

TOWARDS DOMAIN-SPECIFIC EXPLAINABLE AI: MODEL INTERPRETATION OF A SKIN IMAGE CLASSIFIER USING A HUMAN APPROACH

(03) EXPERIMENT AND (04) RESULT

ชื่อผู้เขียน: Fabian Stieler, Fabian Rabe, Bernhard Bauer

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

ปีที่พิมพ์: 2021

03 EXPLAINER FOR SKIN IMAGE CLASSIFIER

paragraph 1: การรวมวิธีในการตีความแบบจำลอง (LIME) รวมกับแนวทางทางการแพทย์

paragraph 2: ส่งผลให้คำอธิบายที่ได้สอดคล้องและมีความหมายในทางการแพทย์มากขึ้น

03 EXPLAINER FOR SKIN IMAGE CLASSIFIER

3.1. PERTURBATION DIMENSIONS

MEDICALLY RELEVANT FEATURES

- (B) BORDER
- (C) COLORS

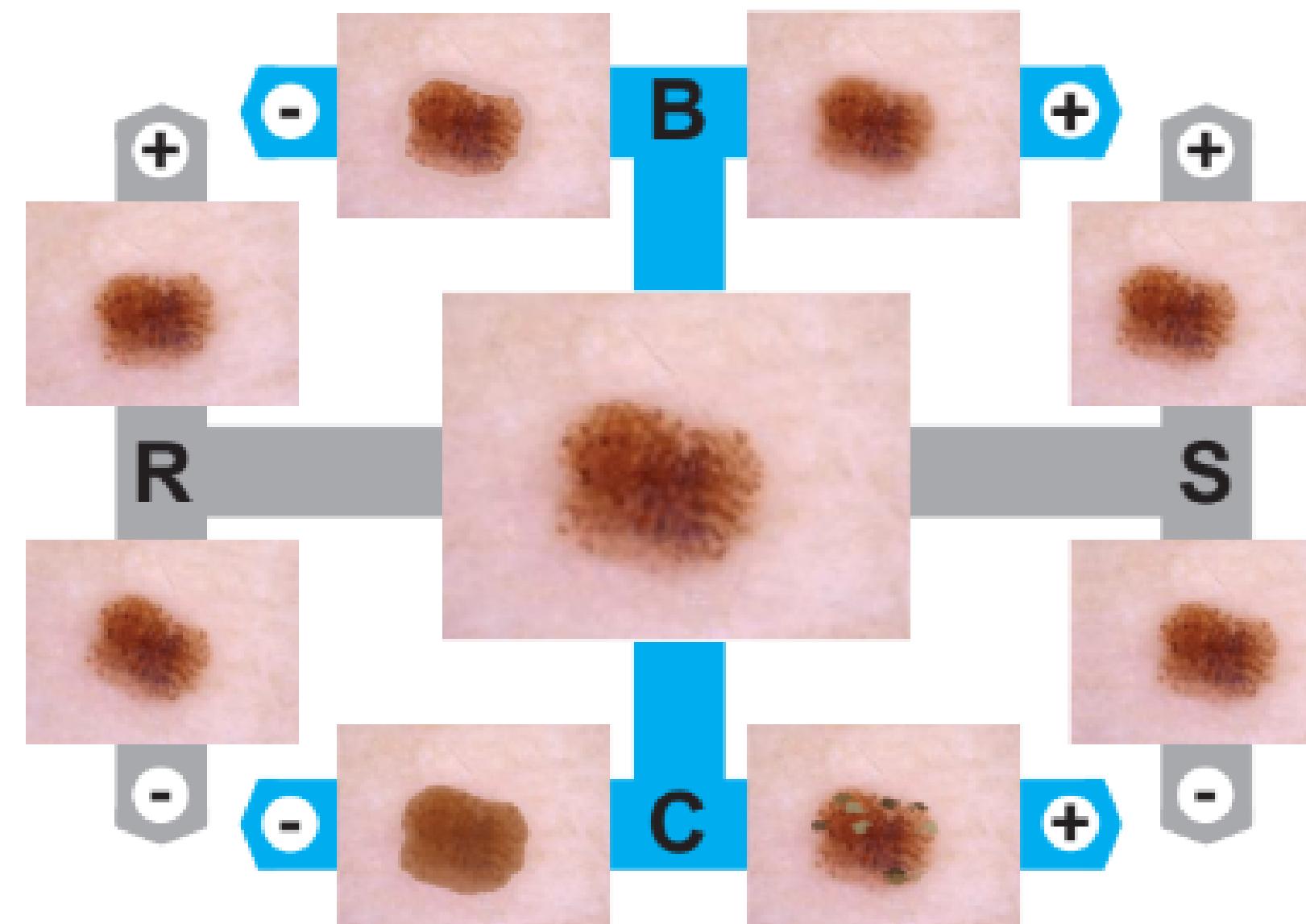
MEDICALLY IRRELEVANT FEATURES*

- (R) ROTATE
- (S) SHIFT

*FONG AND VEDALDI (2017)

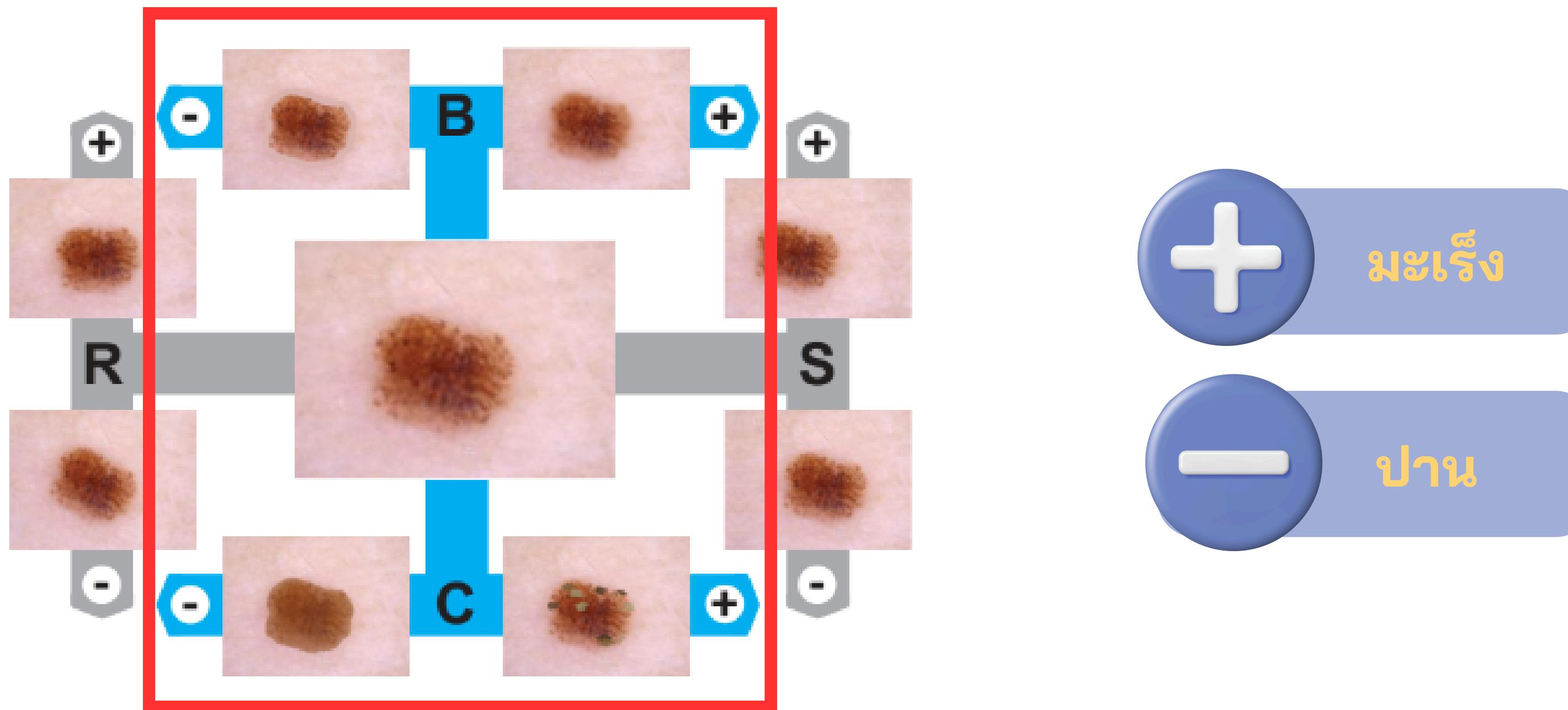
03 EXPLAINER FOR SKIN IMAGE CLASSIFIER

3.1. PERTURBATION DIMENSIONS



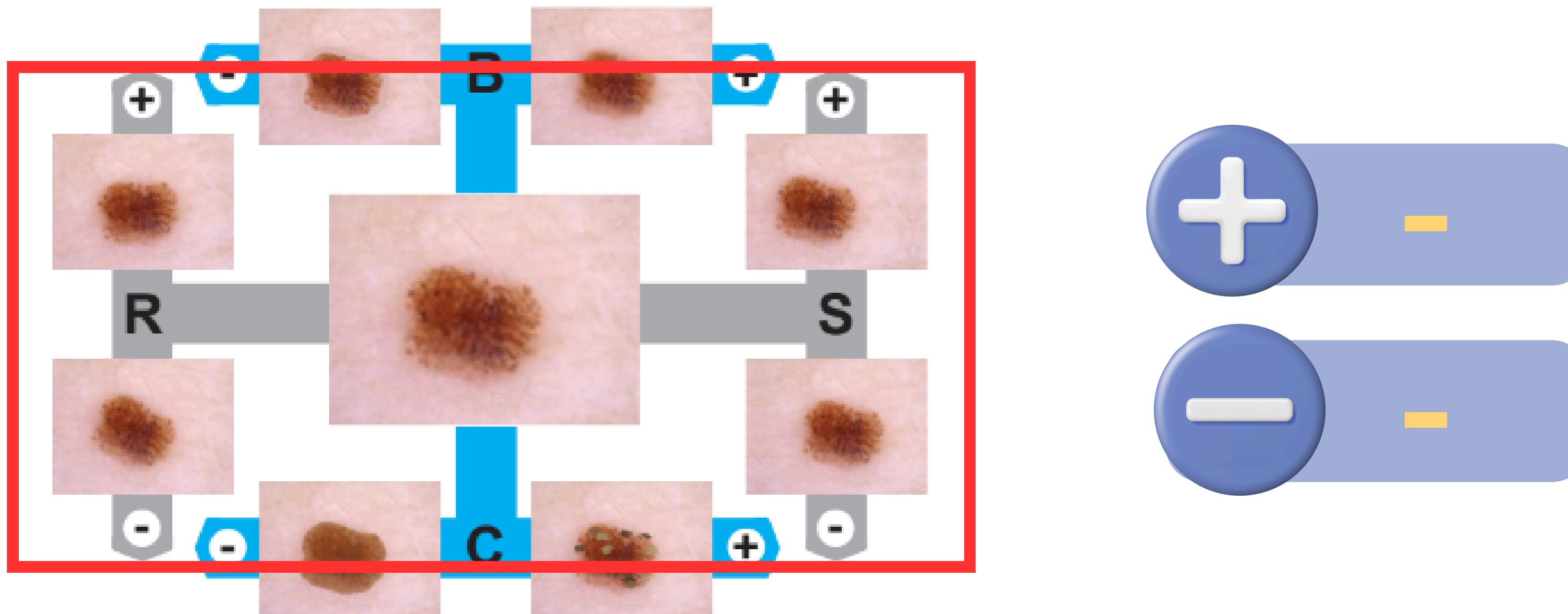
03 EXPLAINER FOR SKIN IMAGE CLASSIFIER

3.1. PERTURBATION DIMENSIONS



03 EXPLAINER FOR SKIN IMAGE CLASSIFIER

3.1. PERTURBATION DIMENSIONS

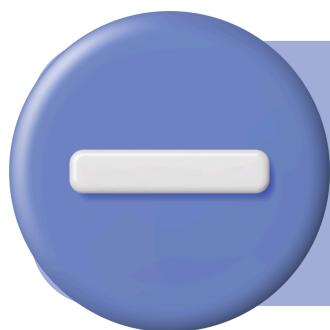


03 EXPLAINER FOR SKIN IMAGE CLASSIFIER

3.1. PERTURBATION DIMENSIONS

MEDICALLY RELEVANT FEATURES

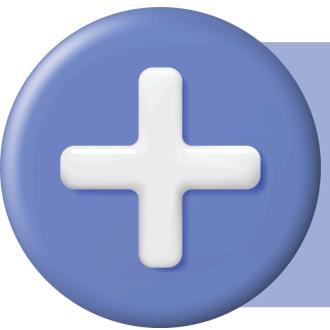
(B) BORDER



ปาน

เพิ่มขอบเขตระหว่างผลกับผิวนังให้มีการแบ่งแยกกันชัดเจน

Original image



มะเร็ง

ลดขอบเขตระหว่างผลกับผิวนังให้มีการแบ่งแยกกันน้อยลง

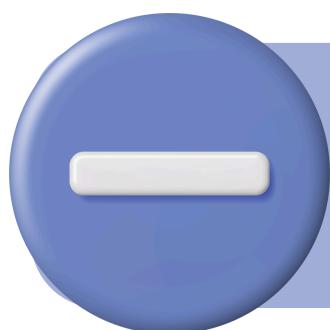


03 EXPLAINER FOR SKIN IMAGE CLASSIFIER

3.1. PERTURBATION DIMENSIONS

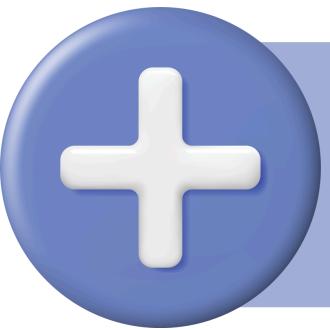
MEDICALLY RELEVANT FEATURES

(C) COLORS



ปน

เพิ่มความสมำเสมอของลีกायในพื้นที่ของขอบเขตแล้ว



มะเร็ง

ลดความสมำเสมอของลีกायในพื้นที่ของขอบเขตแล้ว

Original image

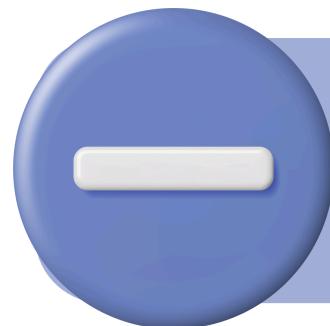


03 EXPLAINER FOR SKIN IMAGE CLASSIFIER

3.1. PERTURBATION DIMENSIONS

MEDICALLY IRRELEVANT FEATURES

(R) ROTATE



หมุนแพลงไปทางขวา มีอัตราของศาที่กำหนด

Original image



หมุนแพลงไปทางซ้าย มีอัตราของศาที่กำหนด

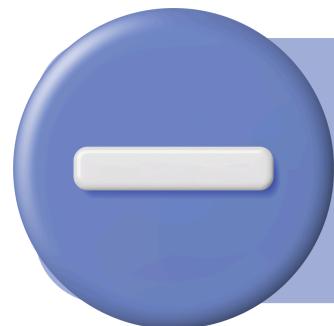


03 EXPLAINER FOR SKIN IMAGE CLASSIFIER

3.1. PERTURBATION DIMENSIONS

MEDICALLY IRRELEVANT FEATURES

(S) SHIFT



เลื่อนแพลไปทางขวา มีอ

Original image



เลื่อนแพลไปทางซ้าย มีอ



03 EXPLAINER FOR SKIN IMAGE CLASSIFIER

3.2. HYPOTHESES

MEDICALLY RELEVANT FEATURES

- A** การปรับมิติภาพไปทางบวก (ปรับให้ภาพดูเป็นมะเร็งมากขึ้น)
- B** การปรับมิติภาพไปทางลบ (ปรับให้ภาพดูเป็นปานมากขึ้น)

MEDICALLY IRRELEVANT FEATURES

- C** การปรับมิติภาพไม่ควรมีผลกระทบต่อการทำนายแบบจำลอง

03 EXPLAINER FOR SKIN IMAGE CLASSIFIER

3.2. HYPOTHESES

สมมติฐาน: Class: nevus

MEDICALLY RELEVANT FEATURES

- A**
 - A₁ ค่าการทำนายในปานจะลดลงใน **Positive perturbation**
 - A₀ ค่าการทำนายในปานจะเพิ่มขึ้นหรือไม่เปลี่ยนแปลง ใน **Positive perturbation**

- B**
 - B₁ ค่าการทำนายในปานจะเพิ่มขึ้นใน **Negative perturbation**
 - B₀ ค่าการทำนายในปานจะลดลงหรือไม่เปลี่ยนแปลง ใน **Negative perturbation**

MEDICALLY IRRELEVANT FEATURES

- C**
 - C₁ การรบกวนมิติที่ไม่ใช่ทางการแพทย์ไม่ส่งผลกระทบต่อการทำนาย
 - C₀ การรบกวนมิติที่ไม่ใช่ทางการแพทย์ส่งผลกระทบต่อการทำนายอย่างมีนัยสำคัญ

03 EXPLAINER FOR SKIN IMAGE CLASSIFIER

3.2. HYPOTHESES

สมมติฐาน: Class: Melanoma

MEDICALLY RELEVANT FEATURES

- A** **A₁** ค่าการทำนายในมะเร็งผิวหนังจะเพิ่มขึ้น
หรือไม่เปลี่ยนแปลง ใน Positive perturbation
- A₀** ค่าการทำนายในมะเร็งผิวหนังจะลดลง ใน Positive perturbation

- B** **B₁** ค่าการทำนายในมะเร็งผิวหนังจะลดลงหรือ
ไม่เปลี่ยนแปลงใน Negative perturbation
- B₀** ค่าการทำนายในมะเร็งผิวหนังจะเพิ่มขึ้นใน
Negative perturbation

MEDICALLY IRRELEVANT FEATURES

- C** **C₁** การรับกวนมิติที่ไม่ใช่ทางการแพทย์ไม่ส่งผลกระทบต่อการทำนาย
- C₀** การรับกวนมิติที่ไม่ใช่ทางการแพทย์ส่งผลกระทบต่อการทำนายอย่างมีนัยสำคัญ

03 EXPLAINER FOR SKIN IMAGE CLASSIFIER

3.3. EXPERIMENTAL SETUP

MODEL MOBILENET

7,818 SAMPLES (HAM10000 DATASET)

6,705 NEVUS SAMPLES

1,113 MELANOMA SAMPLES

TRAIN/TEST: 80/20 (6,257/1,561)

03 EXPLAINER FOR SKIN IMAGE CLASSIFIER

3.3. EXPERIMENTAL SETUP

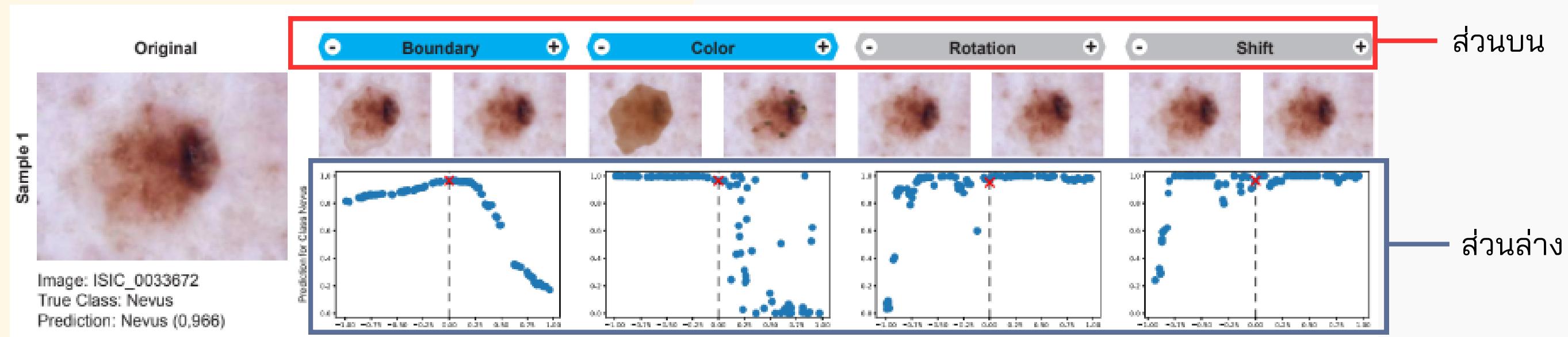
	Nevus	Melanoma	Total
Number of Samples	1,354	216	1,561
True Positives	1,150	144	1,294
False Positives	203	72	275
F_1 -Score	≈ 0.91	≈ 0.57	$\approx 0.74^*$

Table 1. Evaluation results of the classifier. To ensure that class imbalances have no influence, * 'macro' is specified as F_1 average.

04 EMPIRICAL RESULTS

Paragraph 1 : (เกณฑ์การเลือกตัวอย่าง)

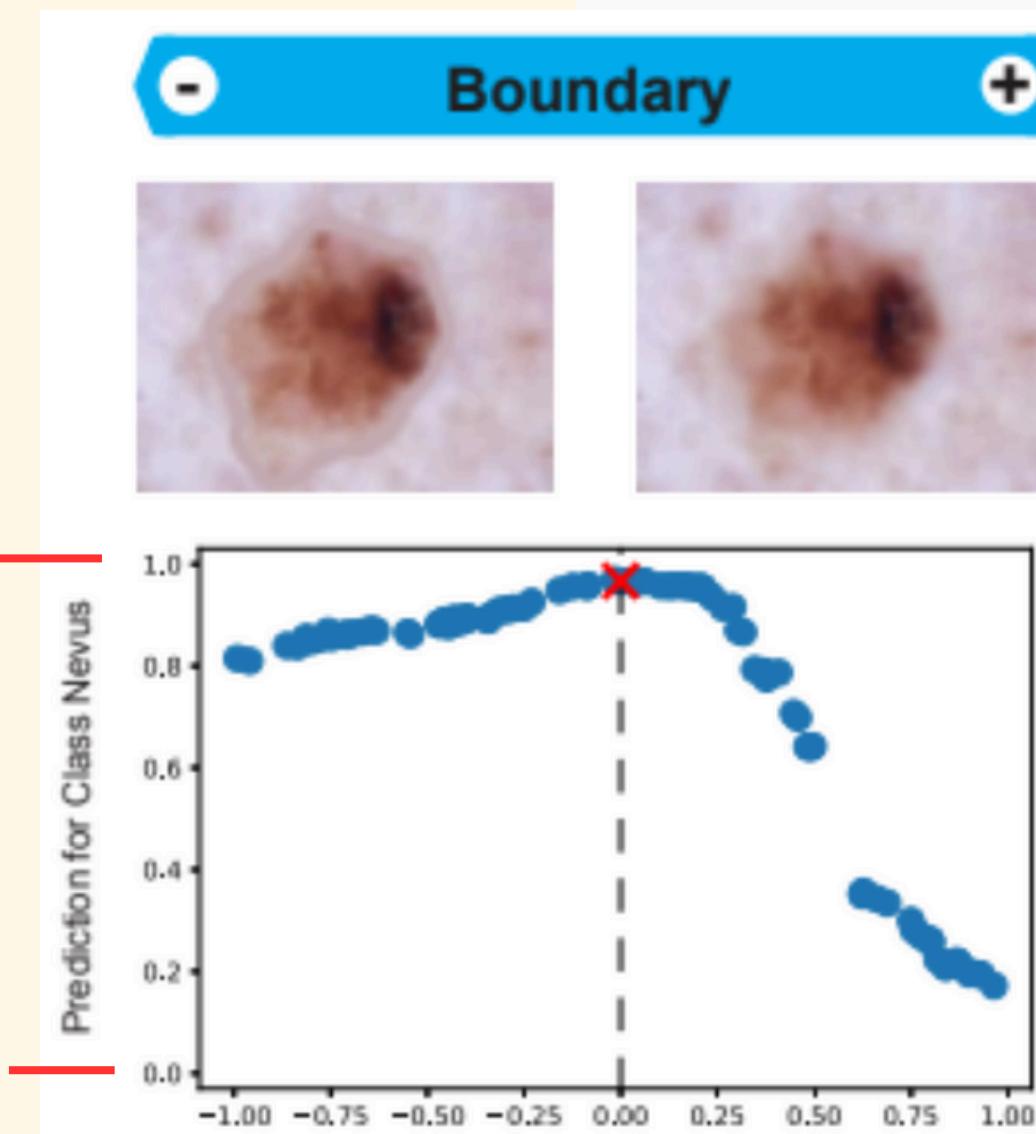
- True positive case เลือก ตัวอย่างที่มีความเชื่อมั่นสูง
- False positive case เลือก ตัวอย่างที่มีความเชื่อมั่นต่ำ



Paragraph 2 : (การอธิบายภาพ)

- ส่วนบน: ชื่อมิติ
- ส่วนล่าง: Scatter plot ระหว่าง ผลการทำนายกับ scale ในการปรับมิติ

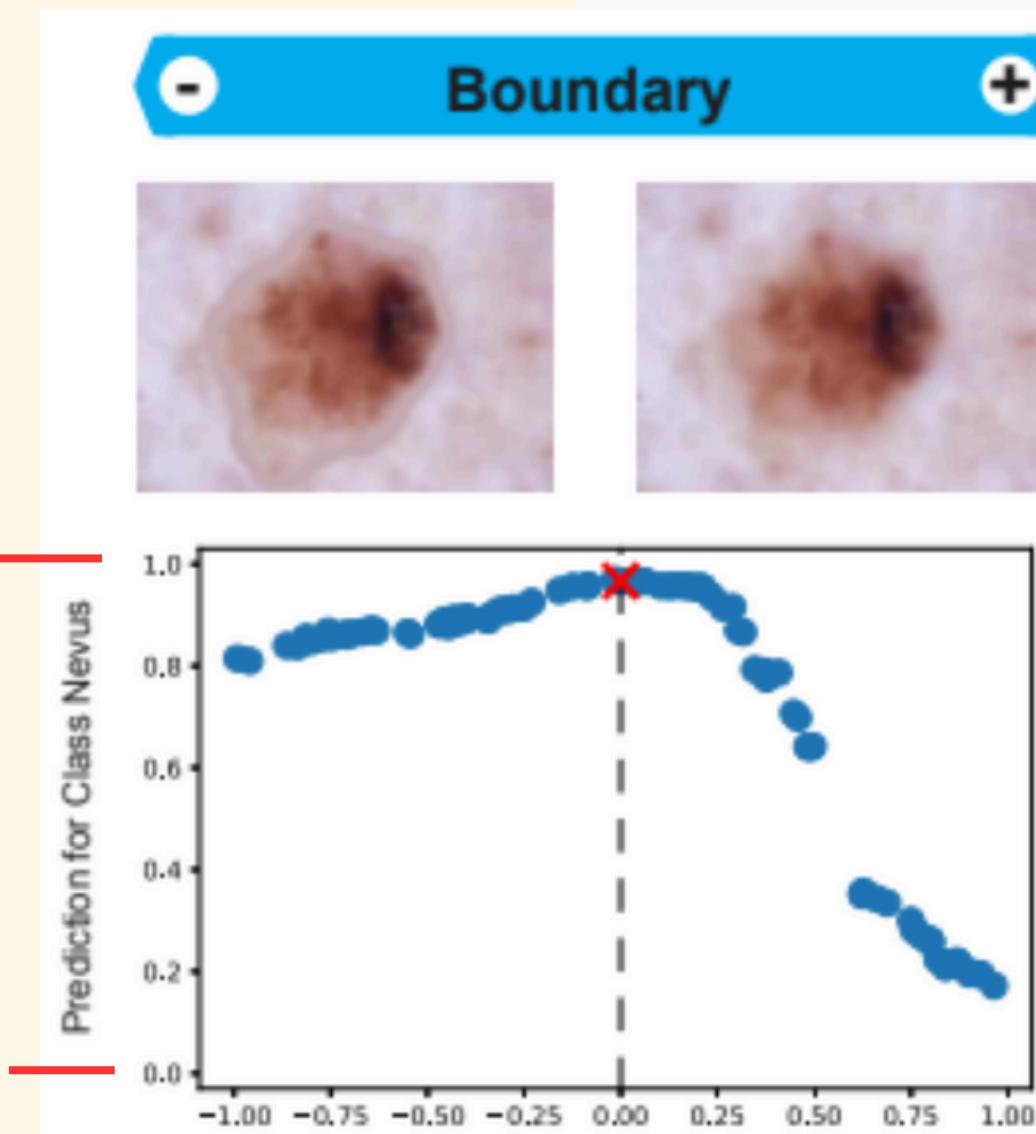
04 EMPIRICAL RESULTS



Paragraph 3 : (การอธิบายพิกัดจุดบนกราฟ scatter plot)

- แกน Y พิกัดจุดอยู่ระหว่าง $[0, 1]$ อ้างอิงถึง class ของตัวอย่าง
- แกน X พิกัดจุดอยู่ระหว่าง $[-1, 1]$ ค่าทางลบอ้างอิงถึงมิติการรบกวนทางลบ
ค่าทางบวกอ้างอิงถึงมิติการรบกวนทางบวก

04 EMPIRICAL RESULTS



Paragraph 4 : (การอธิบายกราฟ scatter plot)

- เลี้นประ เป็นตัวแบ่งกราฟ Scatter plots ที่ 0
- kakabathie แสดงผลการคำนวณของภาพที่ไม่ผ่านการถูกรบกวน
- $n = 50$

04 EMPIRICAL RESULTS

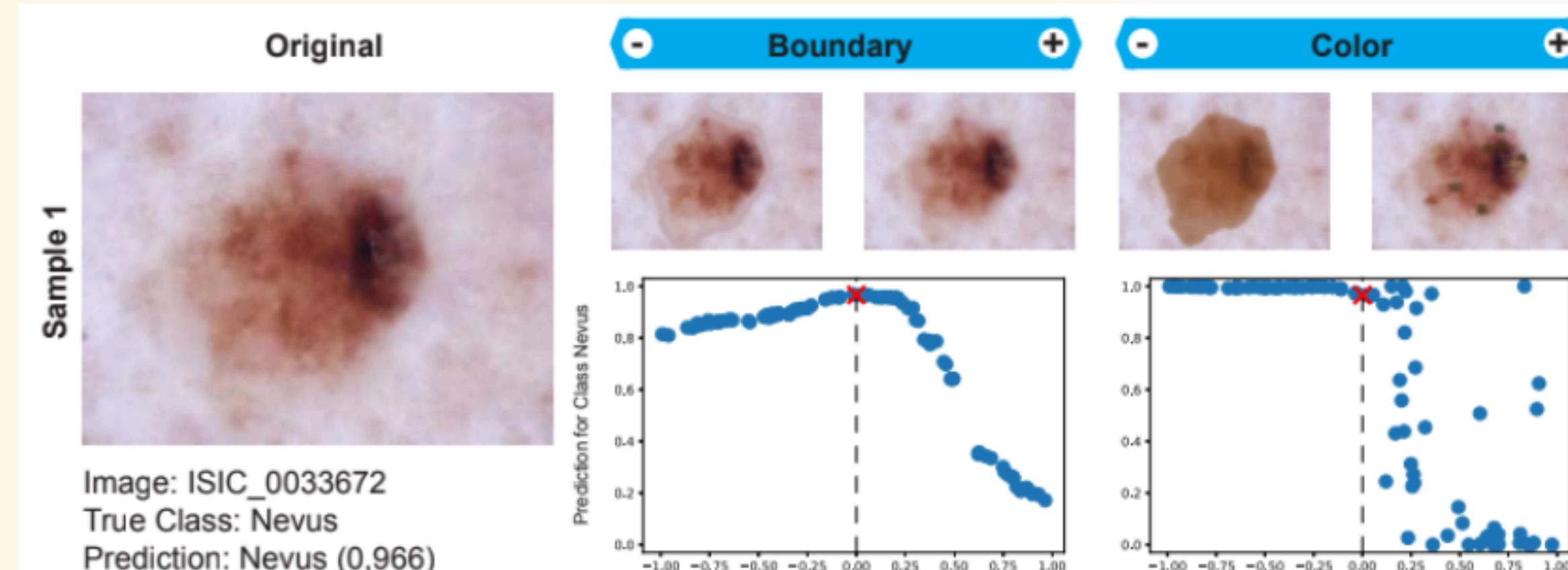
4.1 TRUE POSITIVES

Paragraph 1 :

- เริ่มตรวจสอบคำอธิบายแบบจำลองจากคำตอบที่ตัวแบบทายถูกต้อง
- เพื่อตอบคำถามว่า “แบบจำลองยังคงถูกต้องในมิติใด”

04 EMPIRICAL RESULTS

4.1 TRUE POSITIVES



สมมติฐาน:

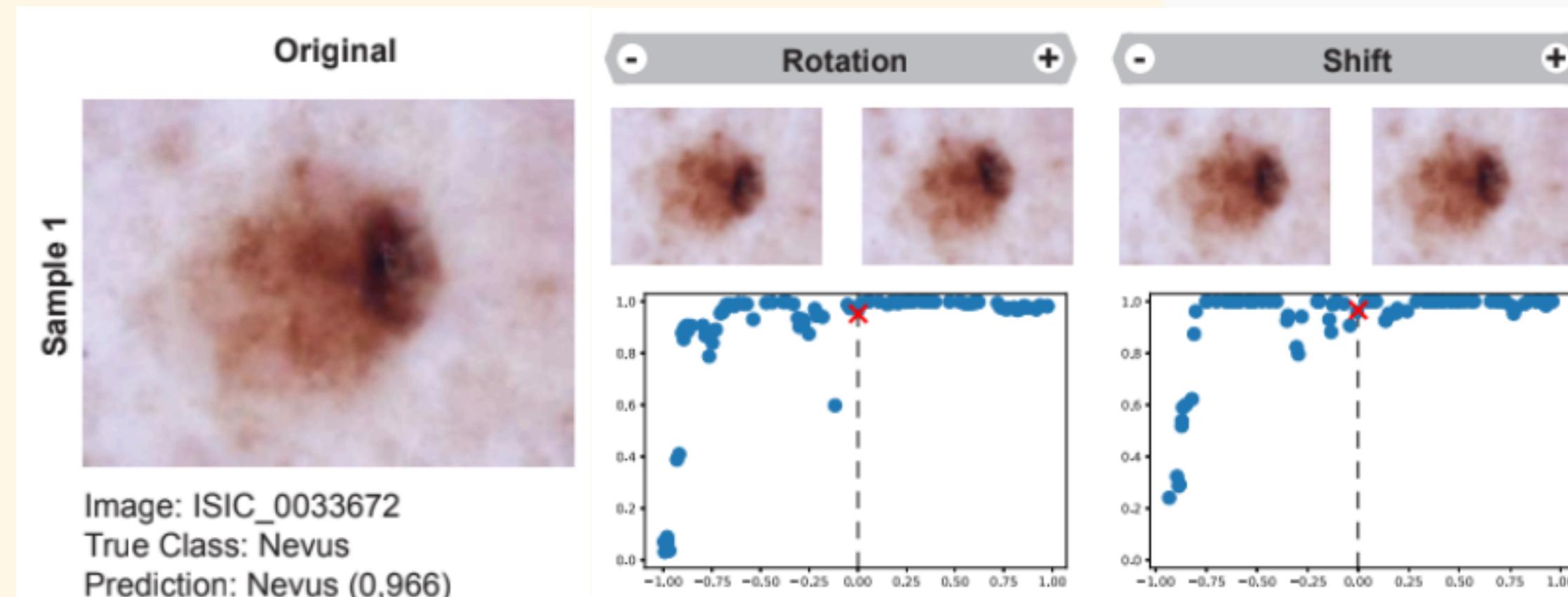
- A₁ ค่าการทำนายในปานจะลดลงใน **Positive perturbation**
- A₀ ค่าการทำนายในปานจะเพิ่มขึ้นหรือไม่เปลี่ยนแปลง ใน **Positive perturbation**
- B₁ ค่าการทำนายในปานจะเพิ่มขึ้นใน **Negative perturbation**
- B₀ ค่าการทำนายในปานจะลดลงหรือไม่เปลี่ยนแปลง ใน **Negative perturbation**

Paragraph 2: (ตรวจสอบสมมติฐานในมิติทางการแพทย์)

- ใน Boundary และ Color dimension ค่าที่ได้จากการทำนายใน Positive perturbation ลดลง --> ยอมรับ A₁
- Color ค่าที่ได้จากการทำนายใน Negative perturbation เพิ่มขึ้น --> ยอมรับ B₁
- Boundary ค่าที่ได้จากการทำนายใน Negative perturbation ลงลง --> ปฏิเสธ B₁

04 EMPIRICAL RESULTS

4.1 TRUE POSITIVES



สมมติฐาน:

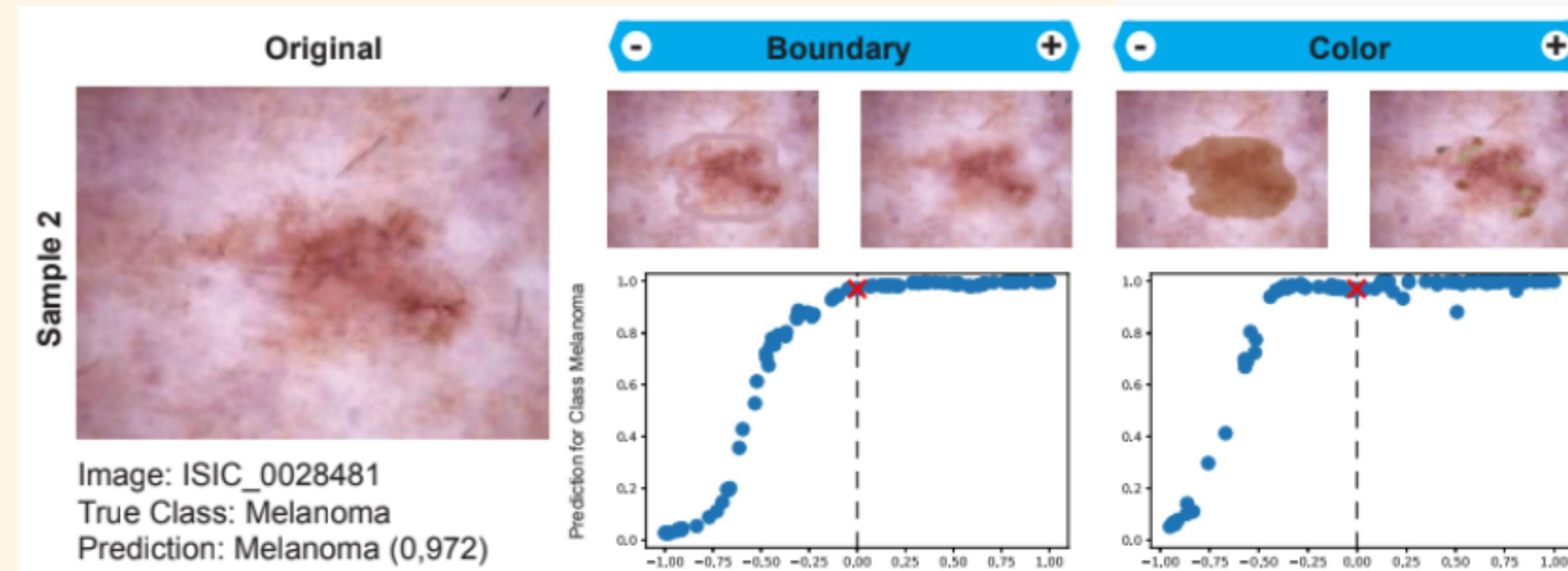
- C₁** การรับกวนมิติที่ไม่ใช่ทางการแพทย์ไม่ส่งผลกระทบต่อการทำนาย
- C₀** การรับกวนมิติที่ไม่ใช่ทางการแพทย์ส่งผลกระทบต่อการทำนายอย่างมีนัยสำคัญ

Paragraph 3: (ตรวจสอบสมมติฐานในมิติที่ไม่เกี่ยวกับทางการแพทย์)

- เมื่อค่าการทำนายจะเปลี่ยนแปลงไปในแต่ละจุด แต่ค่าที่ได้ก็ยังสูง ทั้งใน Positive และ Negative
- Rotation มีค่าการทำนายเฉลี่ยอยู่ที่ $y = 0.931$ (-0.035) --> ยอมรับ C₁
- Shift มีค่าการทำนายเฉลี่ยอยู่ที่ $y = 0.959$ (-0.007) --> ยอมรับ C₁

04 EMPIRICAL RESULTS

4.1 TRUE POSITIVES



สมมติฐาน:

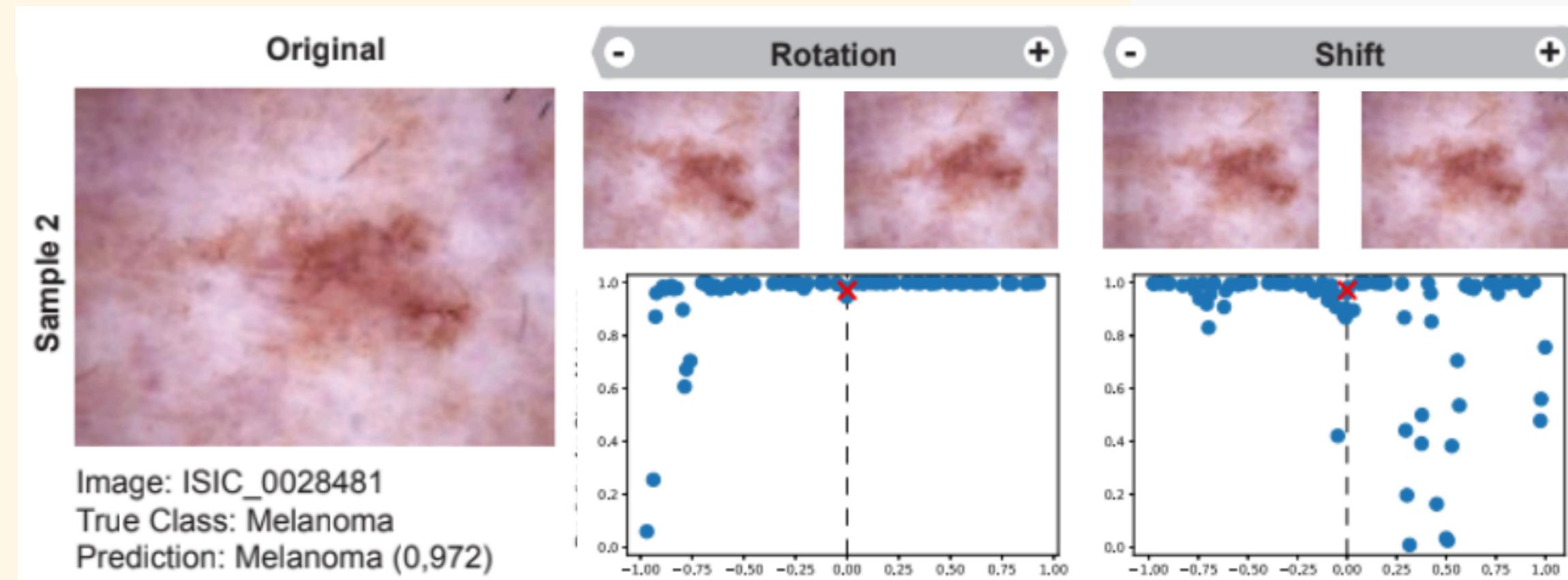
- A₁ ค่าการทำนายในมะเร็งผิวหนังจะเพิ่มขึ้นหรือไม่เปลี่ยนแปลง ใน **Positive perturbation**
- A₀ ค่าการทำนายในมะเร็งผิวหนังจะลดลง ใน **Positive perturbation**
- B₁ ค่าการทำนายในมะเร็งผิวหนังจะลดลงหรือไม่เปลี่ยนแปลงใน **Negative perturbation**
- B₀ ค่าการทำนายในมะเร็งผิวหนังจะเพิ่มขึ้นใน **Negative perturbation**

Paragraph 4: (ตรวจสอบสมมติฐานในมิติทางการแพทย์)

- ใน Boundary และ Color dimension ค่าที่ได้จากการทำนายใน Positive perturbation เพิ่มขึ้น -- > ยอมรับ A₁
- ใน Boundary และ Color dimension ค่าที่ได้จากการทำนายใน Negative perturbation ลดลง -- > ยอมรับ B₁

04 EMPIRICAL RESULTS

4.1 TRUE POSITIVES



สมมติฐาน:

- C₁ การรับกวนมิติที่ไม่ใช่ทางการแพทย์ไม่ส่งผลกระทบต่อการทำนาย
- C₀ การรับกวนมิติที่ไม่ใช่ทางการแพทย์ส่งผลกระทบต่อการทำนายอย่างมีนัยสำคัญ

Paragraph 5: (ตรวจสอบสมมติฐานในมิติที่ไม่เกี่ยวข้องทางการแพทย์)

- Rotation มีค่าการทำนายเฉลี่ยอยู่ที่ $\bar{y} = 0.971$ (-0.001) --> ยอมรับ C₁
- Shift มีค่าการทำนายเฉลี่ยอยู่ที่ $\bar{y} = 0.907$ (0.065) --> ปฏิเสธ C₁

04 EMPIRICAL RESULTS

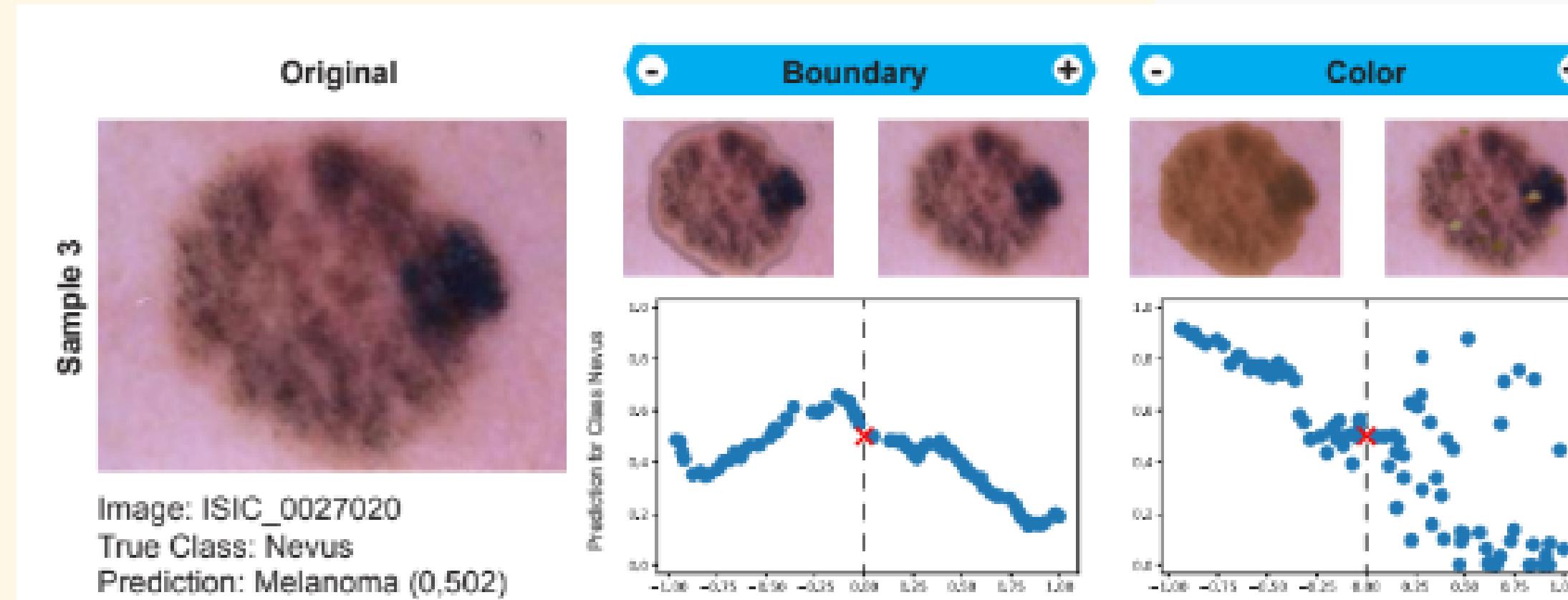
4.2 FALSE POSITIVES

Paragraph 1 :

- ตรวจสอบคำอธิบายแบบจำลองจากคำตอบที่ตัวแบบทายไม่ถูกต้อง
- เพื่อตอบคำถามว่า “ทำไมแบบจำลองถึงทายผิด”

04 EMPIRICAL RESULTS

4.2 FALSE POSITIVES



ສມມຕິຮານ:

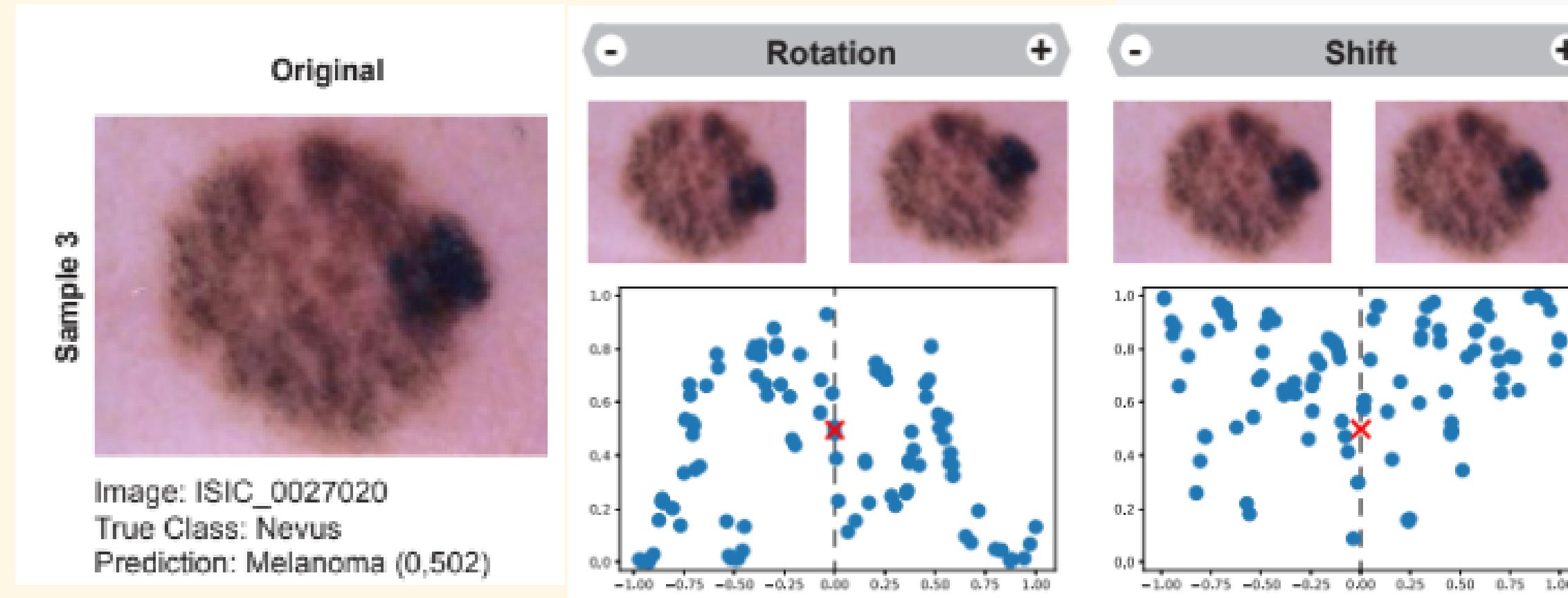
- A₁** ค่าการคำนวณในปานจะลดลงใน **Positive perturbation**
 - A₀** ค่าการคำนวณจะเพิ่มขึ้นหรือไม่เปลี่ยนแปลง ใน **Positive perturbation**
 - B₁** ค่าการคำนวณในปานจะเพิ่มขึ้นใน **Negative perturbation**
 - B₀** ค่าการคำนวณในปานจะลดลงหรือไม่เปลี่ยนแปลง ใน **Negative perturbation**

Paragraph 2: (ตรวจสอบสมมติฐานในมิติทางการแพทย์)

- Boundary Positive perturbationลดลง -->ยอมรับ **A₁**
 - Boundary ค่าที่ได้จากการคำนวณใน Negative perturbation ลงลง -->ปฏิเสธ **B₁**
 - Color ค่าที่ได้จากการคำนวณใน Positive perturbation ขึ้นๆ ลงๆ -->ปฏิเสธ **A₁**
 - Color ค่าที่ได้จากการคำนวณใน Negative perturbation เพิ่มขึ้น -->ยอมรับ **B₁**

04 EMPIRICAL RESULTS

4.2 FALSE POSITIVES



สมมติฐาน:

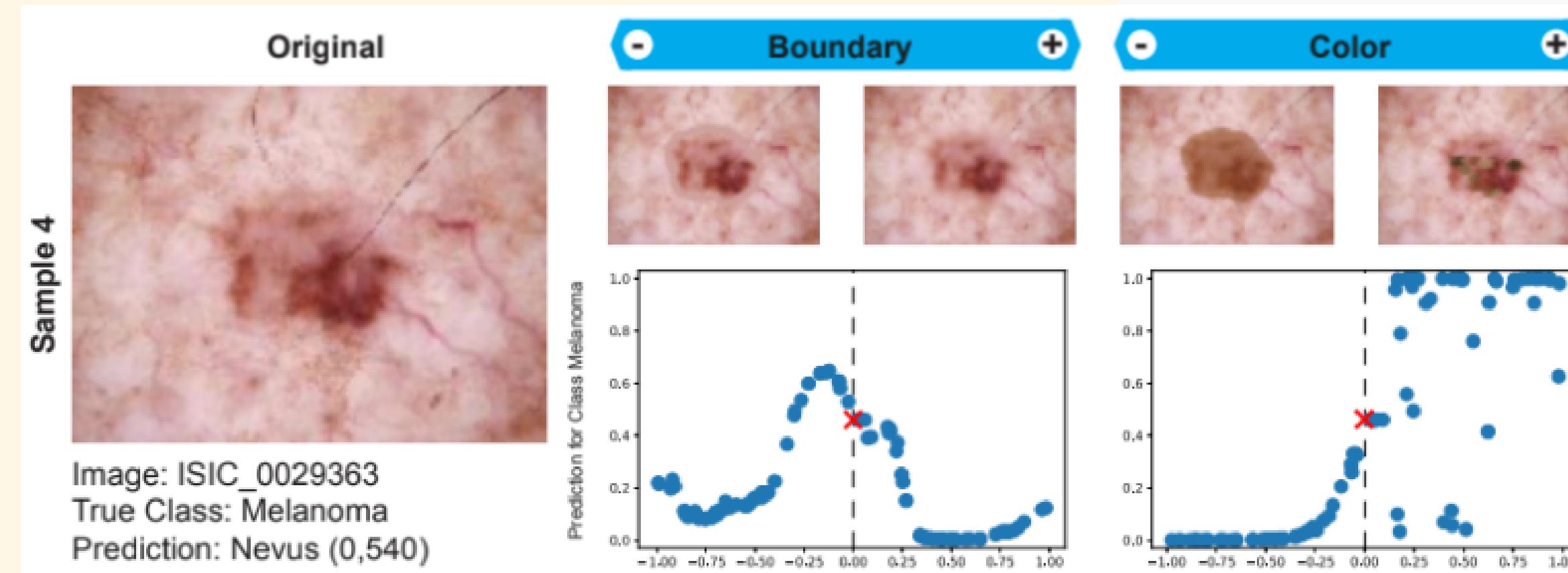
- C₁ การรับกวนมิติที่ไม่ใช่ทางการแพทย์ไม่ส่งผลกระทบต่อการทำนาย
- C₀ การรับกวนมิติที่ไม่ใช่ทางการแพทย์ส่งผลกระทบต่อการทำนายอย่างมีนัยสำคัญ

Paragraph 2: (ตรวจสอบสมมติฐานในมิติที่ไม่เกี่ยวกับทางการแพทย์)

- ค่าการทำนายเฉลี่ย ในทิศทางบวก $y = 0.318$ (-0.184)
- Rotation และ Shift --> ปฏิเสธ C₁

04 EMPIRICAL RESULTS

4.2 FALSE POSITIVES



สมมติฐาน:

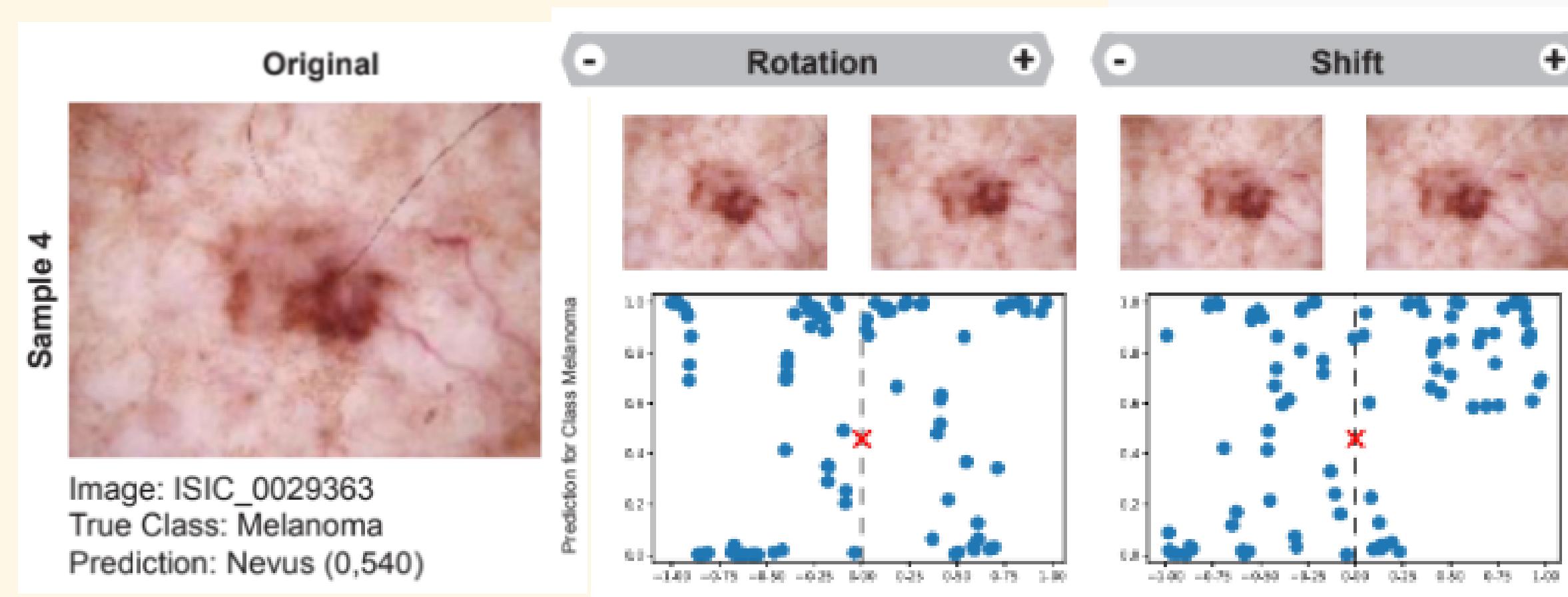
- A₁ ค่าการทำนายในมะเร็งผิวหนังจะเพิ่มขึ้น หรือไม่เปลี่ยนแปลง ใน Positive perturbation
- A₀ ค่าการทำนายในมะเร็งผิวหนังจะลดลง ใน Positive perturbation
- B₁ ค่าการทำนายในมะเร็งผิวหนังจะลดลงหรือ ไม่เปลี่ยนแปลงใน Negative perturbation
- B₀ ค่าการทำนายในมะเร็งผิวหนังจะเพิ่มขึ้นใน Negative perturbation

Paragraph 3: (ตรวจสอบสมมติฐานในมิติทางการแพทย์)

- ใน Boundary ค่าที่ได้จากการทำนายใน Positive และ Negative perturbation ลดลง --> ปฏิเสธ A₁ และ B₁
- ใน Color ค่าที่ได้จากการทำนายใน Positive เพิ่มขึ้น --> ยอมรับ A₁
 $y = 0.688(+0.148)$ สูงกว่าการทำนายของกลุ่มตัวอย่างที่ไม่ถูกกระบวนการ --> ยอมรับ B₁

04 EMPIRICAL RESULTS

4.2 FALSE POSITIVES



สมมติฐาน:

- C₁ การรับกวนมิติที่ไม่ใช่ทางการแพทย์ไม่ส่งผลกระทบต่อการทำนาย
- C₀ การรับกวนมิติที่ไม่ใช่ทางการแพทย์ส่งผลกระทบต่อการทำนายอย่างมีนัยสำคัญ

Paragraph 4:

- ค่าการทำนายเฉลี่ย $\bar{y} = 0.688(+0.148)$
- Rotation และ Shift --> ปฏิเสธ C₁

05 DISCUSSION

Paragraph 1:

- การทดลองทำให้สามารถสรุปได้ว่าปัจจัยใดมีความสำคัญต่อการตัดสินใจของตัวแบบโดยอุดจากพฤติกรรมของตัวแบบ ต่อการปรับเปลี่ยนภาพในแต่ละมิติ

Paragraph 2: (limitation)

- ศึกษาเพียง 1 ตัวแบบ
- ศึกษาแค่การปรับเปลี่ยนภาพที่ลงทะเบียน

Paragraph 3: (limitation)

- พัฒนาวิธีอธิบายตัวแบบสำหรับตัวแบบทางด้านจำแนกประเภทพิวนัง (local) (goal)
- global model explanations (limitation)

Paragraph 4: (limitation)

- ไม่มีหลักฐานที่ชัดเจนเพียงพอระหว่างความสำคัญของคุณลักษณะที่สังเกตได้จากการทดลองและคะแนนจริงตามกฎ ABCD rule

06 CONCLUSION AND FUTURE WORK

- งานวิจัยนี้นำเสนอ XAI ที่สามารถช่วยให้แพทย์เข้าใจการตัดสินใจของโมเดล AI ในการวินิจฉัยโรคจากภาพผิวนั้งได้ดีขึ้น และช่วยสร้างความเชื่อมั่นในการใช้ AI ในการตัดสินใจทางการแพทย์
- การวิจัยในอนาคตสามารถขยายไปยังมิติต่าง ๆ ของภาพที่ยังไม่ได้ศึกษา และนำผลการวิเคราะห์เหล่านี้ไปประยุกต์ใช้กับสาขาแพทย์อื่น ๆ

END