerical values denote
 474 superior performance in their respective columns. The * denotes statistically significant recall ($p < 0.00001$)
 475 compared to the baseline.

Models	AUPRC	B. Acc.	P	R	F	MCC
Warm-IF-Baseline	0.9723	0.8686	0.9214	0.8904	0.9056	0.7258 (0.7138,0.7378)
EWA Ensemble						
Cold-IF, Warm-IF	0.9709	0.8548	0.8999	0.9148	0.9073	0.7159 (0.7037,0.7281)
Cold-IF, Shrink-IF	0.9707	0.8611	0.9100	0.9019	0.9059	0.7190 (0.7069,0.7311)
Warm-IF, Shrink-IF	0.9730	0.8697	0.9200	0.8965	0.9081	0.7305 (0.7185,0.7425)
Cold-IF, Warm-IF, Shrink-IF	0.9712	0.8551	0.8993	0.9170	0.9081	0.7177 (0.7056,0.7298)
F-SLSQP Ensemble						
Cold-IF, Warm-IF	0.9724	0.8680	0.9204	0.8915	0.9057	0.7254 (0.7134,0.7374)
Cold-IF, Shrink-IF	0.9722	0.8635	0.9096	0.9089	0.9092	0.7266 (0.7146,0.7386)
Warm-IF, Shrink-IF	0.9728	0.8604	0.9053	0.9136	0.9094	0.7244 (0.7124,0.7364)
Cold-IF, Warm-IF, Shrink-IF	0.9722	0.8611	0.9068	0.9108	0.9088	0.7238 (0.7118,0.7358)
AGELFS						
Cold-IF, Warm-IF	0.9731	0.8698	0.9218	0.8920	0.9067	0.7284 (0.7164,0.7404)
Cold-IF, Shrink-IF	0.9734	0.8598	0.9026	0.9195*	0.9110	0.7267 (0.7147,0.7387)
Warm-IF, Shrink-IF	0.9727	0.8591	0.9020	0.9192	0.9105	0.7254 (0.7134,0.7374)
Cold-IF, Warm-IF, Shrink-IF	0.9729	0.8558	0.8980	0.9229	0.9103	0.7224 (0.7103,0.7345)

476

477 We select the baseline model based on the best MCC performance reported for the individual models in
 478 Table 5. We observe that the Attention-Guided Ensemble with Learnable Fuzzy Softmax (AGELFS) of the
 479 Cold-IF and Shrink-IF models deliver significantly superior values for recall ($p < 0.00001$) and marginally
 480 higher values for AUPRC and F-score among other ensemble methods. The AGELFS of the Cold-IF and
 481 Warm-IF models delivers higher but not significantly superior values for balanced accuracy and precision.
 482 The learned Fuzziness values for the Softmax Layer in the AGELFS ensemble are 1.113, 1.113, 1.039, and

483 1.044 for the model pairs (Cold-IF, Warm-IF), (Cold-IF, Shrink-IF), (Warm-IF, Shrink-IF), and (Cold-IF,
484 Warm-IF, Shrink-IF), respectively. The Equal Weight Averaging (EWA) ensemble of the Warm-IF and
485 Shrink-IF models yields a marginally higher MCC value compared to the baseline.

486 Table 8 presents the performances achieved with the external adult and pediatric test sets. For
487 brevity, we present here only the key results in a single table while the complete tables are included in the
488 Supplementary (S1 Table, S2 Table, S3 Table, and S4 Table). We observe sub-optimal external
489 generalization compared to the results achieved with the internal test set in Table 7. For the external adult
490 test set, the individual Cold-IF model achieves a relatively higher MCC of 0.4378 among other individual
491 models and so we choose it as the baseline. The F-SLSQP ensemble of the Cold-IF and Warm-IF models
492 demonstrates significantly superior precision ($p < 0.00001$) and the highest AUPRC. The EWA ensemble of
493 the Cold-IF and Warm-IF models achieves a marginally higher balanced accuracy, recall, F-score, and
494 MCC compared to the baseline and other tested combinations. For the Ped-2 test set, the Cold-IF model
495 serves as the baseline. The EWA ensemble significantly improves recall ($p < 0.00001$). We use the Warm-
496 IF as the baseline for the Ped-11 test set. The EWA ensemble of Cold-IF, Warm-IF, and Shrink-IF models
497 demonstrates significantly superior values for recall ($p < 0.00001$). The AGELFS of Cold-IF and Warm-IF
498 models demonstrate higher values for precision; however, these values are not markedly different compared
499 to the individual Warm-IF model, which demonstrates the highest MCC compared to the ensembles. We
500 use the Warm-IF as the baseline for the Ped-18 test set. The EWA ensemble of Cold-IF, Warm-IF, and
501 Shrink-IF models achieves significantly superior recall ($p < 0.00007$), while the AGELFS of Cold-IF and
502 Warm-IF models yield higher, yet not markedly different, balanced accuracy, precision, F-score, and MCC
503 values.

504
505 **Table 8. Performances achieved with the external adult and pediatric test sets.** Bold numerical values
506 denote superior performance in their respective columns. The * denotes statistical significance for the
507 respective column metric compared to the baseline models for each external test set.

Models	AUPRC	B. Acc.	P	R	F	MCC
Adult						
Cold-IF-Baseline	0.8490	0.7170	0.8452	0.5807	0.6884	0.4378 (0.4226,0.4530)
EWA Ensemble						
Cold-IF, Warm-IF	0.8557	0.7272	0.8519	0.5976	0.7024	0.4568 (0.4415,0.4721)
F-SLSQP Ensemble						
Cold-IF, Warm-IF	0.8591	0.6961	0.8929*	0.4697	0.6156	0.4205 (0.4053,0.4357)
Ped-2						
Cold-IF-Baseline	0.4685	0.5480	0.3997	0.8794	0.5496	0.1206 (0.1118,0.1294)
EWA Ensemble						
Cold-IF, Warm-IF, Shrink-IF	0.4367	0.5240	0.3847	0.9335*	0.5449	0.0785 (0.0713,0.0857)
AGELFS						
Cold-IF, Shrink-IF	0.4567	0.5475	0.4003	0.8529	0.5449	0.1135 (0.1050,0.1220)
Ped-11						
Warm-IF-Baseline	0.5381	0.6446	0.4368	0.6976	0.5372	0.2681 (0.2559,0.2803)
EWA Ensemble						
Cold-IF, Warm-IF	0.5381	0.6222	0.3963	0.7918	0.5282	0.2336 (0.2220,0.2452)
Cold-IF, Warm-IF, Shrink-IF	0.5380	0.6180	0.3924	0.7943*	0.5253	0.2267 (0.2152,0.2382)
AGELFS						
Cold-IF, Warm-IF	0.5450	0.6422	0.4374	0.6833	0.5334	0.2636 (0.2515,0.2757)
Cold-IF, Shrink-IF	0.5528	0.6420	0.4297	0.7145	0.5367	0.2635 (0.2514,0.2756)
Cold-IF, Warm-IF, Shrink-IF	0.5411	0.6414	0.4236	0.7394	0.5386	0.2631 (0.2510,0.2752)
Ped-18						
Warm-IF-Baseline	0.6820	0.7324	0.6229	0.8241	0.7095	0.4614 (0.4448,0.4780)
EWA Ensemble						
Cold-IF, Warm-IF, Shrink-IF	0.6698	0.7144	0.5890	0.8616*	0.6997	0.4342 (0.4177,0.4507)
AGELFS						
Cold-IF, Warm-IF	0.6804	0.7368	0.6237	0.8371	0.7148	0.4708 (0.4542,0.4874)
Cold-IF, Warm-IF, Shrink-IF	0.6852	0.7178	0.5939	0.8582	0.7020	0.4397 (0.4232,0.4562)

508

509 We describe below our assessment of potential reasons for the significant improvement in recall ($p < 0.05$)
510 when using the EWA ensemble of Cold-IF, Warm-IF, and Shrink-IF models to predict the external pediatric
511 test sets:

512 (i) Diverse error patterns: The models in the EWA ensemble exhibit different error patterns for the same
513 classification task. The EWA ensemble excels at identifying true positive (TP) samples and enhancing

514 recall. However, an increase in false positive (FP) predictions could counteract precision improvements,
515 resulting in relatively unchanged F-score, MCC, and AUPRC.

516 (ii) Ensemble learning bias-variance tradeoff: Ensemble learning aims to reduce the bias and variance of
517 individual models for better generalization. The EWA ensemble decreases variance without significantly
518 impacting bias. Since recall is sensitive to reducing false negatives (FN) (i.e., variance reduction), it can
519 show significant improvement while other metrics remain unchanged if bias remains relatively constant.

520 (iii) Imbalanced datasets: In imbalanced datasets, EWA ensemble techniques can improve recall for the
521 minority class without significantly affecting other metrics. This is evident in the external pediatric test sets
522 where abnormal CXRs are fewer compared to normal samples. The EWA ensemble model's robustness
523 against overfitting and improved generalization in identifying minority class samples may not lead to
524 significant changes in other metrics. The aforementioned discussions also apply to the significantly superior
525 recall values obtained using the AGELFS of Cold-IF and Shrink-IF models for the internal adult test.

526

527 **Conclusion and future scope**

528 Diverse model initialization techniques are instrumental for deep model optimization thereby
529 affecting convergence speed, reducing the risk of overfitting, and improving generalizability. Our
530 qualitative and quantitative analyses validate the claim that cold-start approaches can decelerate
531 convergence while warm-start methods, such as ImageNet-pretrained weight initialization, enhance
532 convergence and performance. Furthermore, improper weight initialization can introduce biases that
533 inadvertently favor certain classes or feature sets which, in turn, increases the risk of model overfitting to
534 the data and reducing generalizability. To mitigate this risk, we perform ensemble learning and propose
535 novel weight-level ensemble methods to improve performance over individual constituent models. These
536 ensembles can harness a broader range of feature representations, making them more adaptable and
537 effective when handling unseen data. This adaptability is particularly relevant in medical computer vision,

538 where models must demonstrate exceptional generalizability across diverse patient populations and
539 imaging modalities.

540 Future research could explore alternative ensemble methods, such as advanced stacking or voting
541 techniques, to further improve generalization. Further, incorporating demographic factors during model
542 initialization could enable the development of personalized DL models for medical image analysis,
543 extending the scope of this research to other medical imaging tasks and modalities. Pursuing these research
544 directions could help improve medical computer vision DL models for reliable healthcare applications.

545

546 **Acknowledgments**

547 This study is supported by the Intramural Research Program (IRP) of the National Library of Medicine
548 (NLM) and the National Institutes of Health (NIH).

549

550 **References**

- 551 1. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep
552 learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big*
553 *Data*. Springer International Publishing; 2021. doi:10.1186/s40537-021-00444-8
- 554 2. Power SP, Moloney F, Twomey M, James K, O'Connor OJ, Maher MM. Computed tomography
555 and patient risk: Facts, perceptions and uncertainties. *World J Radiol*. 2016;8: 902.
556 doi:10.4329/wjr.v8.i12.902
- 557 3. Kwee TC, Kwee RM. Workload of diagnostic radiologists in the foreseeable future based on
558 recent scientific advances: growth expectations and role of artificial intelligence. *Insights Imaging*.
559 2021;12. doi:10.1186/s13244-021-01031-4
- 560 4. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale Chest X-ray
561 Database and Benchmarks on Weakly-Supervised Classification and Localization of Common

- 562 Thorax Diseases. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
563 2017. pp. 1–19. doi:10.1109/CVPR.2017.369
- 564 5. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilicus S, Chute C, et al. CheXpert: A large chest
565 radiograph dataset with uncertainty labels and expert comparison. 33rd AAAI Conf Artif Intell
566 AAAI 2019, 31st Innov Appl Artif Intell Conf IAAI 2019 9th AAAI Symp Educ Adv Artif Intell
567 EAAI 2019. 2019; 590–597. doi:10.1609/aaai.v33i01.3301590
- 568 6. Jaeger S, Candemir S, Antani S, Wang Y-XJ, Lu P-X, Thoma G. Two public chest X-ray datasets
569 for computer-aided screening of pulmonary diseases. *Quant Imaging Med Surg.* 2014;4: 475–477.
570 doi:10.3978/j.issn.2223-4292.2014.11.20
- 571 7. Jabbour S, Fouhey D, Kazerooni E, Wiens J, Sjoding MW. Combining chest X-rays and electronic
572 health record (EHR) data using machine learning to diagnose acute respiratory failure. *J Am Med*
573 *Informatics Assoc.* 2022;29: 1060–1068. doi:10.1093/jamia/ocac030
- 574 8. Pyrros A, Rodriguez Fernandez J, Borstelmann SM, Flanders A, Wenzke D, Hart E, et al.
575 Validation of a deep learning, value-based care model to predict mortality and comorbidities from
576 chest radiographs in COVID-19. *PLOS Digit Heal.* 2022;1: e0000057.
577 doi:10.1371/journal.pdig.0000057
- 578 9. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding transfer learning for
579 medical imaging. *Adv Neural Inf Process Syst.* 2019;32.
- 580 10. Ash JT, Adams RP. On Warm-Starting Neural Network Training. 2020;33: 1–11.
- 581 11. Xin KZ, Li D, Yi PH. Limited generalizability of deep learning algorithm for pediatric pneumonia
582 classification on external data. *Emerg Radiol.* 2022;29: 107–113. doi:10.1007/s10140-021-01954-
583 x
- 584 12. Lecun YA, Yoshua B, Geoffrey H, Bengio Y, Hinton GE. Deep learning. *Nature.* 2015;521: 436–
585 444. doi:10.1038/nature14539
- 586 13. Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions. 2016. Available:
587 <http://arxiv.org/abs/1610.02357>

- 588 14. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks.
589 Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017.
590 2017. doi:10.1109/CVPR.2017.243
- 591 15. Tan M, Le Q V. EfficientNet: Rethinking model scaling for convolutional neural networks. 36th
592 International Conference on Machine Learning, ICML 2019. 2019.
- 593 16. Suzuki K. Overview of deep learning in medical imaging. Radiological Physics and Technology.
594 2017. pp. 257–273. doi:10.1007/s12194-017-0406-5
- 595 17. Oda S, Awai K, Suzuki K, Yanaga Y, Funama Y, MacMahon H, et al. Performance of radiologists
596 in detection of small pulmonary nodules on chest radiographs: Effect of rib suppression with a
597 massive-training artificial neural network. Am J Roentgenol. 2009. doi:10.2214/AJR.09.2431
- 598 18. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization
599 performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional
600 study. PLoS Med. 2018;15: 1–17. doi:10.1371/journal.pmed.1002683
- 601 19. Sabottke CF, Spieler BM. The effect of image resolution on deep learning in radiography. Radiol
602 Artif Intell. 2020;2: 1–8. doi:10.1148/ryai.2019190015
- 603 20. Rajaraman S, Antani SK, Poostchi M, Silamut K, Hossain MA, Maude RJ, et al. Pre-trained
604 convolutional neural networks as feature extractors toward improved malaria parasite detection in
605 thin blood smear images. PeerJ. 2018;6: e4568. doi:10.7717/peerj.4568
- 606 21. Huang G-H, Fu Q-J, Gu M-Z, Lu N-H, Liu K-Y, Chen T-B. Deep Transfer Learning for the
607 Multilabel Classification of Chest X-ray Images. Diagnostics. 2022;12.
608 doi:10.3390/diagnostics12061457
- 609 22. Yang Q, Liu Y, Cheng Y, Kang Y, Chen T, Yu H. Federated Learning. Synth Lect Artif Intell
610 Mach Learn. 2020;13: 1–207. doi:10.2200/S00960ED2V01Y201910AIM043
- 611 23. Li T, Sahu AK, Talwalkar A, Smith V. Federated Learning: Challenges, Methods, and Future
612 Directions. IEEE Signal Process Mag. 2020;37: 50–60. doi:10.1109/MSP.2020.2975749
- 613 24. Arthur R. Interpretation of the paediatric chest X-ray. Paediatr Respir Rev. 2000;1: 41–50.

- 614 doi:10.1053/prrv.2000.0018
- 615 25. Shih G, Wu CC, Halabi SS, Kohli MD, Prevedello LM, Cook TS, et al. Augmenting the National
616 Institutes of Health Chest Radiograph Dataset with Expert Annotations of Possible Pneumonia.
617 Radiol Artif Intell. 2019. doi:10.1148/ryai.2019180041
- 618 26. Demner-Fushman D, Kohli MD, Rosenman MB, Shooshan SE, Rodriguez L, Antani S, et al.
619 Preparing a collection of radiology examinations for distribution and retrieval. J Am Med
620 Informatics Assoc. 2016. doi:10.1093/jamia/ocv080
- 621 27. Pham VT, Tran CM, Zheng S, Vu TM, Nath S. Chest X-ray abnormalities localization via
622 ensemble of deep convolutional neural networks. Int Conf Adv Technol Commun. 2021;2021-
623 Octob: 125–130. doi:10.1109/ATC52653.2021.9598342
- 624 28. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray: Hospital-Scale Chest X-ray
625 Database and Benchmarks on Weakly Supervised Classification and Localization of Common
626 Thorax Diseases. Advances in Computer Vision and Pattern Recognition. 2019. doi:10.1007/978-
627 3-030-13969-8_18
- 628 29. Candemir S, Antani S. A review on lung boundary detection in chest X-rays. International Journal
629 of Computer Assisted Radiology and Surgery. 2019. doi:10.1007/s11548-019-01917-1
- 630 30. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image
631 segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial
632 Intelligence and Lecture Notes in Bioinformatics). 2015. doi:10.1007/978-3-319-24574-4_28
- 633 31. Rajaraman S, Yang F, Zamzmi G, Xue Z, Antani S. Assessing the Impact of Image Resolution on
634 Deep Learning for TB Lesion Segmentation on Frontal Chest X-rays. Diagnostics. 2023;13.
635 doi:10.3390/diagnostics13040747
- 636 32. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition.
637 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track
638 Proceedings. 2015.
- 639 33. Emmadi SC, Aerra MR, Bantu S. Performance Analysis of VGG-16 Deep Learning Model for

- 640 COVID-19 Detection using Chest X-Ray Images. 2023 10th International Conference on
641 Computing for Sustainable Global Development (INDIACom). 2023. pp. 1001–1007.
- 642 34. Rajaraman S, Candemir S, Kim I, Thoma G, Antani S, Rajaraman S, et al. Visualization and
643 Interpretation of Convolutional Neural Network Predictions in Detecting Pneumonia in Pediatric
644 Chest Radiographs. *Appl Sci*. 2018;8: 1715. doi:10.3390/app8101715
- 645 35. Bougias H, Georgiadou E, Malamateniou C, Stogiannos N. Identifying cardiomegaly in chest X-
646 rays: a cross-sectional study of evaluation and comparison between different transfer learning
647 methods. *Acta radiol*. 2021;62: 1601–1609. doi:10.1177/0284185120973630
- 648 36. Basha N, Kravaris C, Nounou H, Nounou M. Bayesian-optimized Gaussian process-based fault
649 classification in industrial processes. *Comput Chem Eng*. 2023;170: 108126.
650 doi:10.1016/j.compchemeng.2022.108126
- 651 37. Gozzi N, Giacomello E, Sollini M, Kirienko M, Ammirabile A, Lanzi P, et al. Image Embeddings
652 Extracted from CNNs Outperform Other Transfer Learning Approaches in Classification of Chest
653 Radiographs. *Diagnostics*. 2022;12. doi:10.3390/diagnostics12092084
- 654 38. Asgharnezhad H, Shamsi A, Alizadehsani R, Khosravi A, Nahavandi S, Sani ZA, et al. Objective
655 evaluation of deep uncertainty predictions for COVID-19 detection. *Sci Rep*. 2022;12: 1–11.
656 doi:10.1038/s41598-022-05052-x
- 657 39. Rajaraman S, Guo P, Xue Z, Antani SK. A Deep Modality-Specific Ensemble for Improving
658 Pneumonia Detection in Chest X-rays. *Diagnostics*. 2022;12. doi:10.3390/diagnostics12061442
- 659 40. Marques JPPG, Cunha DC, Harada LMF, Silva LN, Silva ID. A cost-effective trilateration-based
660 radio localization algorithm using machine learning and sequential least-square programming
661 optimization. *Comput Commun*. 2021;177: 1–9. doi:10.1016/j.comcom.2021.06.005
- 662 41. Zamzmi G, Rajaraman S, Hsu L-Y, Sachdev V, Antani S. Real-time echocardiography image
663 analysis and quantification of cardiac indices. *Med Image Anal*. 2022;80: 102438.
664 doi:10.1016/j.media.2022.102438
- 665 42. Shakerian R, Yadollahzadeh-Tabari M, Bozorgi Rad SY. Proposing a Fuzzy Soft-max-based

- 666 classifier in a hybrid deep learning architecture for human activity recognition. IET Biometrics.
667 2022;11: 171–186. doi:10.1049/bme2.12066
- 668 43. Eckmann P, Bandrowski A. PreprintMatch: A tool for preprint to publication detection shows
669 global inequities in scientific publication. PLoS One. 2023;18: 1–20.
670 doi:10.1371/journal.pone.0281659
- 671 44. Van Der Maaten L, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008.
672 doi:10.1007/s10479-011-0841-3

673

674 **Supporting information**

675 **S1 Figure. Confusion matrices for each model pair while predicting the internal adult test set.**

676 **S2 Figure. MCC comparison along with the p -value for each model pair while predicting the internal
677 adult test set.**

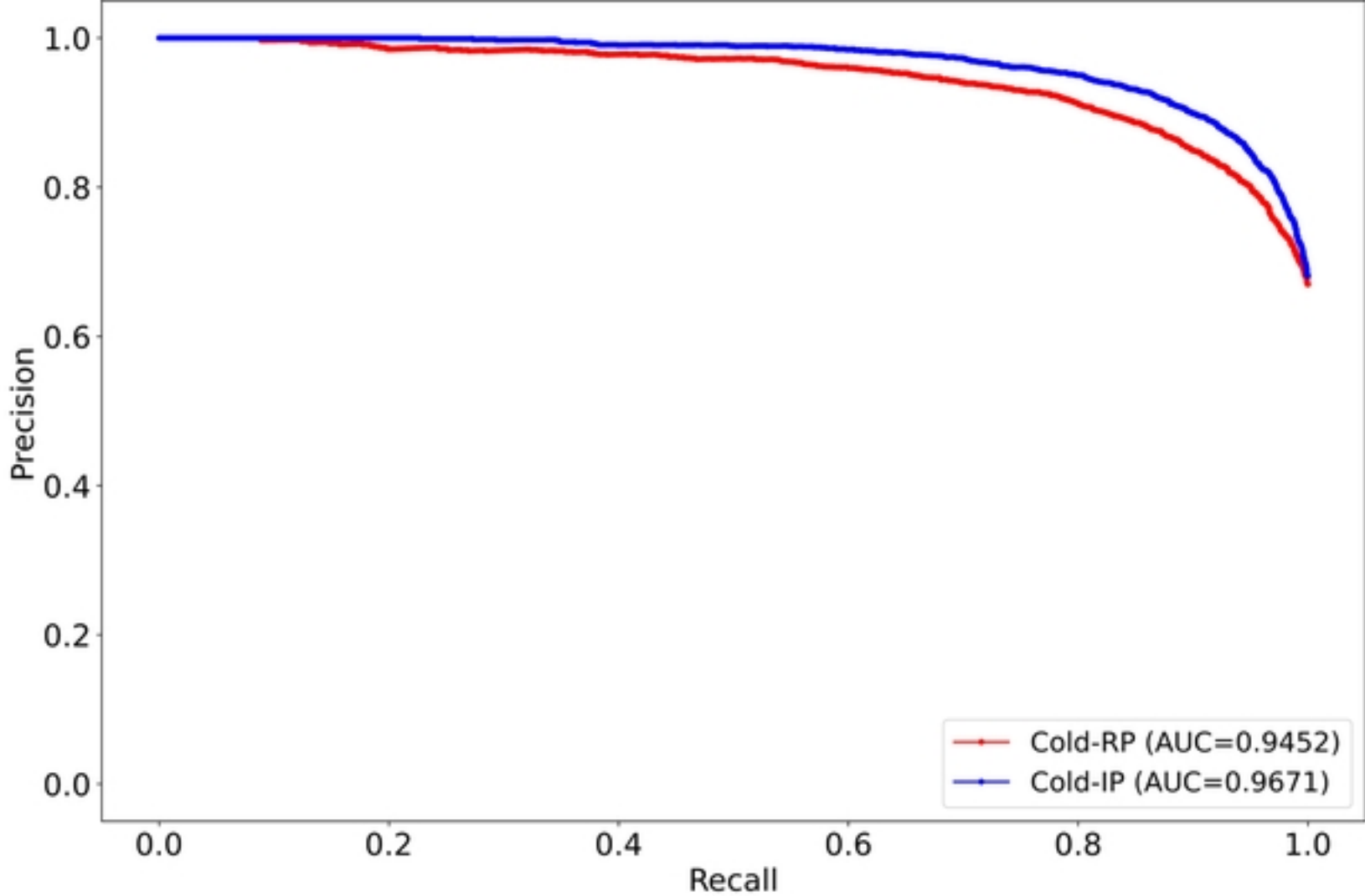
678 **S3 Figure. Histograms show the distribution of Softmax activations for each model pair.** (a) Cold-RF
679 and Cold-IF; (b) Warm-RF and Warm-IF, and (c) Shrink-RF and Shrink-IF.

680 **S1 Table. Performances achieved with the external adult test.** Bold numerical values denote superior
681 performance in their respective columns. The * denotes statistically significant precision ($p < 0.00001$)
682 compared to the baseline.

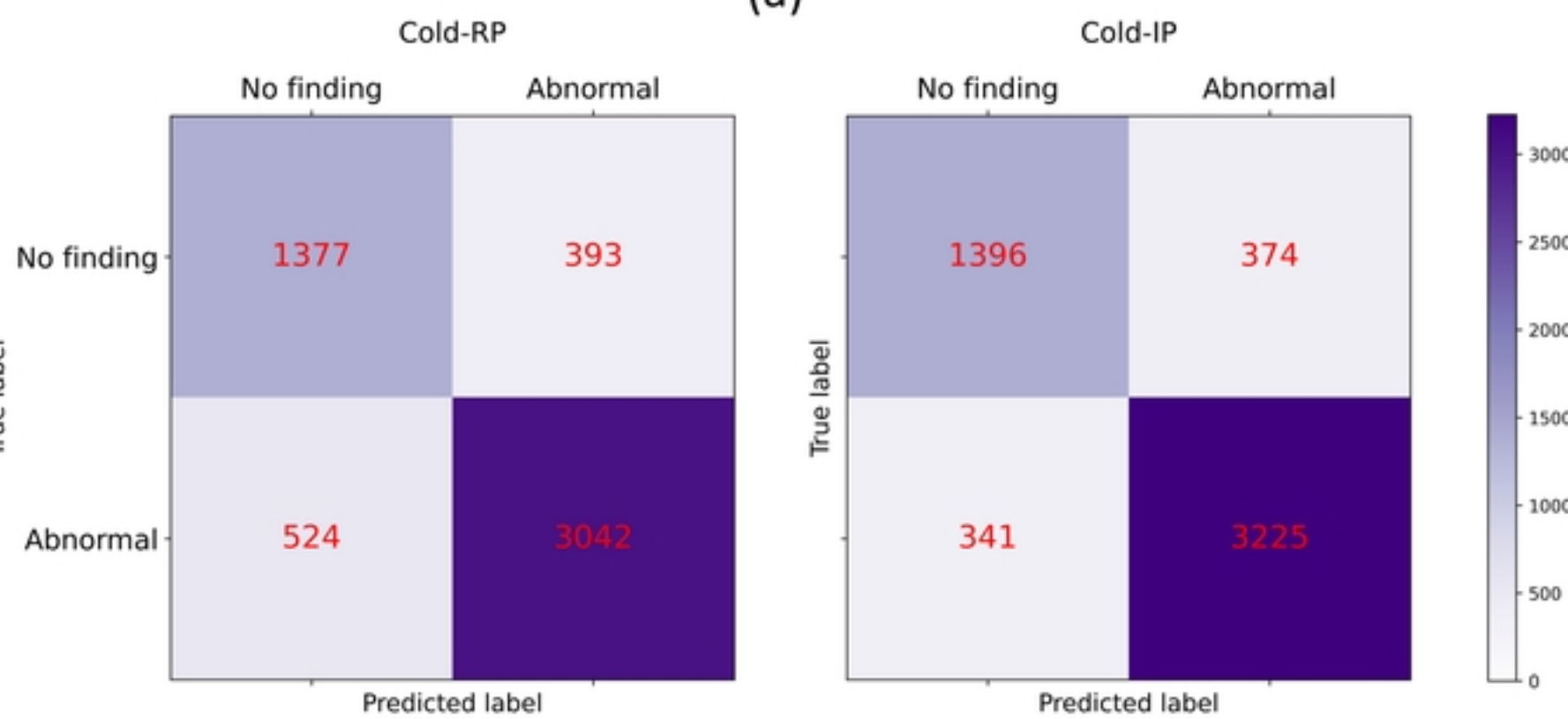
683 **S2 Table. Performances achieved with the external Ped-2 test.** Bold numerical values denote superior
684 performance in their respective columns. The * denotes statistically significant recall ($p < 0.00001$)
685 compared to the baseline.

686 **S3 Table. Performances achieved with the external Ped-11 test.** Bold numerical values denote superior
687 performance in their respective columns. The * denotes statistically significant recall ($p < 0.00001$)
688 compared to the baseline.

689 **S4 Table. Performances achieved with the external Ped-18 test.** Bold numerical values denote superior
690 performance in their respective columns. The * denotes statistically significant recall ($p < 0.00007$)
691 compared to the baseline.

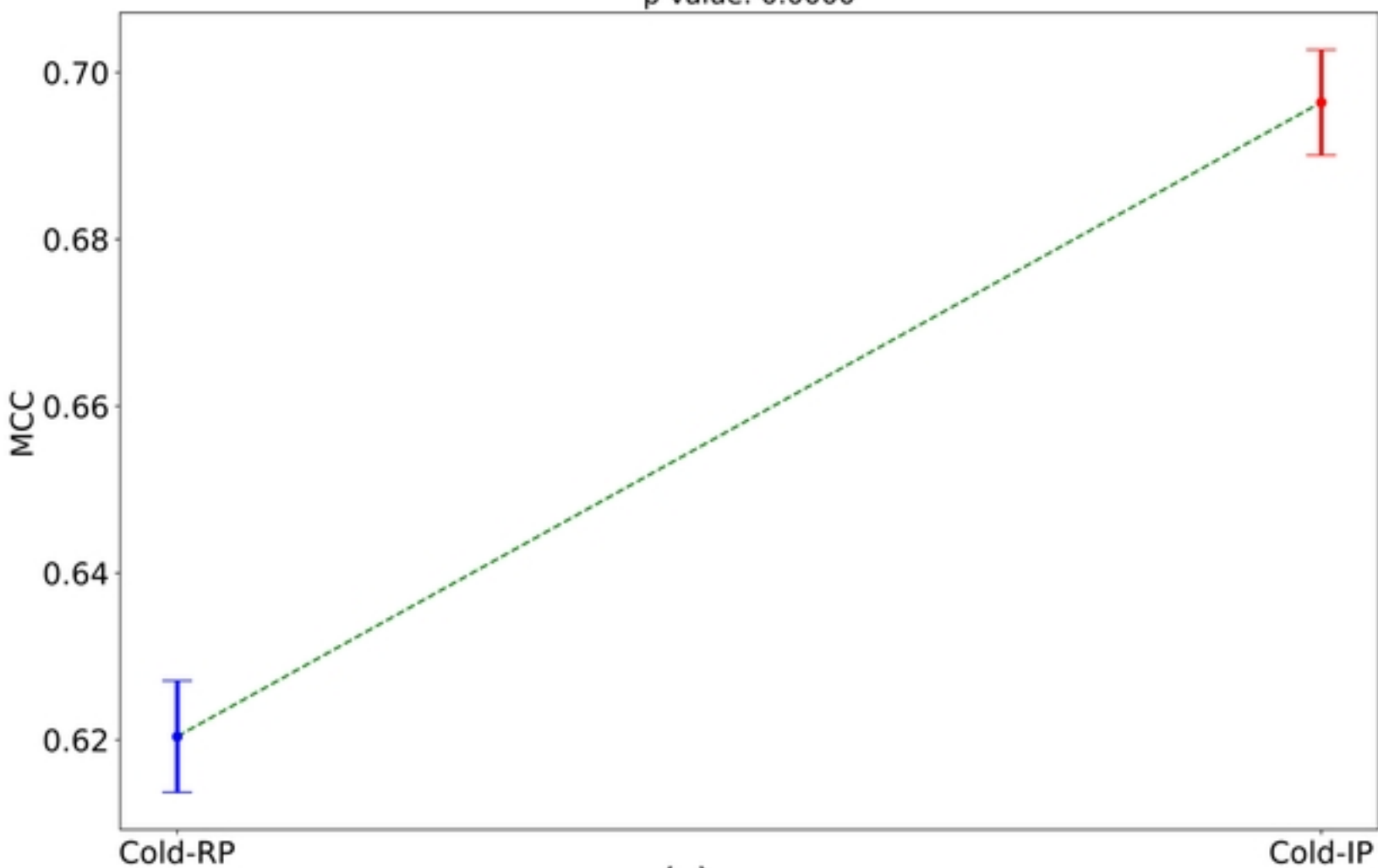


(a)



(b)

MCC Comparison with Confidence Intervals
p-value: 0.0000



(c)

Figure 1

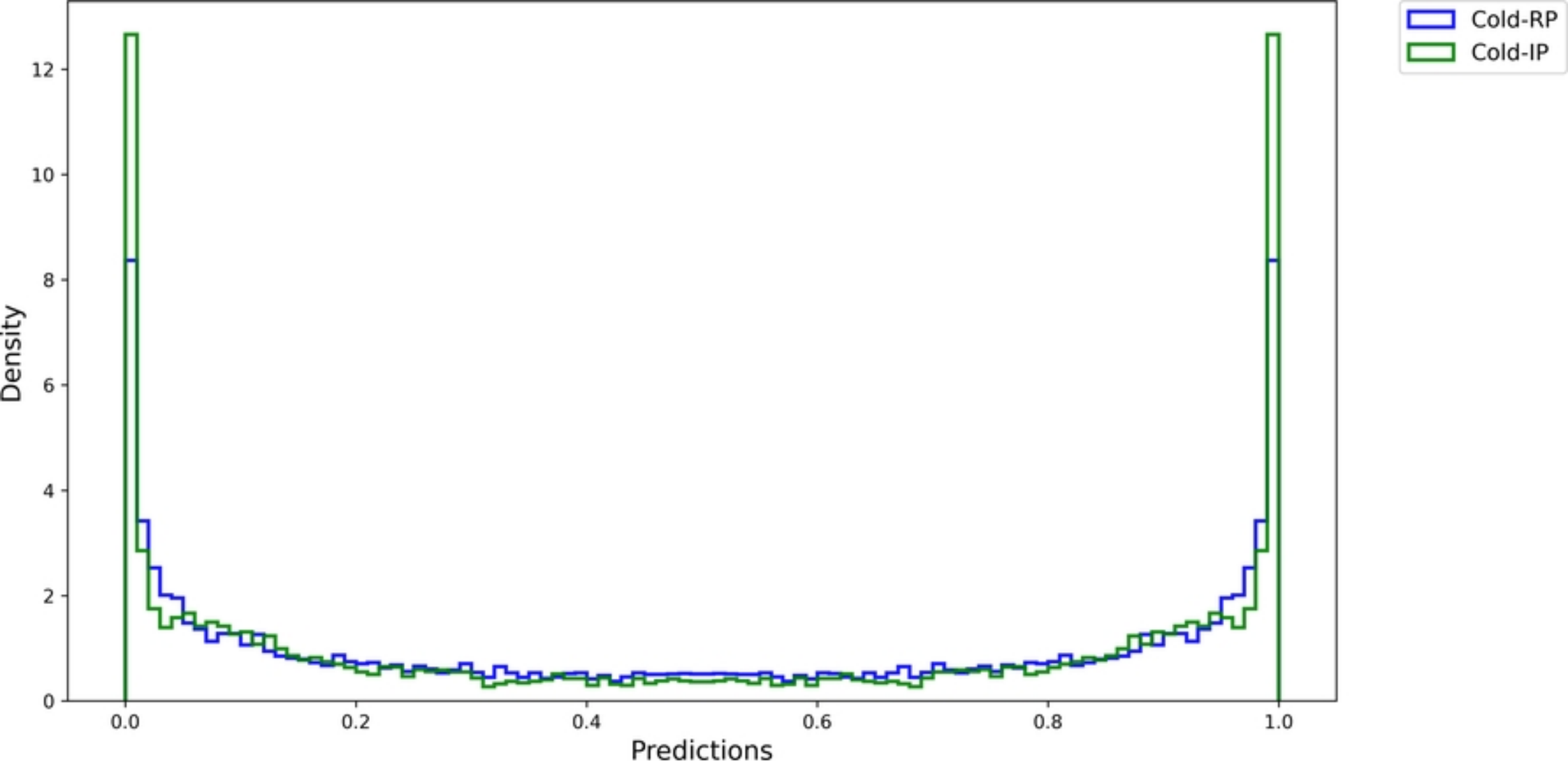
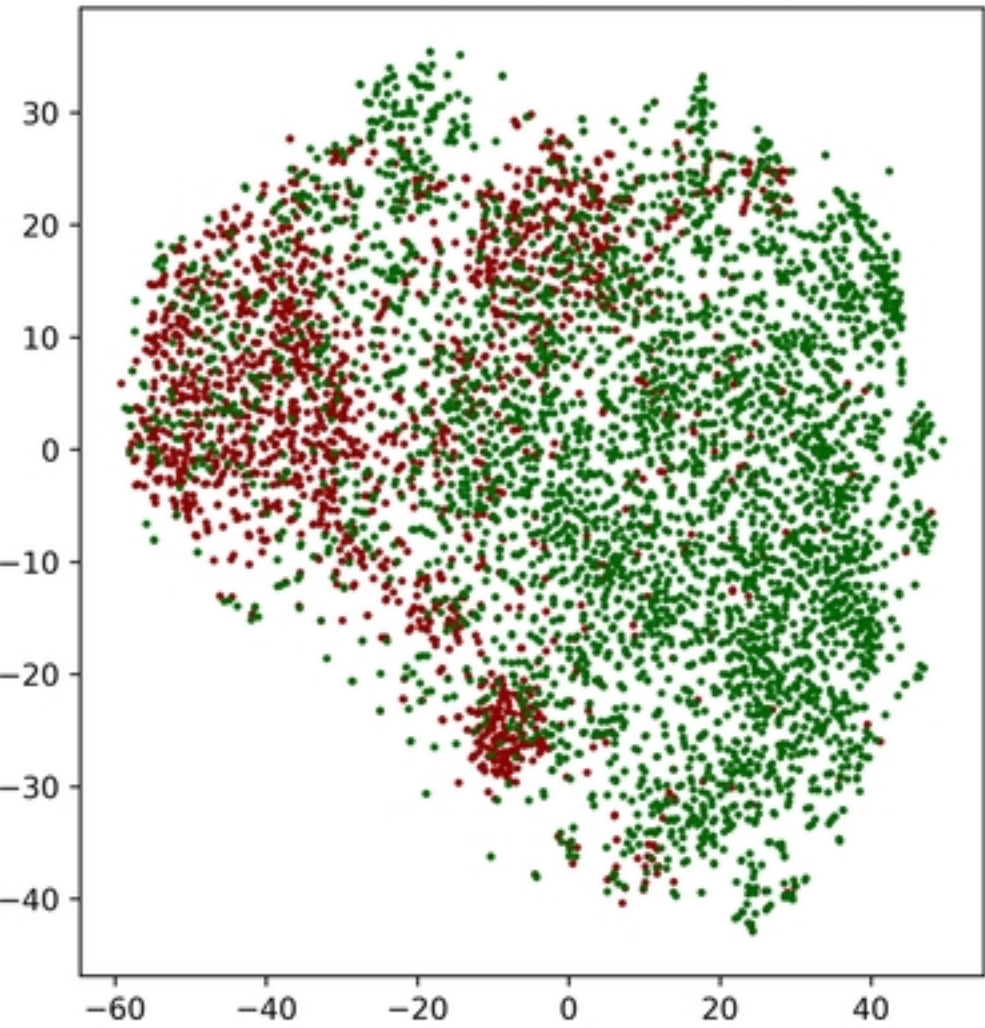


Figure 2

t-SNE of Cold-RP (Perplexity: 50, Learning Rate: 100)



t-SNE of Cold-IP (Perplexity: 50, Learning Rate: 100)

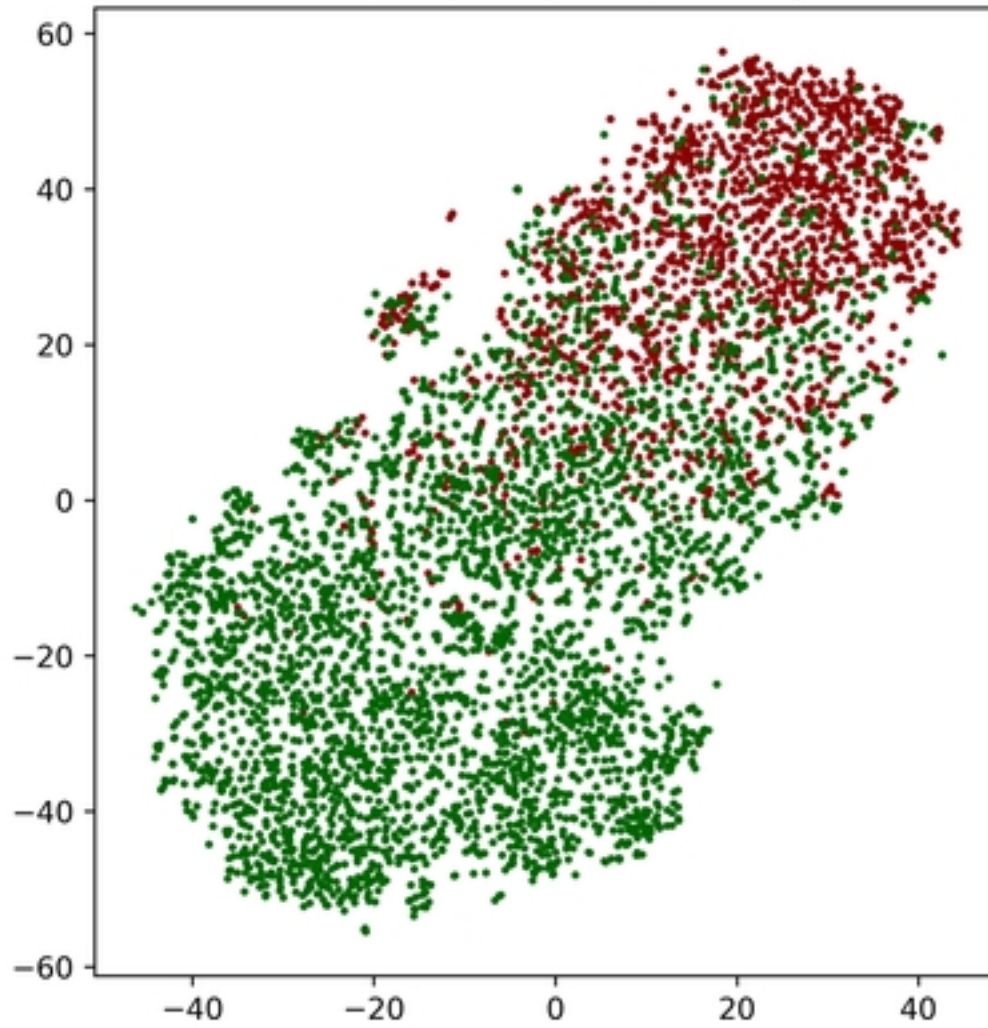


Figure 3

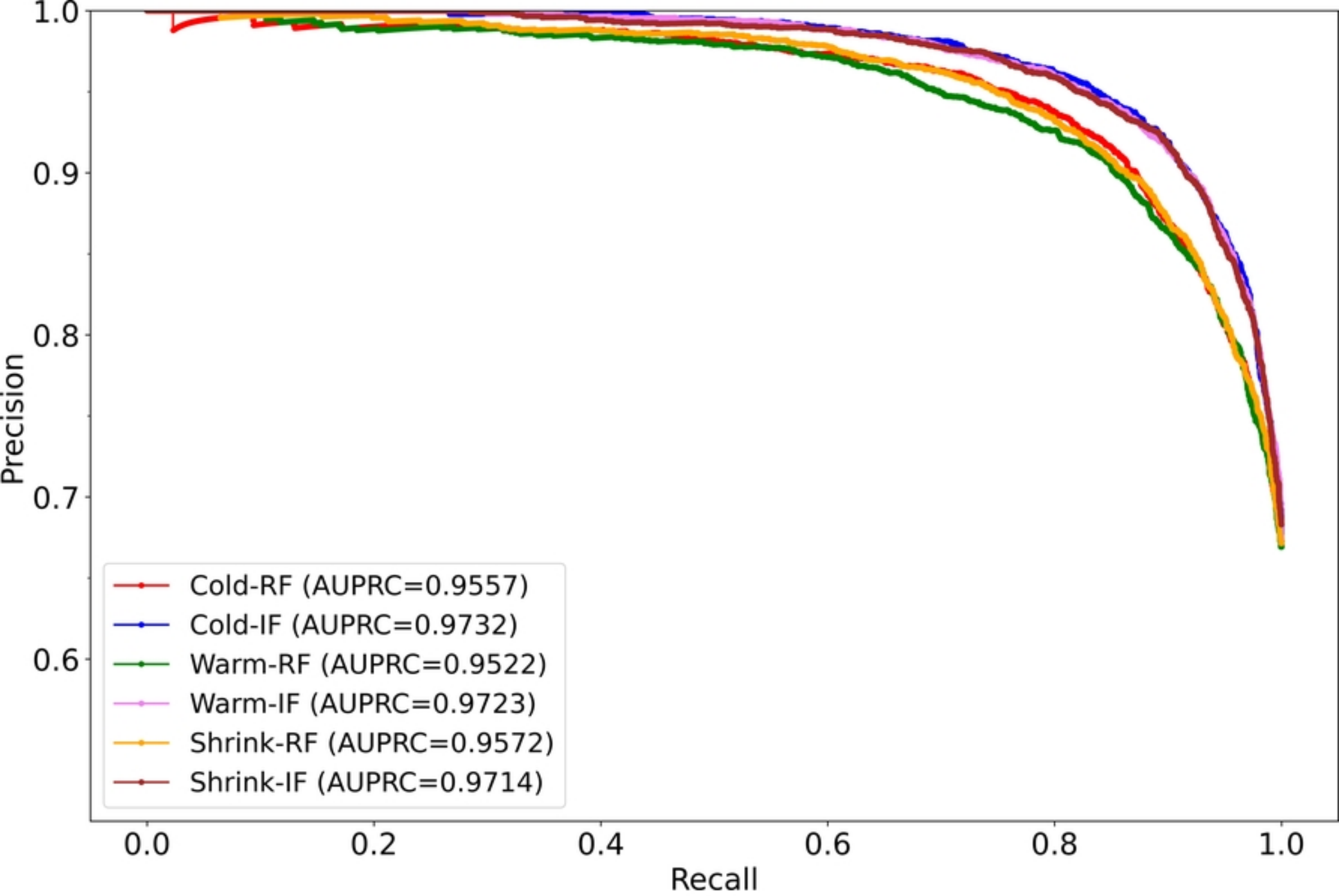
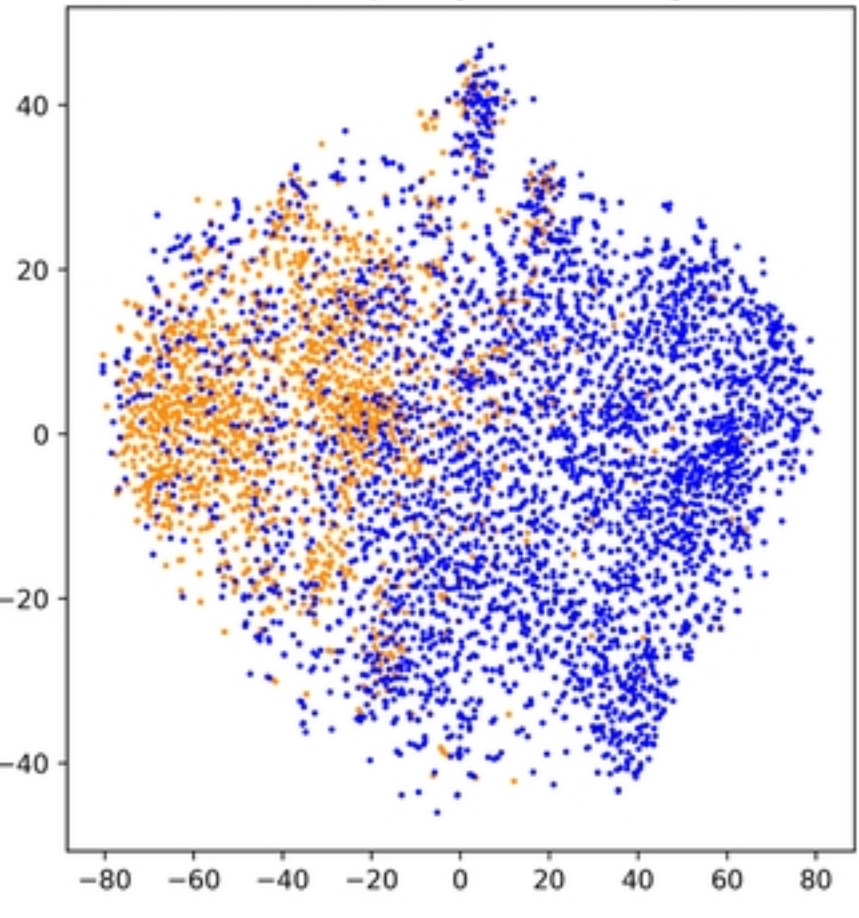
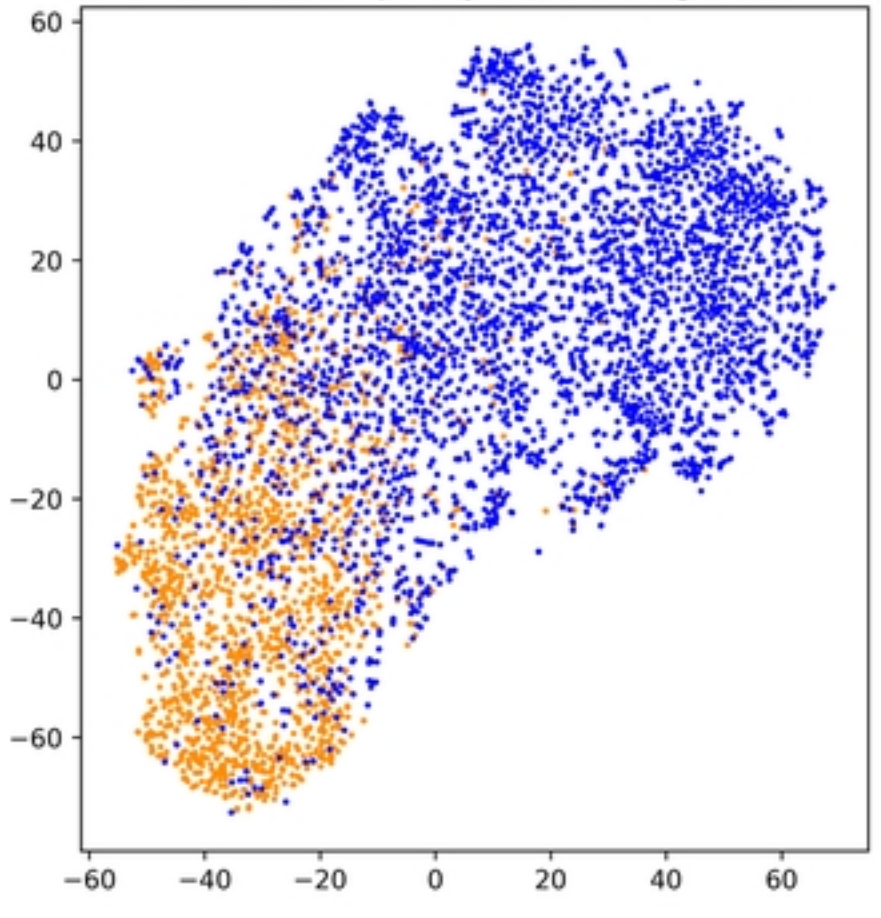


Figure 4

t-SNE of Cold-RF (Perplexity: 40, Learning Rate: 1000)



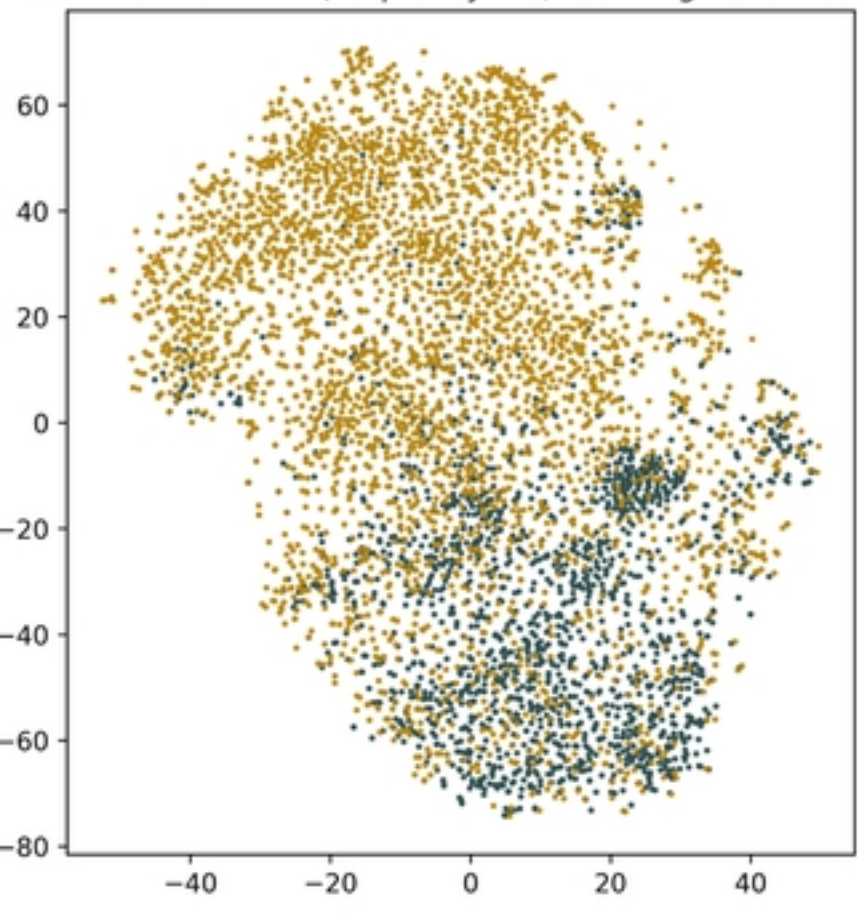
t-SNE of Cold-IF (Perplexity: 40, Learning Rate: 1000)



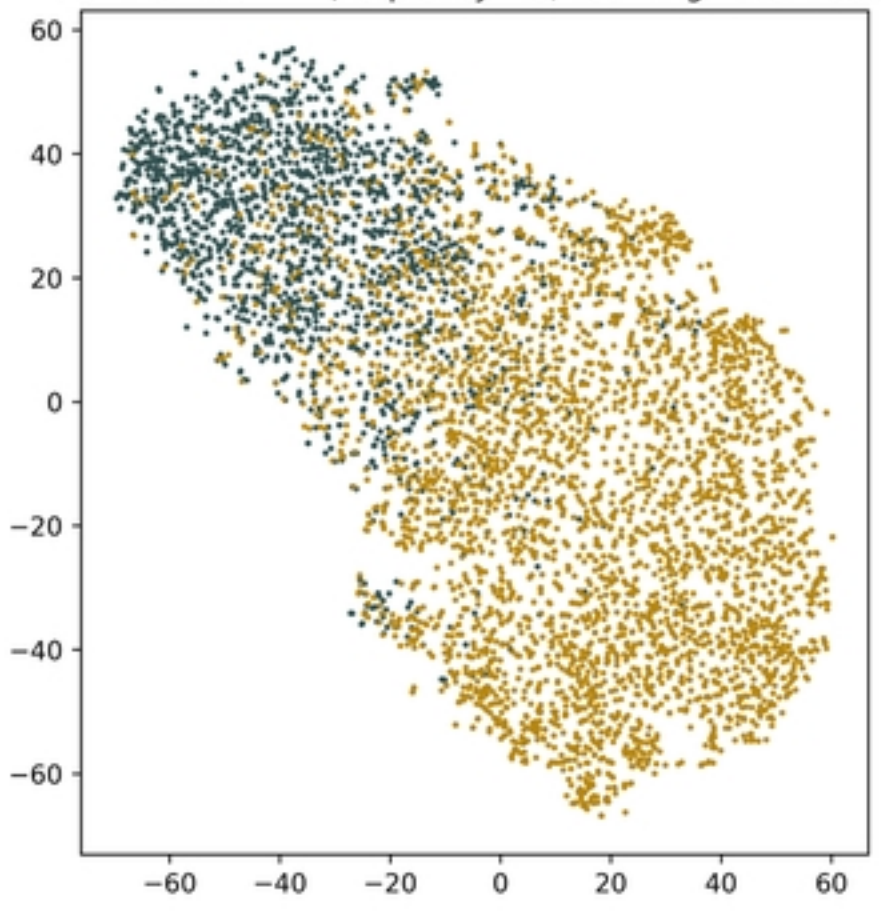
● No Finding
● Abnormal

(a)

t-SNE of Warm-RF (Perplexity: 40, Learning Rate: 1000)



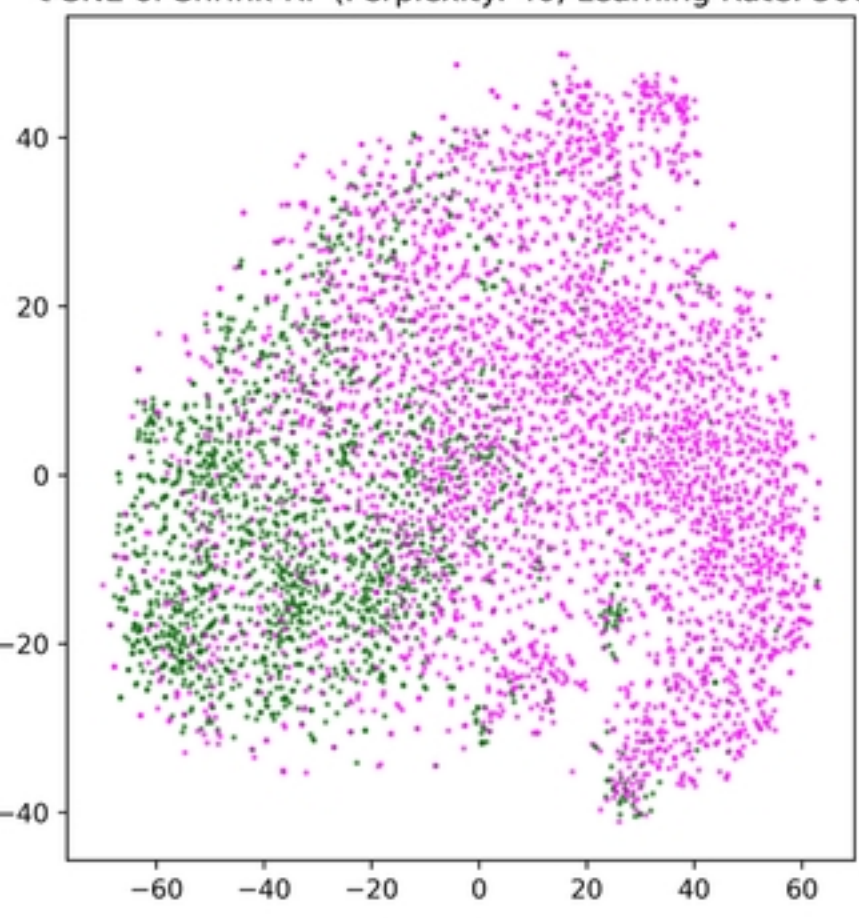
t-SNE of Warm-IF (Perplexity: 40, Learning Rate: 1000)



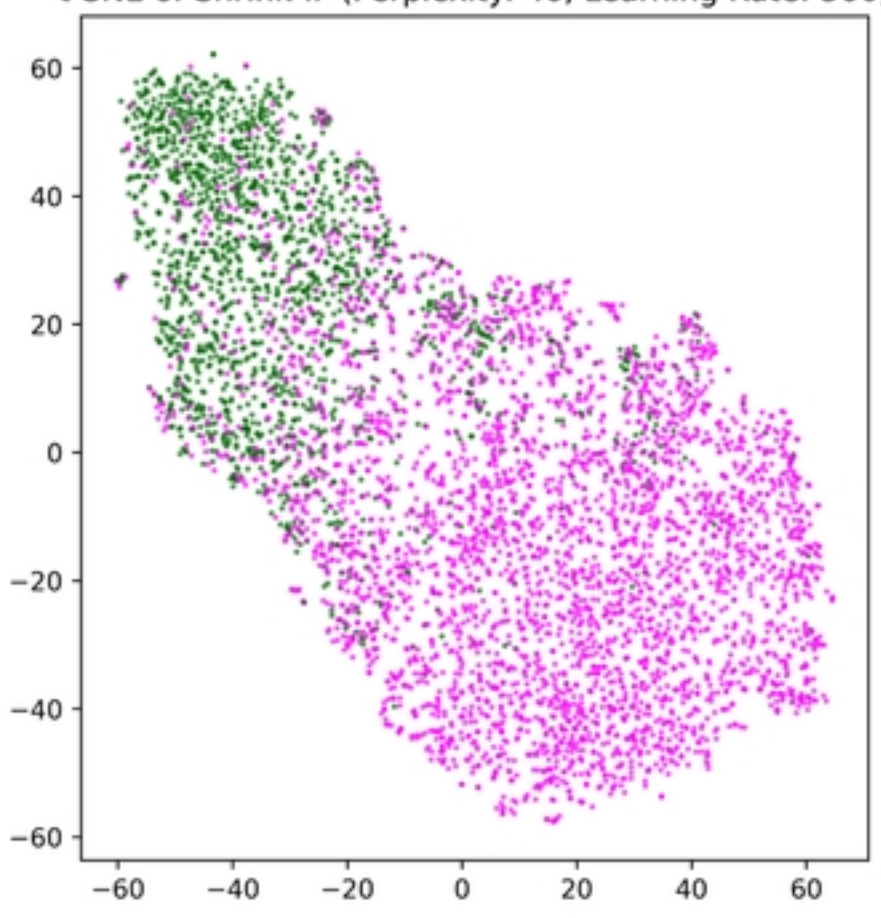
● No Finding
● Abnormal

(b)

t-SNE of Shrink-RF (Perplexity: 40, Learning Rate: 500)



t-SNE of Shrink-IF (Perplexity: 40, Learning Rate: 500)



● No Finding
● Abnormal

(c)

Figure 5

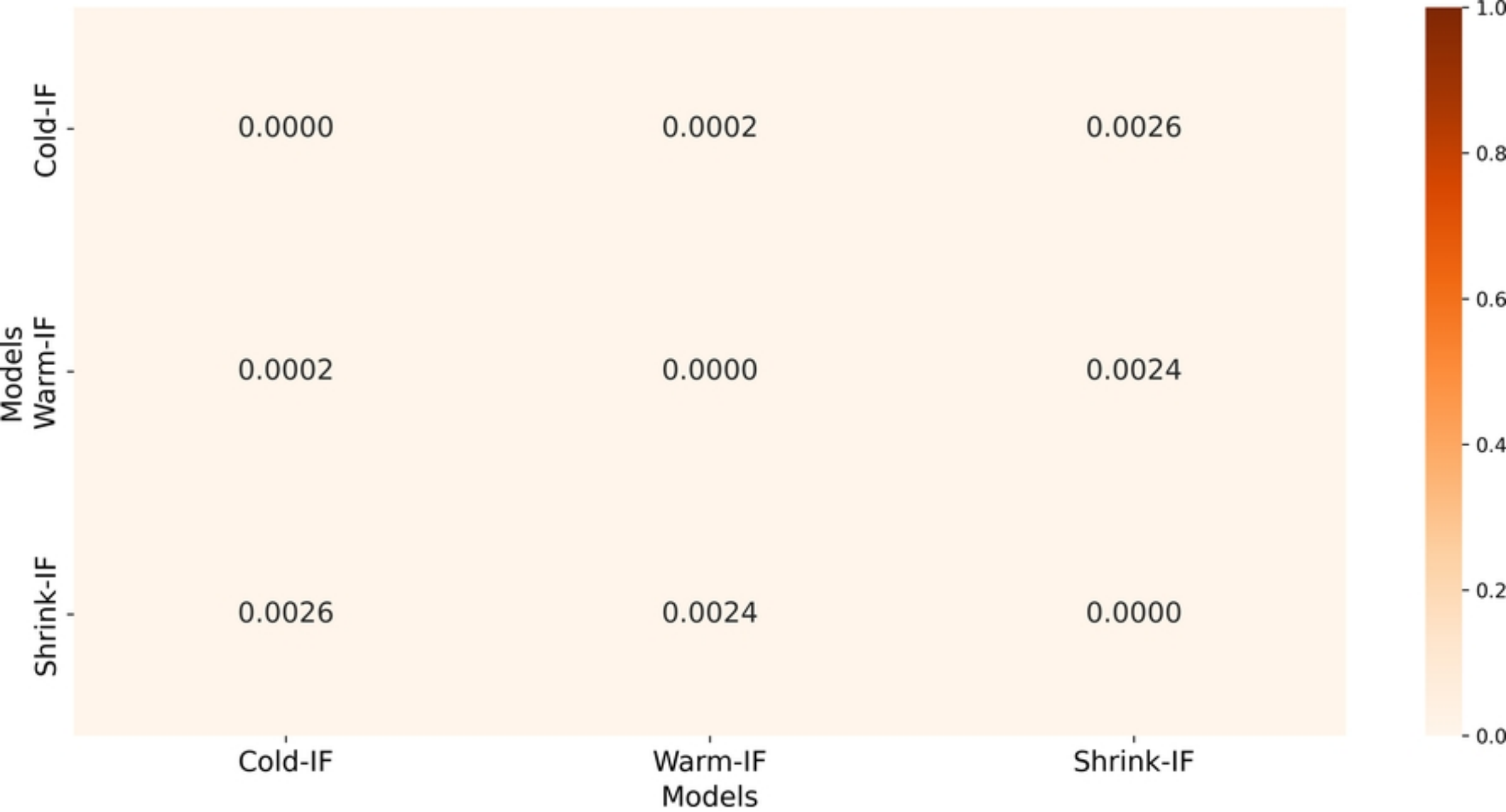


Figure 6

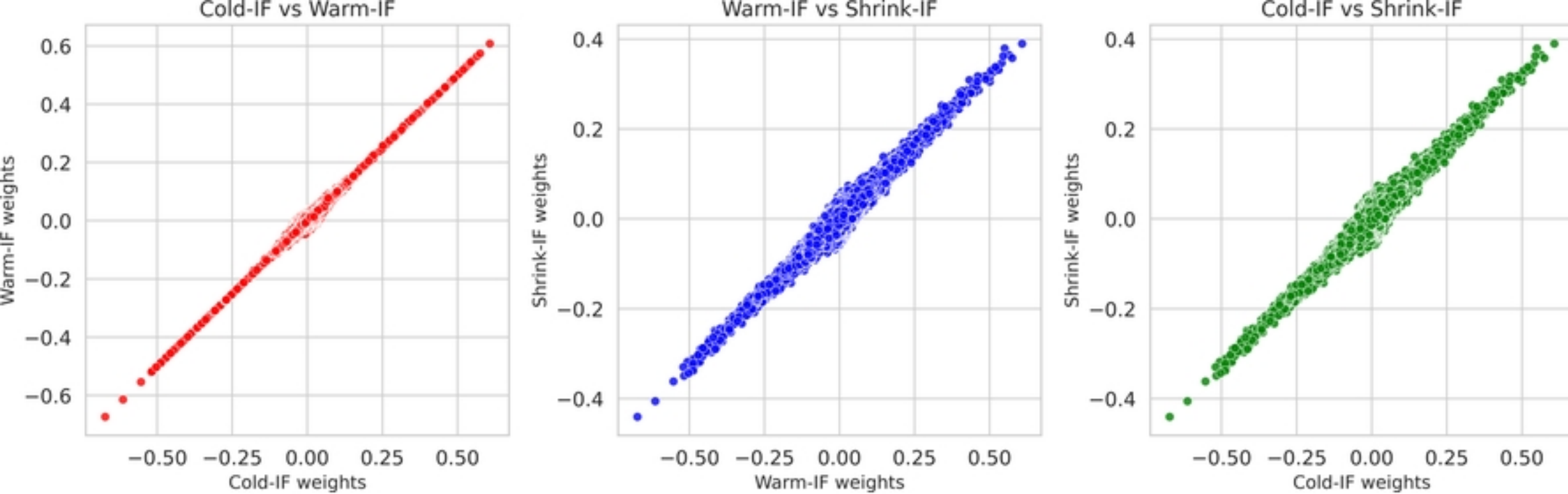
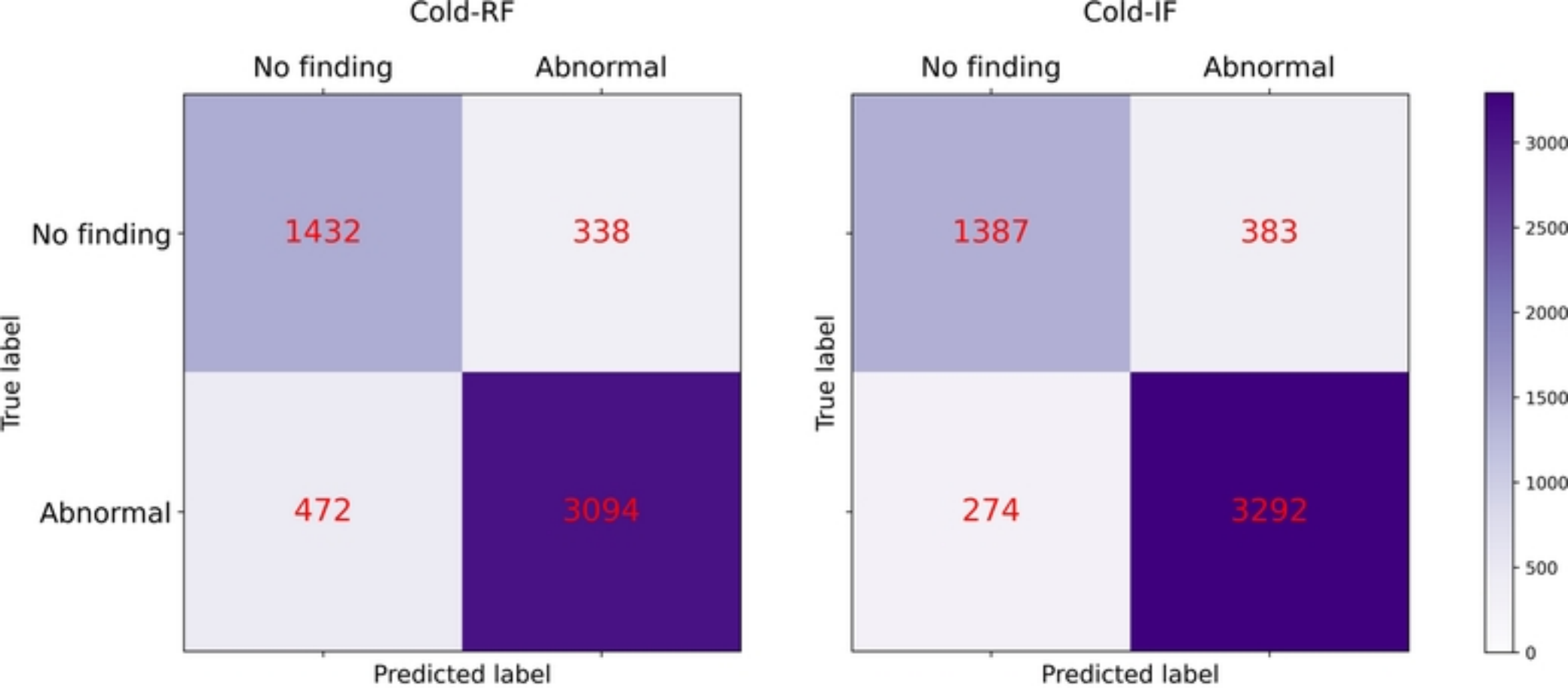
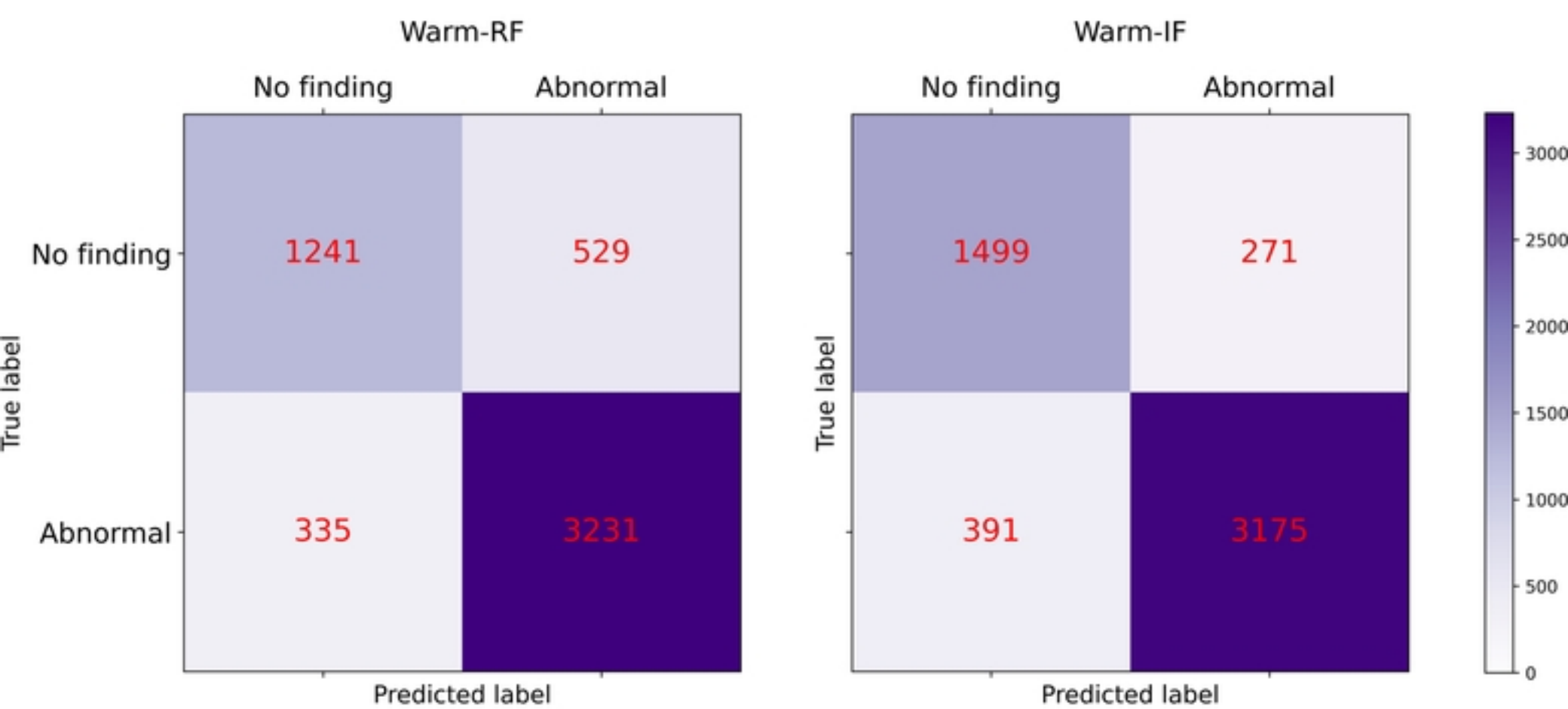


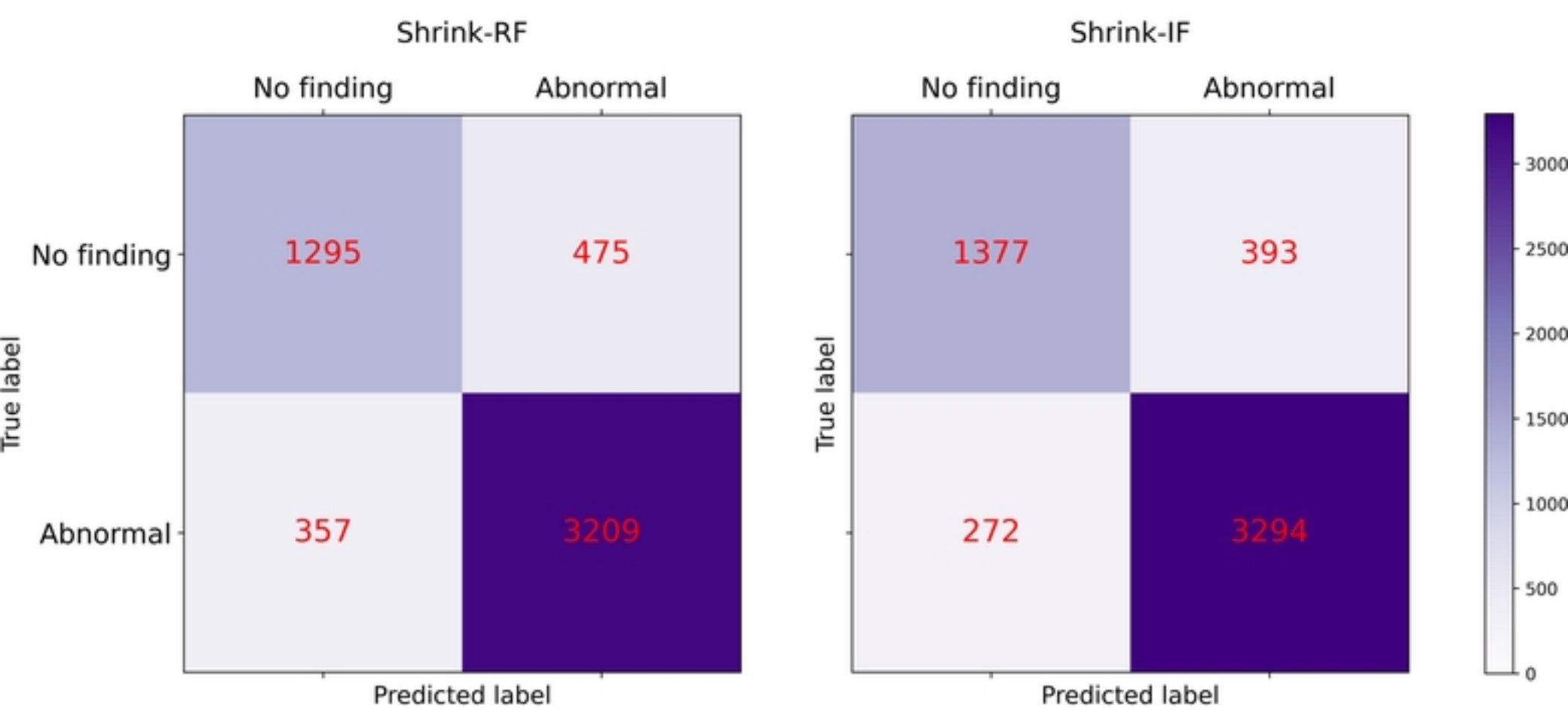
Figure 7



(a)

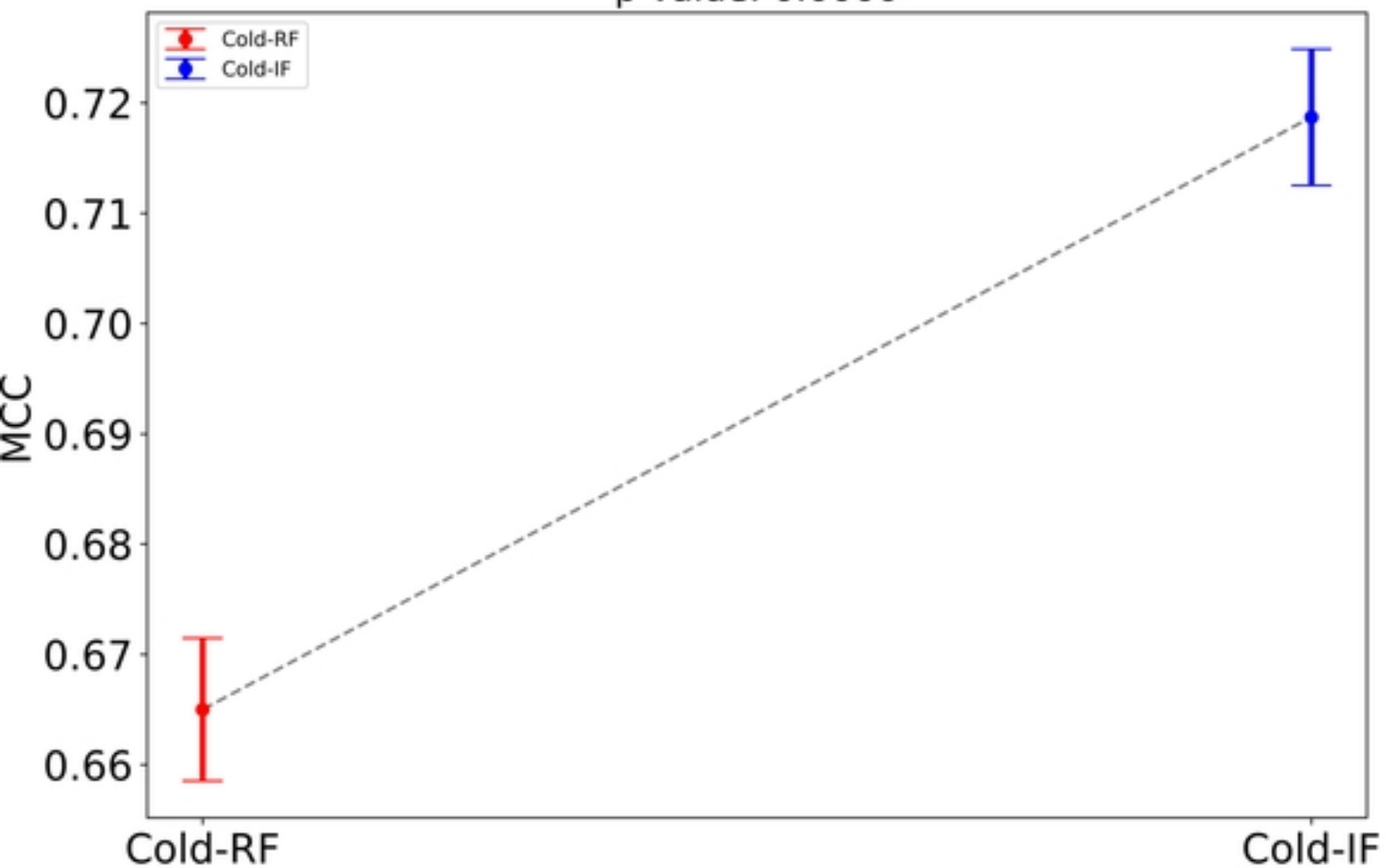


(b)

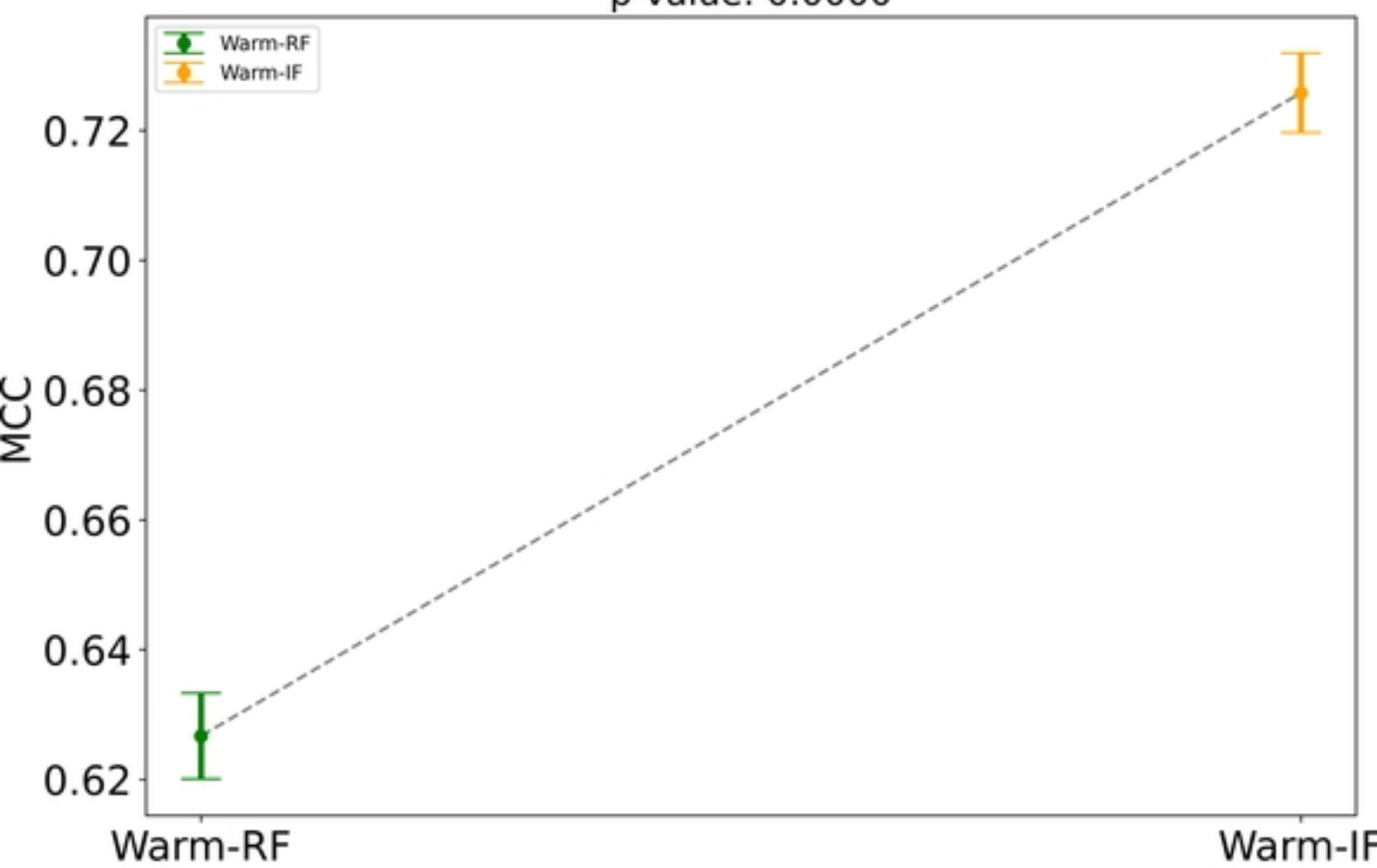


(c)

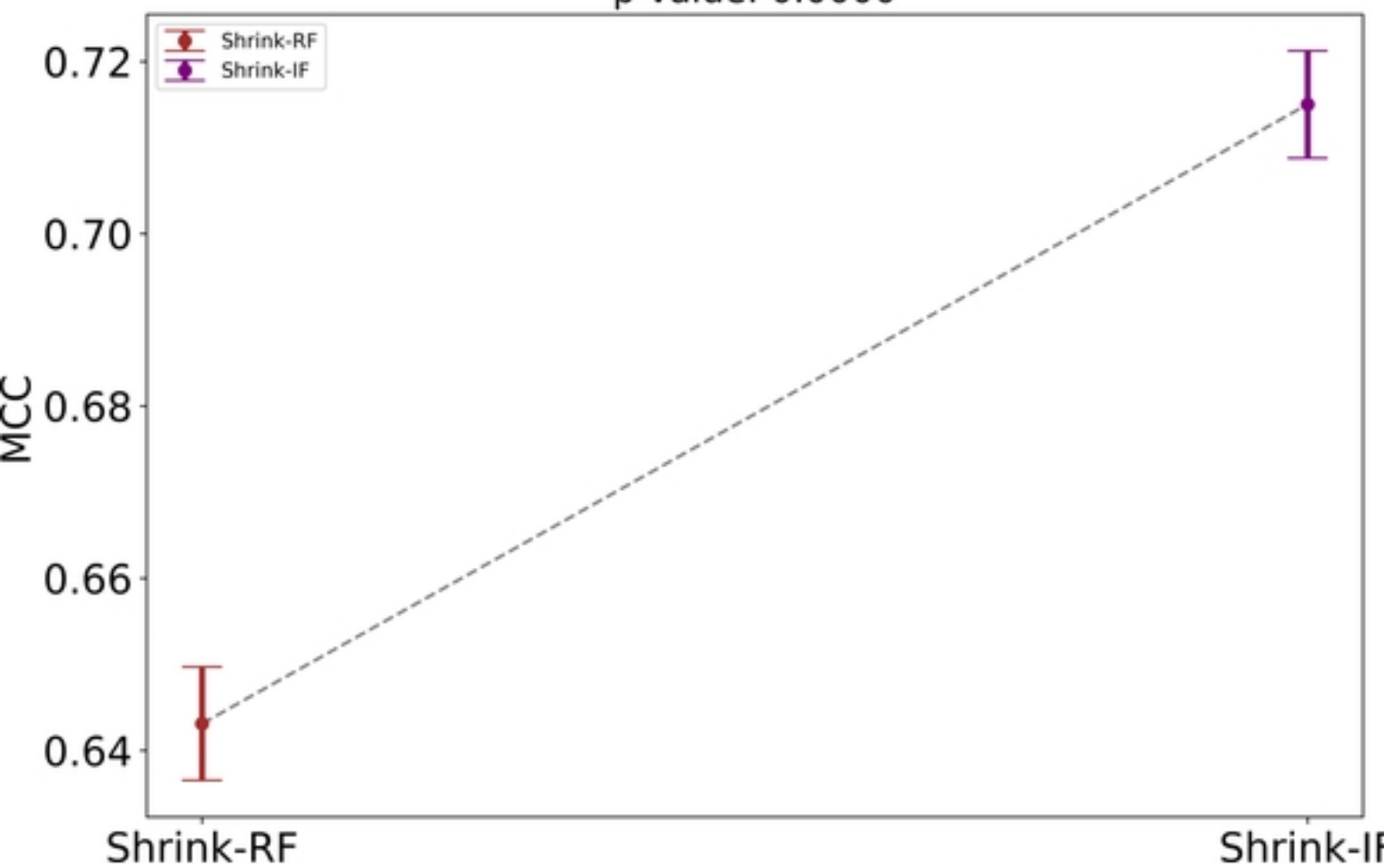
Cold-RF vs Cold-IF MCC Comparison with Confidence Intervals
p-value: 0.0000



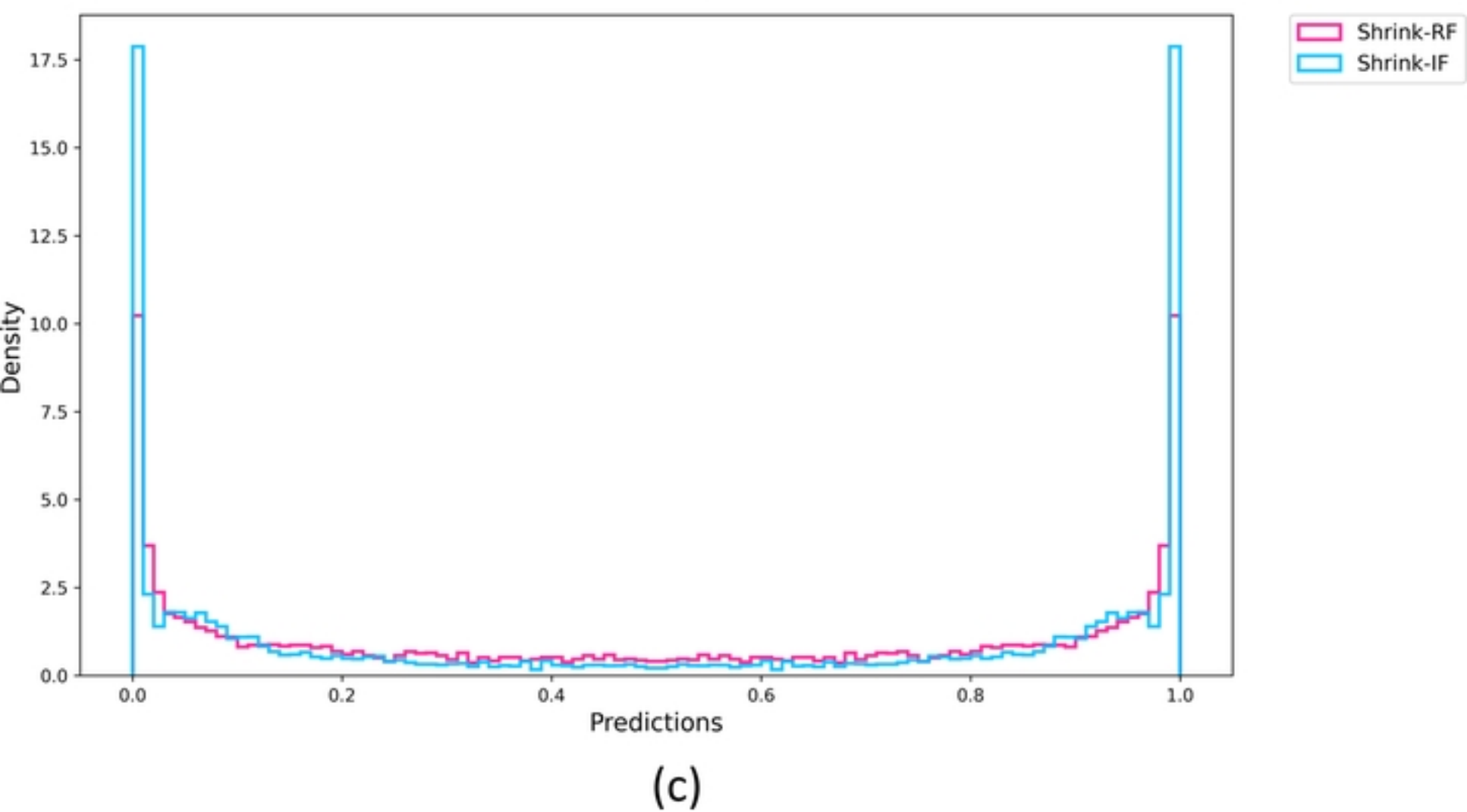
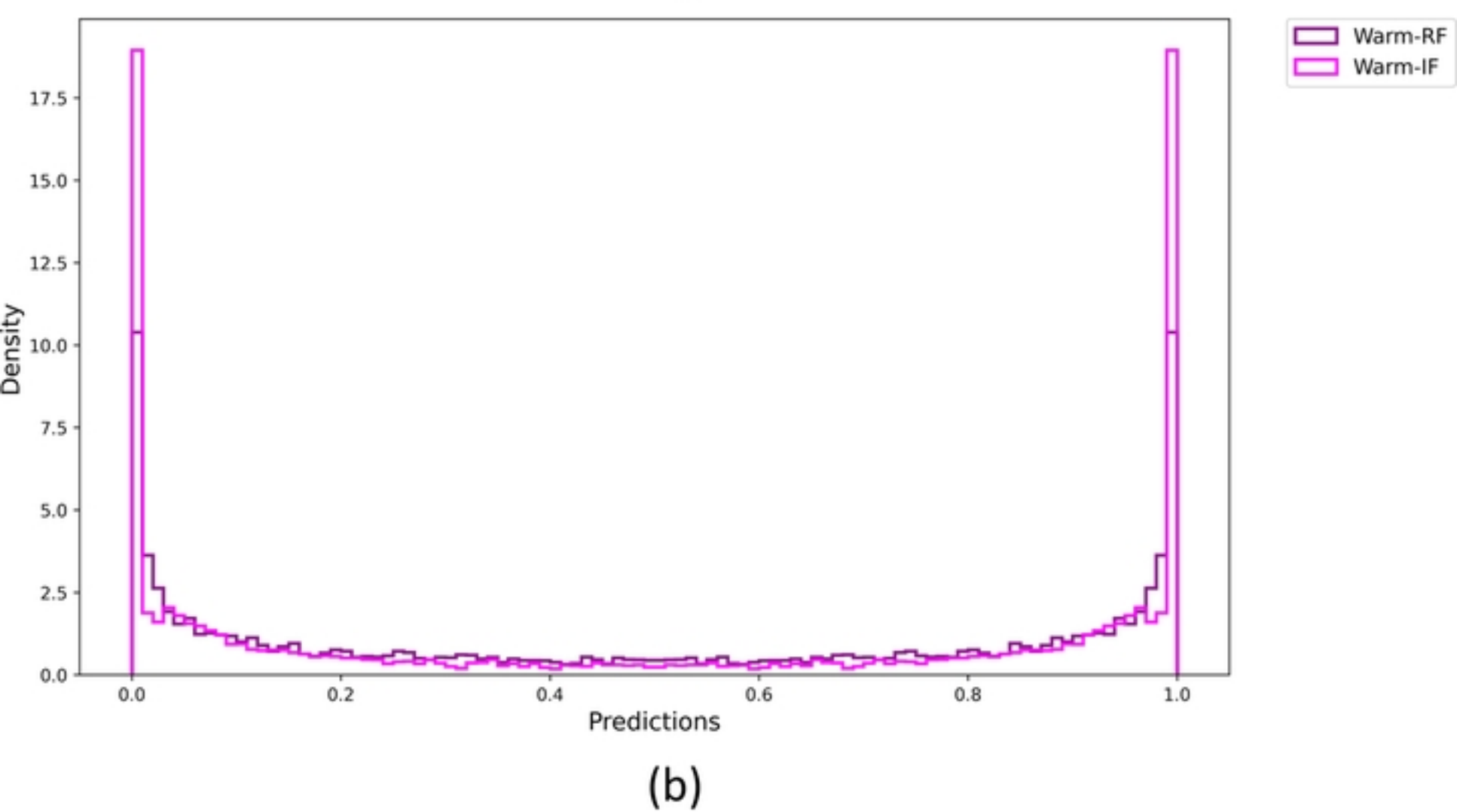
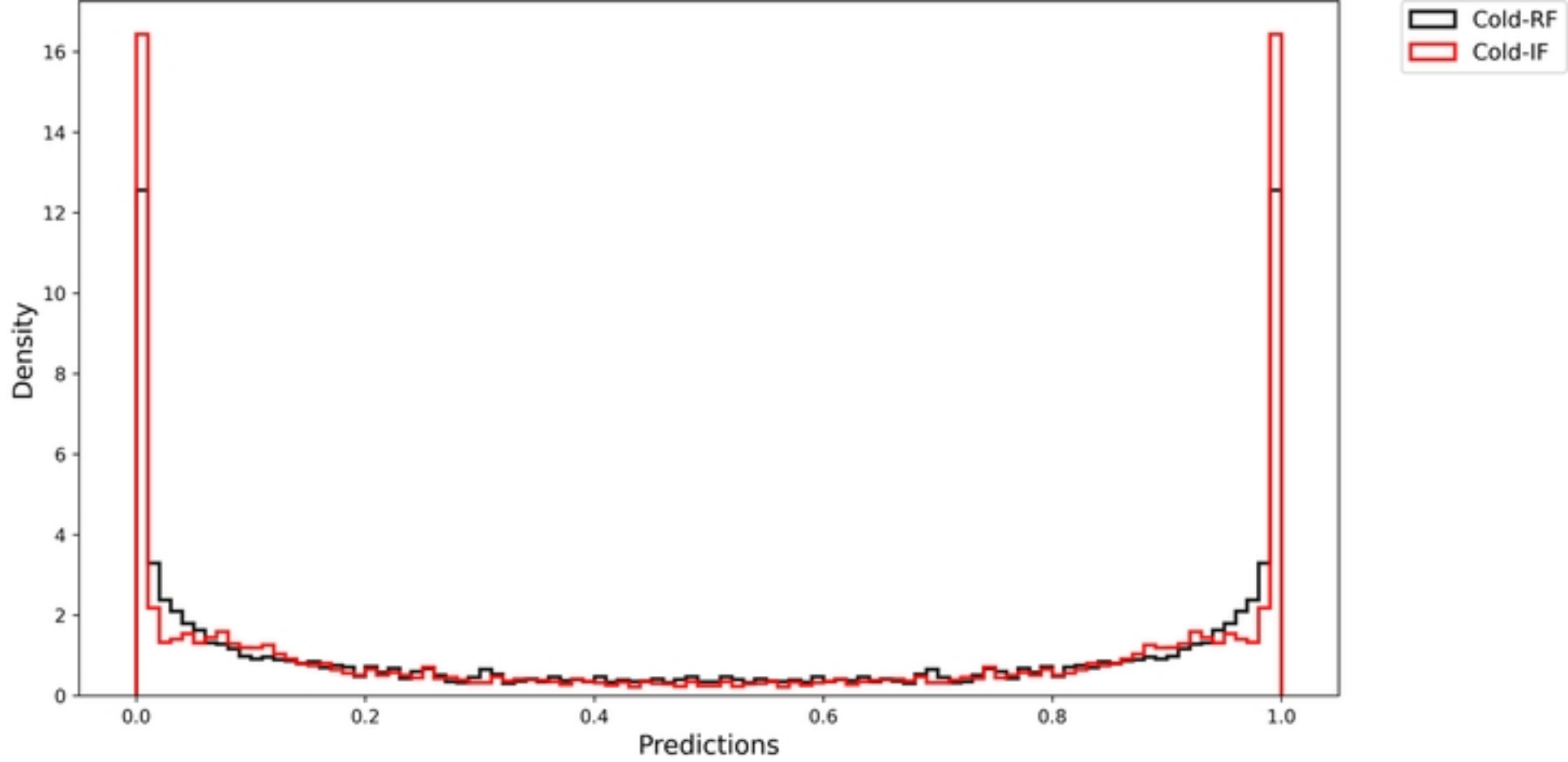
Warm-RF vs Warm-IF MCC Comparison with Confidence Intervals
p-value: 0.0000



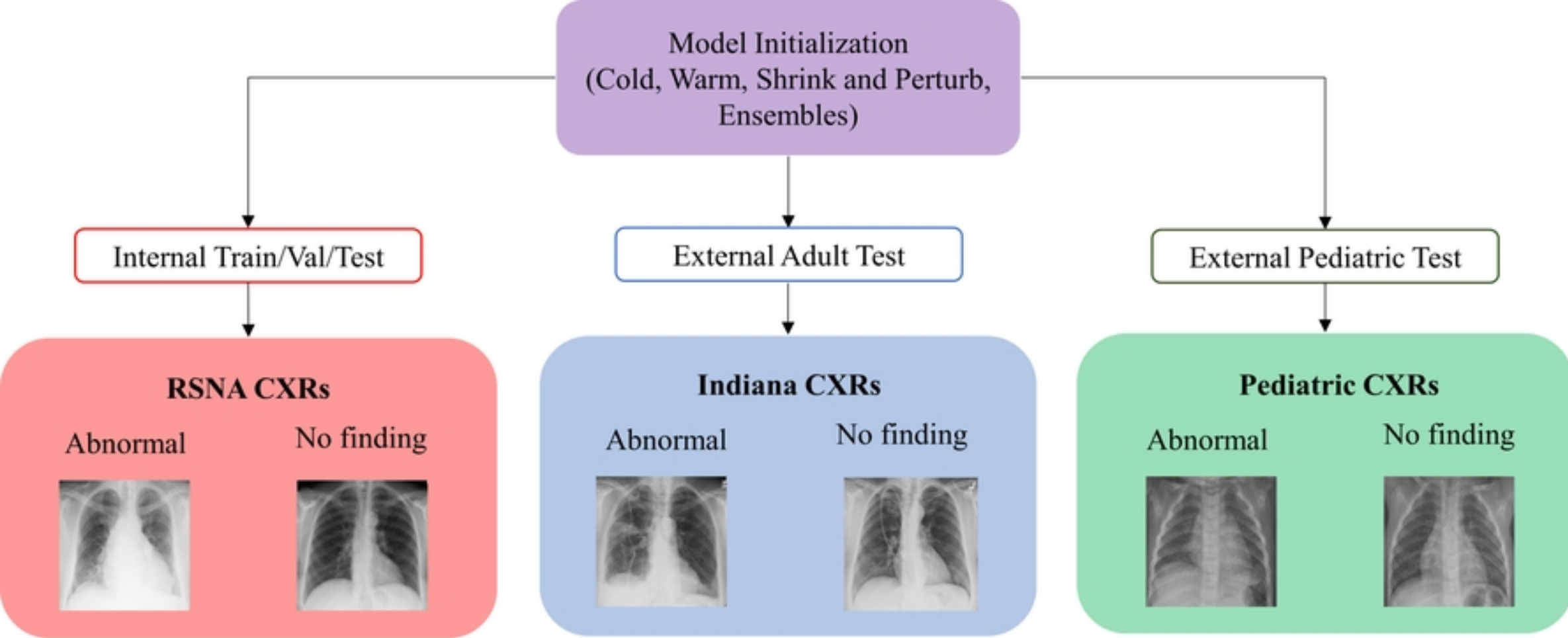
Shrink-RF vs Shrink-IF MCC Comparison with Confidence Intervals
p-value: 0.0000



Supplementary Figure 2



Supplementary Figure 3



Striking Diagram