

חיזוי סרטן ריאה באמצעות מודל למידת מכונה ובדיקות מעבדה

ליהי שוורץ¹, נאור מתניה², מתנאל לוי³, טדי לזבניק⁴, שירי קושניר⁵, נוגה יוסף⁶, אסף חוגי⁷, דקל שלומי⁸

¹מרפאת פלינר, המחלקת להתמחות ברפואת המשפחה, מחוז דן-פתח תקווה, חטיבת הקהילה, שירותי בריאות כללית, פתח תקווה, ישראל;
²המחלקה למדעי המחשב, אוניברסיטת בר-אילן, רמת גן, ישראל; ³בית הספר לרפואה ע"ש אדלסון, אוניברסיטת אריאל, אריאל, ישראל;
⁴המחלקה למתמטיקה, אוניברסיטת אריאל, אריאל, ישראל והמחלקה לביוכימיה של הסרטן, מכון הסרטן, קולג' לונדון, לונדון, בריטניה; ⁵רשות המחקר, מרכז רפואי רבין, קמפוס בילינסון, פתח תקווה, ישראל; ⁶יחידת המחקר, מחוז דן-פתח תקווה, חטיבת קהילה, שירותי בריאות כללית, רמת גן, ישראל; ⁷בית הספר למדעי המחשב ומרכז המחקר לבינה מלאכותית ומדעי הנתונים, אוניברסיטת אריאל, אריאל, ישראל; ⁸בית הספר לרפואה ע"ש אדלסון, אוניברסיטת אריאל, אריאל ומרפאת ריאות, מחוז דן-פתח תקווה, שירותי בריאות כללית.

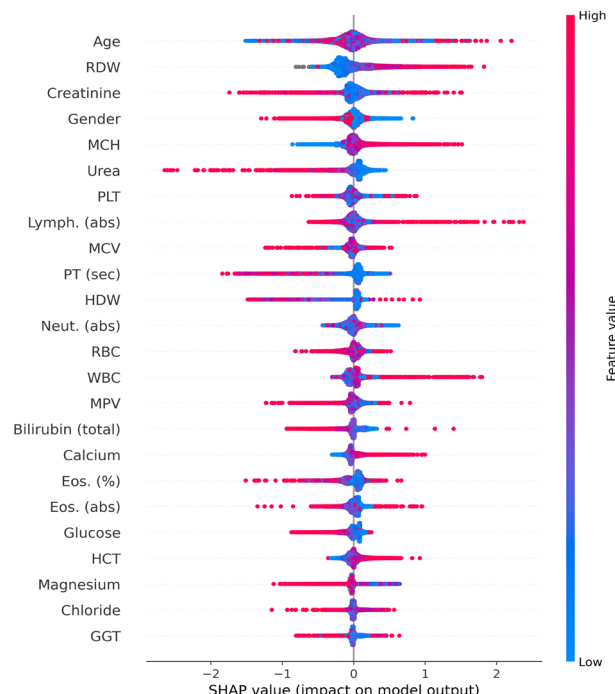
חיזוי ממאירות מותאם אישית הינה מטרה עליונה בגילוי מוקדם. הקריטריונים המרכזיים לפיהם מבוצעת סקירה שנתית לאיתור סרטן הריאה באמצעות CT בקרינה נמוכה הם גיל והיסטוריית עישון. כיום אין המלצות מבוססות לבדיקות סקר באוכלוסיות אחרות כגון בלא מעשנים אך קיימים מחשבוני שונים עם רמת דיוק בינונית. למידת מכונה (Machine Learning) עשויה להיות אופציה מבטיחה לחיזוי דפוסי מורכבים לצורך גילוי סרטן ריאה בקרב אוכלוסיות שונות תוך שימוש במדדים שונים מתוך התיק הרפואי והתאמת הסיכון האישי.

מודל למידת מכונה לחיזוי עתידי של סרטן ריאה המבוסס על בדיקות מעבדה שגרתיות ומאפיינים סוציודמוגרפיים (גיל ומגדר) אשר משווה בין קבוצת חולים בסרטן ריאה לקבוצת ביקורת.

מלבד גיל ומגדר, המודל התבסס על 22 בדיקות מעבדה אשר נמצאו בבדיקת מקדימות על ידי המודל התורמים העיקריים לניבוי. עבור כלל אוכלוסיית המחקר, רמת הדיוק (Accuracy) של המודל לניבוי תחלואה עתידית בסרטן ריאה הייתה 71.2%, שטח מתחת לעקומה (AUC) - 78.7%, רגישות (Sensitivity) של 63% וערך מנבא חיובי (PPV) של 67.2%. רמת דיוק גבוהה יותר נצפתה דווקא בקרב נשים לעומת גברים (71.8% לעומת 70.8% בהתאמה) ובלא מעשנים לעומת חשיפה לעישון (73.6% לעומת 70.1% בהתאמה). הגורם המנבא המוביל היה גיל (13.6%) ולאחריו התפלגות גודל תאי הדם האדומים (RDW) (5.1%), קראטינין (5%), מגדר (3.6%), כמות ההמוגלובין בכדוריות הדם האדומות (MCH) (3.3%), ורמת אוראה בדם (מעט מתחת ל-3%). מרבית הבדיקות המעבדה תרמו במודל מורכב של חיזוי אישי באופן שונה בין אדם לאדם (מעלות או מורידות את הסיכון) בהתאם לקשרים המיוחדים בין כלל הבדיקות. לעומת זאת, מספר בדיקות (RDW, MCH PT, המטוקריט, אוריה וסידן) תרמו למודל הניבוי באופן מעט יותר דיכוטומי.

Shapley Additive Explanations (SHAP)

גורמי הסיכון ל-LC מוצגים על גבי הגרף לפי תרומתם למודל, כשהחשוב ביותר מופיע בראש הרשימה. עבור כל גורם סיכון, ציר ה-x מציין את תרומת המשתנה לסיכון לסרטן ריאה, כאשר ערכים מעל 0 מעלים את הסיכון וערכים מתחת ל-0 מפחיתים את הסיכון. ציר ה-y מציין את תרומתו הממוצעת של כל משתנה למודל הניבוי, כאשר אות משמעותי יותר מייצג תרומה גדולה יותר. בנוסף, תרשים ה-"beeswarm" מוסיף קידוד צבעוני לכל נקודה בהתאם להשוואת ערך המשתנה של אותם נבדקים לממוצע באוכלוסייה כולה. כך, ככל שהצבע אדום יותר, כך ערך התוצאה הספציפית גבוה יותר מהממוצע של ערכי המשתנה לאוכלוסייה הנבדקת, וככל שהצבע כחול יותר, כך ערך התוצאה נמוך יותר מהממוצע. עבור מין המטופל (gender), אדם מייצג נשים וכחול מייצג גברים.



ניתן להשתמש במודל למידת המכונה לעיל, המבוסס על בדיקות מעבדה לניבוי סרטן ריאה, במידת הצלחה סבירה. יש צורך בביצוע מחקרים נוספים באוכלוסיות שונות וכן ב מחקרי מדע בסיסי על מנת לבנות מודל המבוסס על מדדים שונים הכוללים תוצאות בדיקות מעבדה לניבוי ממאירות ריאה.