**Lung Cancer Risk Prediction with Machine Learning Models**

**Abstract:**

ปอดเป็นศูนย์กลางของการควบคุมการหายใจและทำหน้าที่ให้แน่ใจว่าเซลล์ทุกเซลล์ในร่างกายได้รับออกซิเจน ในขณะเดียวกัน ปอดยังกรองอากาศเพื่อป้องกันไม่ให้สารที่ไม่มีประโยชน์และเชื้อโรคเข้าสู่ร่างกาย ร่างกายมนุษย์มีระบบป้องกันที่ออกแบบมาอย่างพิเศษเพื่อปกป้องปอด อย่างไรก็ตาม ระบบเหล่านี้ไม่เพียงพอที่จะขจัดความเสี่ยงของโรคต่าง ๆ ที่ส่งผลต่อปอดได้อย่างสมบูรณ์ การติดเชื้อ การอักเสบ หรือภาวะแทรกซ้อนที่รุนแรงกว่า เช่น การเจริญเติบโตของเนื้องอกมะเร็ง สามารถส่งผลกระทบต่อปอดได้ ในงานวิจัยนี้ เราใช้วิธีการเรียนรู้ของเครื่อง (Machine Learning) ในการสร้างแบบจำลองที่มีประสิทธิภาพเพื่อระบุบุคคลที่มีความเสี่ยงสูงที่จะเป็นมะเร็งปอด และทำให้สามารถทำการแทรกแซงล่วงหน้าเพื่อหลีกเลี่ยงภาวะแทรกซ้อนในระยะยาว ข้อเสนอของบทความนี้คือการใช้ Rotation Forest ซึ่งแสดงให้เห็นถึงประสิทธิภาพสูงและประเมินโดยตัวชี้วัดที่เป็นที่รู้จักดี เช่น ความแม่นยำ (precision) การเรียกคืน (recall) F-Measure ความถูกต้อง (accuracy) และพื้นที่ใต้เส้นโค้ง (Area Under the Curve หรือ AUC) โดยเฉพาะอย่างยิ่ง การประเมินการทดลองพบว่าโมเดลที่นำเสนอมีค่า AUC อยู่ที่ 99.3% และ F-Measure, precision, recall และ accuracy อยู่ที่ 97.1%

Keywords: การดูแลสุขภาพ; มะเร็งปอด; การทำนาย; การเรียนรู้ของเครื่อง; การวิเคราะห์ข้อมูล

**1. Introduction**

ปอดเป็นอวัยวะหลักของการหายใจ ร่างกายมนุษย์มีปอดสองข้าง โดยอยู่ด้านละข้างของทรวงอก ปอดด้านซ้ายมีขนาดเล็กกว่าปอดด้านขวาเนื่องจากต้องเว้นที่ให้กับหัวใจ ระหว่างการหายใจ ทรวงอกจะยกขึ้นและยุบลง นั่นเป็นเพราะเมื่อหายใจเข้า ปอดจะขยายตัว และเมื่อหายใจออก ปอดจะหดตัว ปอดมีหน้าที่เพิ่มปริมาณออกซิเจนในเลือด หัวใจจะส่งเลือดที่มีออกซิเจนต่ำและมีคาร์บอนไดออกไซด์สูงไปยังปอด เลือดในปอดจะถูก "ชำระล้าง" โดยดูดซับออกซิเจนและปล่อยคาร์บอนไดออกไซด์ คาร์บอนไดออกไซด์จะถูกกำจัดออกในขณะที่หายใจออก ในขณะที่ออกซิเจนจะเข้าสู่ปอดเมื่อหายใจเข้า [1,2].

นอกจากนี้ อากาศจะต้องผ่านทางโพรงจมูก (หรือโพรงปากในกรณีที่เราหายใจทางปาก) คอหอย กล่องเสียง หลอดลมใหญ่ และหลอดลมฝอย ก่อนจะเข้าสู่ปอด ภายในปอด หลอดลมฝอยจะแตกแขนงออกเป็นหลอดลมฝอยที่เล็กลงเรื่อยๆ และไปสิ้นสุดที่ถุงลม ภายในถุงลมมีหลอดเลือดฝอยจำนวนมากซึ่งจะปล่อยคาร์บอนไดออกไซด์ออกสู่ถุงลมและรับออกซิเจน มนุษย์จะไม่หยุดหายใจจนกว่าจะเสียชีวิต เพราะปอดทำหน้าที่จัดหาออกซิเจนให้กับเลือด ซึ่งเป็นสิ่งสำคัญต่อชีวิตมนุษย์ [2].

ในประเทศที่พัฒนาแล้ว โรคปอดเป็นหนึ่งในสาเหตุหลักของการเสียชีวิต ปัจจัยต่าง ๆ เช่น การสูบบุหรี่ สารพิษจากสิ่งแวดล้อม และการอักเสบเรื้อรัง ส่งผลกระทบที่เป็นอันตรายซึ่งมักนำไปสู่ความเสียหายถาวร ปอดมีความสามารถในการทำความสะอาดตัวเองผ่านกระบวนการและกลไกต่าง ๆ เช่น เสมหะ อย่างไรก็ตาม สำหรับผู้ที่สูบบุหรี่ ระบบนี้ไม่เพียงพอ ปัจจัยด้านสิ่งแวดล้อม พันธุกรรม หรือการสืบทอด หรือการผสมผสานของสิ่งเหล่านี้ สามารถส่งผลกระทบต่อปอดและส่งเสริมการเกิดโรคต่าง ๆ ได้ โรคที่เกิดขึ้นในระบบทางเดินหายใจแบ่งออกได้เป็นหลายประเภท [1,3].

โดยเฉพาะอย่างยิ่ง โรคปอดอุดกั้นเรื้อรัง (COPD) ประกอบด้วยโรคหลอดลมอักเสบเรื้อรังและถุงลมโป่งพอง บ่อยครั้งที่โรคทั้งสองนี้เกิดขึ้นพร้อมกัน ส่งผลให้เกิดสภาวะที่ซับซ้อนเรียกว่าโรคปอดอุดกั้นเรื้อรัง การสูบบุหรี่เป็นสาเหตุหลักของโรคปอดอุดกั้นเรื้อรัง [4]. หลอดลมอักเสบเรื้อรังมีลักษณะเฉพาะคือการอักเสบและความเสียหายต่อเยื่อบุหลอดลม ซึ่งหลอดลมทำหน้าที่เชื่อมต่อหลอดลมใหญ่กับปอด อาการหลักคือไอเรื้อรัง การผลิตเสมหะมากขึ้น และหายใจลำบาก อาการหลักของถุงลมโป่งพองคือไอ หายใจลำบาก ความทนทานต่อการออกกำลังกายลดลง และต้องใช้ความพยายามในการทำกิจกรรมต่างๆ [5].

นอกจากนี้ โรคหืดเป็นภาวะเรื้อรังที่ส่งผลกระทบต่อหลอดลมและหลอดลมฝอย อาการที่พบมากที่สุดของโรคหืดคือหายใจลำบากและเสียงหวีดเนื่องจากการตีบของทางเดินหายใจ [6]. โรคซิสติกไฟโบรซิสเป็นภาวะที่ถ่ายทอดทางพันธุกรรมที่ส่งผลกระทบต่อเสมหะและเหงื่อของผู้ป่วย เนื่องจากปัญหาที่เกิดขึ้น เสมหะสะสมอยู่ในปอดและเป็นสาเหตุของการติดเชื้อในปอดบ่อยครั้ง ทีละน้อยเกิดความเสียหายถาวรต่อปอด และเกิดภาวะล้มเหลวของระบบหายใจที่รุนแรง วัณโรคเป็นการติดเชื้อที่เกิดจากเชื้อแบคทีเรียชนิดหนึ่งที่ส่งผลกระทบต่อปอดเป็นหลัก เชื้อแบคทีเรียนี้ทำให้เกิดการอักเสบในเนื้อเยื่อปอดและทำลายเนื้อเยื่อนั้น [7]. สุดท้าย โรคปอดบวมครอบคลุมกลุ่มของโรคติดเชื้อที่เกิดจากการติดเชื้อในปอดจากเชื้อโรคต่าง ๆ เช่น แบคทีเรีย ไวรัส ปรสิต และเชื้อรา [8].

มะเร็งปอดเป็นสาเหตุหลักของการเสียชีวิตจากโรคร้ายในทั้งสองเพศ ควรสังเกตว่าการเสียชีวิตจากมะเร็งปอดมีมากกว่าการเสียชีวิตจากมะเร็งลำไส้ใหญ่ มะเร็งปากมดลูก และมะเร็งเต้านมรวมกัน อาการที่พบบ่อยที่สุดของมะเร็งปอดคือการไอ ซึ่งต้องให้ความสำคัญเป็นพิเศษ เนื่องจากผู้ป่วยมะเร็งปอดส่วนใหญ่มักเป็นผู้สูบบุหรี่และมีโรคปอดอุดกั้นเรื้อรังซึ่งทำให้เกิดการไออยู่แล้ว สิ่งที่สำคัญยิ่งกว่าคือการเปลี่ยนแปลงลักษณะของการไอ (มีความต่อเนื่องมากขึ้น รุนแรงขึ้น และอาจมาพร้อมกับเสมหะหรือเสมหะปนเลือด) นอกจากนี้ อาการที่เกิดจากมะเร็งปอดรวมถึงเสมหะ เจ็บหน้าอก หายใจลำบาก เบื่ออาหาร น้ำหนักลด มีไข้ และไอปนเลือด [9–11].

การตรวจเอกซเรย์คอมพิวเตอร์ทรวงอก (CT) หรือการเอกซเรย์ทรวงอกเป็นวิธีทั่วไปในการวินิจฉัยมะเร็งปอด บางครั้งการใช้การถ่ายภาพด้วยคลื่นแม่เหล็กไฟฟ้า (MRI) หรือการถ่ายภาพแบบปล่อยโฟตอนโพสิตรอน (PET) จะถูกใช้ในขั้นตอนการประเมินการแพร่กระจายของมะเร็ง เนื่องจากวิธีนี้ช่วยกำหนดการจัดการรักษาที่ดีที่สุด การส่องกล้องหลอดลมและการตรวจชิ้นเนื้อ (การตรวจชิ้นเนื้อด้วยเข็ม การตรวจชิ้นเนื้อผ่าตัด) จำเป็นในการวินิจฉัยมะเร็งปอดอย่างแท้จริง รวมถึงให้ข้อมูลเกี่ยวกับชนิดทางพยาธิวิทยาด้วย [12,13].

ในหลายประเทศ จำนวนผู้ที่เคยสูบบุหรี่มีจำนวนมาก และมะเร็งปอดหลายประเภทส่งผลกระทบต่อผู้ที่เคยสูบบุหรี่เช่นกัน ในสหรัฐอเมริกาเพียงแห่งเดียว มีผู้ที่เคยสูบบุหรี่มากกว่า 50 ล้านคน (คือ ผู้ที่เลิกสูบบุหรี่แล้ว) [14] ดังนั้นวิธีการตรวจคัดกรองมะเร็งปอดจึงเป็นมาตรการที่มีหลักฐานสนับสนุนเพื่อการตรวจพบและรักษามะเร็งปอดก่อนที่มะเร็งจะลุกลามถึงระยะที่แพร่กระจายถึงชีวิตในผู้สูบบุหรี่ทั้งปัจจุบันและอดีต การสนับสนุนให้เลิกสูบบุหรี่เป็นสิ่งสำคัญสำหรับผู้ที่ยังสูบบุหรี่อยู่ แต่ความเสี่ยงของมะเร็งปอดเป็นภัยตลอดชีวิตสำหรับทุกคนที่เคยสูบบุหรี่ ความเสี่ยงที่ผู้ป่วยจะเสียชีวิตจากมะเร็งปอดขึ้นอยู่กับระยะที่มะเร็งอยู่ หากมีการตรวจพบในระยะเริ่มต้น โรคนี้สามารถรักษาให้หายได้ แต่หากอยู่ในระยะลุกลามแล้ว อัตราการรอดชีวิตเฉลี่ยจะน้อยกว่าสองปี การตรวจพบมะเร็งปอดในระยะเริ่มต้นมีความสัมพันธ์กับอัตราการรักษาให้หายสูง ในขณะที่มะเร็งปอดที่ตรวจพบในระยะท้ายมักจะมีอัตราการรอดชีวิตเฉลี่ยต่ำกว่าสองปี [15–17].

ปัจจุบัน เทคโนโลยีปัญญาประดิษฐ์ (AI) และการเรียนรู้ของเครื่อง (ML) มีบทบาทสำคัญในด้านการดูแลสุขภาพ เนื่องจากความสามารถในการประยุกต์ใช้ AI/ML ที่หลากหลายในการทำนายความเสี่ยงของภาวะสุขภาพต่างๆ จึงจำเป็นต้องกำหนดข้อบังคับต่าง ๆ เพื่อประเมินและสนับสนุนการพัฒนาซอฟต์แวร์เครื่องมือที่ใช้ AI/ML เพื่อการทำนายและวินิจฉัยโรคในระยะแรก [18,19] โรคที่เครื่องมือเหล่านี้เกี่ยวข้องมากที่สุด ได้แก่ โรคเบาหวาน (ในรูปแบบการจัดประเภท [20] หรือการทำนายค่ากลูโคสต่อเนื่องในรูปแบบอนุกรมเวลา [21]) โรคความดันโลหิตสูง [22] โควิด-19 [23] ภาวะไขมันในเลือดสูง [24] โรคปอดอุดกั้นเรื้อรัง (COPD) [25] โรคหลอดเลือดสมอง [26] โรคหัวใจและหลอดเลือด (CVDs) [27] ภาวะตับวายเฉียบพลัน (ALF) [28] โรคนอนไม่หลับ [29] โรคไวรัสตับอักเสบซี [30] กลุ่มอาการเมตาบอลิก [31] โรคไตเรื้อรัง (CKD) [32] และอื่น ๆ

ในบริบทของการศึกษานี้ เราจะมุ่งเน้นที่มะเร็งปอด สำหรับโรคนี้ มีงานวิจัยทางวิทยาศาสตร์หลายชิ้นที่ดำเนินการจากมุมมองของการเรียนรู้ของเครื่อง (ML) ที่นี่ เรานำเสนอวิธีการออกแบบโมเดลการจำแนกประเภท ML ที่มีประสิทธิภาพเพื่อทำนายการเกิดมะเร็งปอดโดยใช้พฤติกรรมและอาการ/สัญญาณที่พบบ่อยที่สุดเป็นคุณลักษณะนำเข้าของโมเดล การมีส่วนร่วมของเราคือการประเมินเปรียบเทียบตัวจำแนกประเภทหลายตัวเพื่อพัฒนาโมเดลที่มีความไวและความสามารถในการแยกแยะสูงที่สุดในการระบุกลุ่มที่มีความเสี่ยงสูง สำหรับการประเมินโมเดล เราได้พิจารณาตัวชี้วัดประสิทธิภาพ ได้แก่ precision, recall, F-Measure, accuracy และ AUC นอกจากนี้ เรายังนำเสนอกราฟ AUC ROC ที่บันทึกไว้ด้วย สุดท้าย การวิเคราะห์ประสิทธิภาพในหลายด้านพบว่า Rotation Forest เป็นโมเดลที่มีประสิทธิภาพสูงสุด จึงเป็นข้อเสนอหลักของบทความวิจัยนี้

ส่วนถัดไปของบทความนี้ถูกจัดรูปแบบดังนี้ ในส่วนที่ 2 จะกล่าวถึงงานวิจัยที่เกี่ยวข้องกับหัวข้อที่กำลังศึกษา การนำเสนอชุดข้อมูลและการวิเคราะห์วิธีการที่ใช้จะอยู่ในส่วนที่ 3 นอกจากนี้ ในส่วนที่ 4 เราจะอภิปรายผลการทดลองที่ได้ สุดท้าย บทสรุปและแนวทางในอนาคตจะถูกกล่าวถึงในส่วนที่ 5

**2. Related Work**

ในที่นี้ เราจะนำเสนอภาพรวมโดยย่อของผลงานวิจัยที่เกี่ยวข้องล่าสุดเกี่ยวกับการทำนายการเกิดมะเร็งปอดโดยใช้เทคนิคและโมเดลการเรียนรู้ของเครื่อง (ML)

เริ่มต้นจาก [33] ผู้เขียนได้แสดงให้เห็นถึงวิธีการที่มีประสิทธิภาพในการตรวจจับและจำแนกมะเร็งปอดโดยใช้ภาพเอกซเรย์คอมพิวเตอร์ (CT scan) พวกเขาใช้โมเดลการจำแนกเจ็ดแบบ ได้แก่ ต้นไม้ตัดสินใจ (decision tree), ป่าแบบสุ่ม (random forest), เครื่องเวกเตอร์สนับสนุน (support vector machine), นาอีฟเบย์ (naive Bayes), k-nearest neighbors, การไล่ระดับความชันแบบสุ่ม (stochastic gradient descent) และ perceptron แบบหลายชั้น (multi-layer perceptron) สำหรับการฝึกอบรมและทดสอบตัวจำแนกเหล่านี้ มีการใช้ชุดข้อมูลทางคลินิกจำนวน 15,750 รายการ ซึ่งมีภาพที่เกี่ยวข้องกับมะเร็งปอดที่ไม่เป็นอันตราย 6,910 ภาพ และที่เป็นอันตราย 8,840 ภาพ ผลลัพธ์ที่ได้แสดงให้เห็นว่า ตัวจำแนกประเภท perceptron แบบหลายชั้นมีความแม่นยำสูงสุดที่ 88.55% เมื่อเทียบกับตัวจำแนกอื่น ๆ

ในทำนองเดียวกัน ใน [34] ผู้เขียนได้ใช้เครือข่ายประสาทเทียม, เครือข่ายฟังก์ชันฐานรัศมี (radial basis function network), เครื่องเวกเตอร์สนับสนุน, การถดถอยโลจิสติก, ป่าแบบสุ่ม, J48, นาอีฟเบย์ และ k-nearest neighbors ในการทำนายมะเร็งปอด พวกเขาแสดงให้เห็นว่า เครือข่ายฟังก์ชันฐานรัศมีให้ความแม่นยำสูงสุดที่ 81.25% ในชุดข้อมูลมะเร็งปอด นอกจากนี้ วัตถุประสงค์หลักของ [35] คือการวินิจฉัยมะเร็งปอดในระยะแรกโดยการประเมินประสิทธิภาพของอัลกอริธึมการจำแนกประเภท ผู้เขียนได้ใช้การจำแนกประเภทด้วยอัลกอริธึมต่าง ๆ เช่น นาอีฟเบย์, เครื่องเวกเตอร์สนับสนุน, ต้นไม้ตัดสินใจ และการถดถอยโลจิสติก ในชุดข้อมูลมะเร็งปอดจาก UCI การถดถอยโลจิสติกให้ความแม่นยำสูงสุดที่ 96.9% ในขณะที่ในชุดข้อมูลมะเร็งปอดจาก data.world เครื่องเวกเตอร์สนับสนุนให้ความแม่นยำสูงสุดที่ 99.2%

เป้าหมายของงานวิจัย [36] คือการเพิ่มความแม่นยำในการทำนายและค่า Root Mean Square Error (RMSE) ของเวลารอดชีวิตของผู้ป่วยมะเร็งปอด (6 เดือน, 7–24 เดือน, หรือ >24 เดือน) โดยการผสานโมเดลการจำแนกประเภทป่าแบบสุ่มเข้ากับการถดถอยเชิงเส้นทั่วไปและเครื่องเพิ่มประสิทธิภาพเชิงไล่ระดับ (gradient-boosted machines) ป่าแบบสุ่มให้ผลดีที่สุดสำหรับช่วงเวลารอดชีวิต 6 เดือน (RMSE 10.52) และ >24 เดือน (RMSE 20.51) ในขณะที่เครื่องเพิ่มประสิทธิภาพเชิงไล่ระดับเป็นโมเดลที่ดีที่สุดสำหรับช่วง 7–24 เดือน (RMSE 15.65)

นอกจากนี้ ใน [37] ผู้เขียนได้ใช้ตัวจำแนกที่เป็นที่รู้จักกันดีหลายตัว เช่น เครื่องเวกเตอร์สนับสนุน, ต้นไม้ตัดสินใจ C4.5, perceptron แบบหลายชั้น, เครือข่ายประสาทเทียม, และนาอีฟเบย์ กับชุดข้อมูลอ้างอิงจากคลังข้อมูล UCI เพื่อการทำนายมะเร็งปอดในระยะแรก นอกจากนี้ ยังใช้โมเดลรวม เช่น ป่าแบบสุ่มและการลงคะแนนเสียงส่วนใหญ่เพื่อเปรียบเทียบประสิทธิภาพ ผลลัพธ์ที่ได้แสดงให้เห็นว่า ต้นไม้เพิ่มประสิทธิภาพเชิงไล่ระดับ (gradient-boosted tree) มีประสิทธิภาพสูงสุด โดยมีความแม่นยำที่ 90%

ผู้เขียนใน [38] ตั้งเป้าสร้างโมเดลการจำแนกประเภทการทำเหมืองข้อมูลเพื่อทำนายว่าผู้ป่วยมีมะเร็งปอดหรือไม่ โดยใช้ชุดข้อมูล [39] ด้วยวิธีการ Cross Industry Standard Process for Data Mining (CRISP-DM) และซอฟต์แวร์ RapidMiner โมเดลและวิธีการสุ่มตัวอย่างต่าง ๆ ถูกสร้างขึ้น อัลกอริธึมเครือข่ายประสาทเทียมให้ผลดีที่สุด โดยมีความแม่นยำที่ 92% ค่า recall ที่ 94.2% และค่า precision ที่ 90.8% เมื่อเปรียบเทียบกับโมเดลอื่น ๆ

สุดท้ายในงานวิจัย [40] ผู้เขียนได้ออกแบบกลไกเพื่อระบุไบโอมาร์คเกอร์ที่เหมาะสมสำหรับการวินิจฉัยมะเร็งปอดในระยะแรก โดยผสมผสานกลไกเมตาบอลอมิกส์ที่ได้รับการยอมรับและวิธีการเรียนรู้ของเครื่อง การศึกษานี้อ้างอิงจากชุดข้อมูลที่ประกอบด้วยผู้ป่วยมะเร็งปอด 110 ราย และผู้เข้าร่วมที่มีสุขภาพดี 43 ราย เพื่อช่วยในการแยกแยะผู้ป่วยมะเร็งปอดในระยะแรกจากบุคคลที่มีสุขภาพดี ได้เลือกไบโอมาร์คเกอร์เฉพาะ 6 รายการหลังจากการวิเคราะห์ ROC โดยมีค่า AUC, Sensitivity และ Specificity เท่ากับ 0.989, 0.981 และ 1 ตามลำดับ ตัวกรองที่อิงตามความสัมพันธ์อย่างรวดเร็ว (Fast Correlation-Based Filter หรือ FCBF) ถูกใช้เพื่อค้นหาไบโอมาร์คเกอร์ทางเมตาบอลิซึมที่สำคัญ 5 อันดับแรก เมื่อประเมินโมเดลแล้ว นาอีฟเบย์ (Naïve Bayes) ได้รับการเสนอให้เป็นโมเดลที่เหมาะสมที่สุดสำหรับการทำนายเนื้องอกในปอดในระยะแรก

**3. Materials and Methods**

ในส่วนนี้ เราจะอธิบายชุดข้อมูลที่เราใช้เป็นพื้นฐานและขั้นตอนหลักของวิธีการที่ใช้ในการทำนายความเสี่ยงของมะเร็งปอด ซึ่งได้แก่ การปรับสมดุลของข้อมูลและการจัดอันดับคุณลักษณะในข้อมูลที่สมดุล นอกจากนี้ เราจะนำเสนอความถี่ในการเกิดของคุณลักษณะเชิงสัญลักษณ์ที่สัมพันธ์กับประเภทของมะเร็งปอด รวมทั้งอธิบายโมเดลการเรียนรู้ของเครื่อง (ML) และตัวชี้วัดประสิทธิภาพ

**3.1. Dataset Description**

งานวิจัยนี้อ้างอิงจากชุดข้อมูลสาธารณะ [39] โดยมีผู้เข้าร่วมจำนวน 309 คน และคุณลักษณะทั้งหมด (15 รายการเป็นข้อมูลนำเข้าสำหรับโมเดล ML และ 1 รายการเป็นคลาสเป้าหมาย) ได้อธิบายไว้ดังนี้:

* เพศ [41]: คุณลักษณะนี้ระบุว่าเพศของผู้เข้าร่วมเป็นชายหรือหญิง
* อายุ (ปี) [42]: คุณลักษณะนี้บันทึกอายุของผู้เข้าร่วม
* การสูบบุหรี่ [43]: คุณลักษณะนี้ระบุว่าผู้เข้าร่วมสูบบุหรี่หรือไม่
* นิ้วเหลือง [44]: คุณลักษณะนี้ระบุว่าผู้เข้าร่วมมีนิ้วเหลืองหรือไม่
* ความกังวล [45]: คุณลักษณะนี้แสดงว่าผู้เข้าร่วมมีความกังวลหรือไม่
* ความกดดันจากเพื่อน [46]: คุณลักษณะนี้บันทึกว่าผู้เข้าร่วมรู้สึกกดดันจากเพื่อนหรือไม่
* โรคเรื้อรัง [47]: คุณลักษณะนี้แสดงว่าผู้เข้าร่วมมีโรคเรื้อรังหรือไม่
* ความเหนื่อยล้า [48]: คุณลักษณะนี้บ่งชี้ว่าผู้เข้าร่วมมีอาการเหนื่อยล้าหรือไม่
* ภูมิแพ้ [49]: คุณลักษณะนี้ระบุว่าผู้เข้าร่วมมีอาการแพ้หรือไม่
* หายใจหวีด [50]: คุณลักษณะนี้แสดงว่าผู้เข้าร่วมมีอาการหายใจหวีดหรือไม่
* การดื่มแอลกอฮอล์ [51]: คุณลักษณะนี้แสดงว่าผู้เข้าร่วมดื่มแอลกอฮอล์หรือไม่
* อาการไอ [52]: คุณลักษณะนี้แสดงว่าผู้เข้าร่วมมีอาการไอหรือไม่
* หายใจลำบาก [53]: คุณลักษณะนี้ระบุว่าผู้เข้าร่วมมีอาการหายใจลำบากหรือไม่
* กลืนลำบาก [54]: คุณลักษณะนี้แสดงว่าผู้เข้าร่วมมีอาการกลืนลำบากหรือไม่
* เจ็บหน้าอก [55]: คุณลักษณะนี้บันทึกว่าผู้เข้าร่วมมีอาการเจ็บหน้าอกหรือไม่
* มะเร็งปอด: คุณลักษณะนี้ระบุว่าผู้เข้าร่วมได้รับการวินิจฉัยว่าเป็นมะเร็งปอดหรือไม่

คุณลักษณะทั้งหมดเป็นข้อมูลเชิงสัญลักษณ์ ยกเว้นอายุที่เป็นข้อมูลเชิงตัวเลข

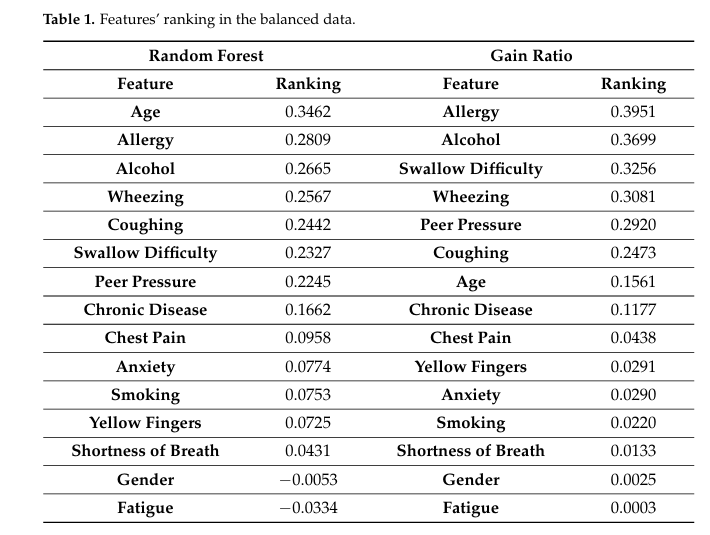
**3.2. Data Preprocessing**

เราต้องสังเกตว่าไม่มีการประมวลผลใด ๆ บนชุดข้อมูลที่เราใช้ เนื่องจากไม่มีค่าที่ขาดหายหรือค่าผิดปกติใด ๆ เพื่อแก้ไขปัญหาการกระจายตัวของคลาสที่ไม่สมดุลระหว่างผู้เข้าร่วมในกลุ่มมะเร็งปอด (87.4%) และกลุ่มที่ไม่เป็นมะเร็งปอด เราได้ใช้วิธี SMOTE [56] ซึ่งเป็นวิธีที่ใช้กันอย่างแพร่หลาย โดยใช้ตัวจำแนก 5-NN เพื่อสร้างข้อมูลสังเคราะห์ [57] สำหรับคลาสที่มีจำนวนน้อยกว่า เช่น กลุ่มที่ไม่เป็นมะเร็งปอด ซึ่งถูกสุ่มตัวอย่างมากขึ้นเพื่อให้จำนวนตัวอย่างในสองกลุ่มมีการกระจายเท่าเทียมกัน (50%–50%)

**3.3. Features Analysis**

ในบริบทของการวิเคราะห์คุณลักษณะ ขั้นแรกเราได้วัดคะแนนความสำคัญของคุณลักษณะทั้งหมดที่เกี่ยวข้องกับคลาสเป้าหมาย สำหรับวัตถุประสงค์นี้ เราได้พิจารณาวิธีการจัดอันดับคุณลักษณะสองวิธี ได้แก่ gain ratio และ random forest เราได้ใช้วิธี gain ratio (GR) [58] ซึ่งให้คะแนนตามสมการ GR(fi) = H(c) - H(c | fi) / H(fi) โดยที่ H(c) คือ entropy ของตัวแปรที่บันทึกค่าของคลาส, H(c | fi) และ H(fi) คือ conditional entropy ของคลาสที่ให้คุณลักษณะ และ entropy ของคุณลักษณะ fi (i = 1, 2, ..., 15) ตามลำดับ ส่วน random forest คำนวณค่า Gini impurity เพื่อวัดความสามารถของคุณลักษณะในการแยกแยะตัวอย่างในสองคลาสอย่างเหมาะสม [59]

คะแนนการจัดอันดับในลำดับจากมากไปน้อยแสดงอยู่ในตารางที่ 1 เราจะเห็นว่าวิธีการทั้งสองได้จัดอันดับคุณลักษณะหกจากทั้งหมดสิบห้าในลำดับความสำคัญเดียวกันตามคะแนนที่ได้มา ขณะที่คุณลักษณะบางส่วนมีการจัดอันดับที่ใกล้เคียงหรือกลับกัน คุณลักษณะที่มีความสำคัญน้อยหรือไม่มีเลยจะได้รับคะแนนใกล้เคียงกับศูนย์หรือเป็นลบ อย่างไรก็ตาม คุณลักษณะทั้งหมดเป็นสัญญาณสำคัญในการเกิดมะเร็งปอดและการจัดการโดยแพทย์ ดังนั้น โมเดลจะถูกฝึกและตรวจสอบความถูกต้องโดยพิจารณาคุณลักษณะทั้งหมด



ในตารางที่แสดงการจัดอันดับคุณลักษณะในชุดข้อมูลที่สมดุลนี้ มีการเปรียบเทียบระหว่างวิธีการ Random Forest และ Gain Ratio ในการจัดลำดับความสำคัญของคุณลักษณะสำหรับการทำนายความเสี่ยงมะเร็งปอด โดยสรุปได้ดังนี้:

1. **Random Forest** ใช้การวัดความสำคัญของคุณลักษณะโดยการคำนวณค่าความไม่บริสุทธิ์ของ Gini (Gini impurity) ซึ่งวัดความสามารถของแต่ละคุณลักษณะในการแยกข้อมูลออกเป็นคลาสต่างๆ อย่างเหมาะสม
2. **Gain Ratio** ใช้การวัดอัตราส่วน gain ซึ่งพิจารณาการลดลงของ entropy ในคลาสเป้าหมายตามคุณลักษณะนั้นๆ

ในตารางนี้เราจะเห็นว่า:

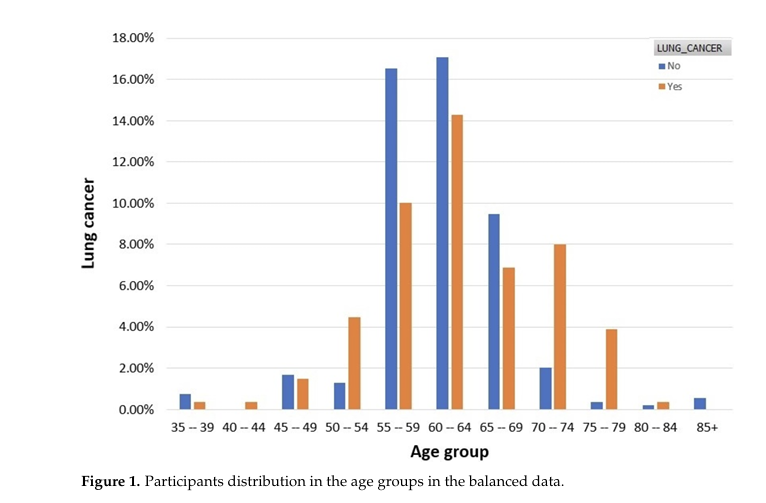
* **คุณลักษณะที่มีความสำคัญมากที่สุดในวิธี Random Forest** คือ "Age" (อายุ) โดยได้คะแนนความสำคัญสูงสุด (0.3462) รองลงมาเป็น "Allergy" (ภูมิแพ้) และ "Alcohol" (การดื่มแอลกอฮอล์)
* **คุณลักษณะที่มีความสำคัญมากที่สุดในวิธี Gain Ratio** คือ "Allergy" (ภูมิแพ้) โดยได้คะแนนสูงสุด (0.3951) รองลงมาเป็น "Alcohol" และ "Swallow Difficulty" (กลืนลำบาก)

สิ่งที่น่าสังเกตคือ บางคุณลักษณะ เช่น "Allergy" และ "Alcohol" ถูกจัดอันดับสูงในทั้งสองวิธี ขณะที่บางคุณลักษณะ เช่น "Gender" (เพศ) และ "Fatigue" (อาการเหนื่อยล้า) มีคะแนนต่ำหรือเป็นค่าลบในวิธี Random Forest แต่ยังคงมีคะแนนบวกในวิธี Gain Ratio

ทั้งนี้ คุณลักษณะทั้งหมดมีความสำคัญในการทำนายความเสี่ยงมะเร็งปอด แต่โมเดลจะฝึกและตรวจสอบความถูกต้องโดยคำนึงถึงคุณลักษณะทุกตัว

ผู้เขียนกล่าวถึง **รูปที่ 1** ซึ่งแสดงการกระจายตัวของผู้เข้าร่วมการทดลองตามกลุ่มอายุ จากการสังเกตพบว่า **มะเร็งปอดมักพบในผู้ที่มีอายุระหว่าง 50 ถึง 79 ปี** ซึ่งในกลุ่มอายุ **60–64 ปี** เป็นกลุ่มที่มีความถี่ของผู้ป่วยสูงที่สุด

กล่าวอีกนัยหนึ่ง ผู้ที่มีอายุในช่วงกลางของชีวิต มีความเสี่ยงสูงที่สุดที่จะเป็นมะเร็งปอดจากข้อมูลการศึกษา



ใน **รูปที่ 1** ซึ่งแสดงการกระจายตัวของผู้เข้าร่วมตามกลุ่มอายุในข้อมูลที่สมดุลเกี่ยวกับมะเร็งปอด มีการเปรียบเทียบระหว่างกลุ่มผู้ที่มีและไม่มีมะเร็งปอด โดยแยกสีเป็นสองแท่งสำหรับแต่ละกลุ่มอายุ:

* แท่งสีฟ้าแทนผู้ที่ **ไม่มีมะเร็งปอด** (Lung cancer = No)
* แท่งสีส้มแทนผู้ที่ **มีมะเร็งปอด** (Lung cancer = Yes)

**ข้อสังเกตสำคัญ**:

1. **กลุ่มอายุ 60-64 ปี** มีอัตราผู้ป่วยมะเร็งปอดสูงที่สุด แสดงให้เห็นว่าผู้ที่อยู่ในกลุ่มอายุดังกล่าวมีความเสี่ยงสูงสุดที่จะถูกวินิจฉัยว่าเป็นมะเร็งปอด
2. **กลุ่มอายุ 55-59 ปี และ 65-69 ปี** มีอัตราผู้ป่วยมะเร็งปอดสูงรองลงมา ซึ่งบ่งชี้ว่าคนในวัยกลางคนถึงวัยชราตอนต้นมีแนวโน้มที่จะป่วยเป็นมะเร็งปอดสูง
3. อัตราการเกิดมะเร็งปอดต่ำมากในกลุ่มอายุที่น้อยกว่า 50 ปี และในกลุ่มอายุที่มากกว่า 70 ปี แม้ว่าจะยังมีผู้ป่วยอยู่บ้างในกลุ่มอายุเหล่านั้น แต่มีจำนวนไม่มากเท่ากับช่วงอายุ 50-79 ปี
4. กลุ่มอายุที่มีความเสี่ยงมากที่สุดคือช่วง **50-79 ปี** โดยเฉพาะอย่างยิ่งในกลุ่มอายุ **60-64 ปี** ซึ่งเป็นกลุ่มที่พบผู้ป่วยมะเร็งปอดมากที่สุด

ภาพนี้แสดงให้เห็นถึงความเสี่ยงที่สูงขึ้นของการเกิดมะเร็งปอดในกลุ่มคนวัยกลางคนถึงวัยชรา และตอกย้ำถึงความสำคัญของการตรวจสุขภาพปอดในช่วงอายุดังกล่าวเพื่อการวินิจฉัยและป้องกันล่วงหน้า

ใน **ตารางที่ 2** มีการแสดงการแสดงออกของคุณลักษณะในแต่ละกลุ่ม (คลาส) โดยสรุปได้ดังนี้:

1. **เพศ**: ผู้ชายมีโอกาสที่จะได้รับการวินิจฉัยว่าเป็นมะเร็งปอดใกล้เคียงกับผู้หญิง ซึ่งหมายความว่าเพศไม่เป็นปัจจัยที่มีความแตกต่างมากนักในการเกิดโรค
2. จากตารางนี้สามารถสรุปได้ว่า **แต่ละคุณลักษณะที่ตรวจสอบ** มีการแสดงออกในผู้ป่วยมะเร็งปอดระหว่าง 26% ถึง 35% ซึ่งหมายถึงอาการต่าง ๆ เช่น การไอ, การหายใจลำบาก ฯลฯ ถูกพบในผู้ป่วยที่ได้รับการวินิจฉัยว่าเป็นมะเร็งปอดในระดับประมาณนี้
3. สิ่งที่น่าสนใจคือมีผู้เข้าร่วมจำนวนมากที่มี **อาการเสี่ยง** โดยที่ยังไม่ได้รับการวินิจฉัยว่าเป็นมะเร็งปอด ซึ่งแสดงให้เห็นว่าการเฝ้าระวังอาการและการตรวจติดตามอาการเสี่ยงเหล่านี้อย่างใกล้ชิด อาจช่วย **ป้องกัน** หรือ **จำกัดผลกระทบ** ที่ไม่พึงประสงค์จากโรคมะเร็งปอดได้

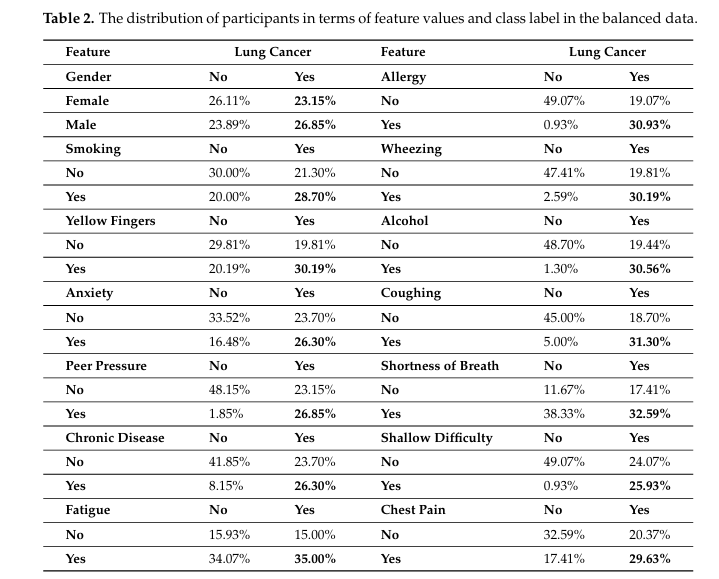
กล่าวอีกนัยหนึ่ง การติดตามและตรวจสอบอาการที่เป็นปัจจัยเสี่ยงของมะเร็งปอดเป็นสิ่งสำคัญ ไม่ว่าจะมีการวินิจฉัยโรคแล้วหรือไม่ เพื่อป้องกันหรือบรรเทาผลกระทบที่ไม่พึงประสงค์จากโรคนี้

**3.4. Machine Learning Models**

ในบทความวิจัยนี้ ผู้เขียนได้ใช้ **โมเดลการเรียนรู้ของเครื่อง (ML) หลายรูปแบบ** เพื่อระบุว่าโมเดลใดที่ทำงานได้ดีกว่าโมเดลอื่น ๆ โดยประเมินจากประสิทธิภาพในการทำนายของพวกมัน โดยเฉพาะอย่างยิ่ง โมเดลที่ถูกพิจารณา ได้แก่:

1. **Naive Bayes (NB)** [60] และ **Bayesian Network (BayesNet)** [61] ซึ่งเป็นตัวจำแนกแบบน่าจะเป็น (probabilistic classifiers)
2. **Logistic Regression (LR)** [62] และ **Logistic Model Tree (LMT)** [63] ซึ่งเป็นโมเดลที่ใช้การถดถอยโลจิสติก
3. **Support Vector Machine (SVM)** [64] ซึ่งเป็นตัวจำแนกที่ใช้แกนในการจัดการข้อมูล
4. **Stochastic Gradient Descent (SGD)** [65] ซึ่งเป็นวิธีการเรียนรู้สำหรับตัวจำแนกเชิงเส้นภายใต้ฟังก์ชัน loss ของ SVM
5. โมเดลที่ใช้ **ต้นไม้ตัดสินใจ (decision tree)** เช่น **J48** [66], **Random Tree (RT)** [67], **Rotation Forest (RotF)** [68], และ **Reduced Error Pruning Tree (RepTree)** [69]
6. โมเดลจาก **อัลกอริธึมการรวมหลายโมเดล (ensemble ML algorithms)** เช่น **Random Forest (RF)** [71] และ **AdaBoostM1** [72]
7. โมเดล **เครือข่ายประสาทเทียมอย่างง่าย (Artificial Neural Network)** อย่าง **Multi-layer Perceptron (MLP)** [73] และ **K-nearest neighbors (K-NN)** [74] ซึ่งเป็นตัวจำแนกที่ใช้ระยะทางเป็นตัววัด

โมเดลเหล่านี้ถูกทดสอบเพื่อค้นหาว่าโมเดลใดสามารถทำนายได้แม่นยำที่สุดจากการประเมินประสิทธิภาพ ซึ่งแสดงถึงความหลากหลายในวิธีการจัดการข้อมูลของแต่ละโมเดล



ใน **ตารางที่ 2** แสดงการกระจายของผู้เข้าร่วมการศึกษาในแง่ของค่าคุณลักษณะต่างๆ และคลาสที่เกี่ยวข้องกับมะเร็งปอดในข้อมูลที่สมดุล โดยมีการเปรียบเทียบผู้ที่ **เป็นมะเร็งปอด (Yes)** และ **ไม่เป็นมะเร็งปอด (No)** สำหรับแต่ละคุณลักษณะ

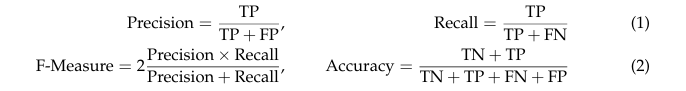
**คำอธิบายแต่ละคุณลักษณะ:**

1. **Gender (เพศ)**:
   * ผู้หญิงที่เป็นมะเร็งปอด: 23.15% และไม่เป็น: 26.11%
   * ผู้ชายที่เป็นมะเร็งปอด: 26.85% และไม่เป็น: 23.89%
   * สรุป: ทั้งเพศชายและหญิงมีโอกาสที่จะเป็นมะเร็งปอดใกล้เคียงกัน
2. **Smoking (การสูบบุหรี่)**:
   * ผู้ที่สูบบุหรี่และเป็นมะเร็งปอด: 28.70% และไม่เป็น: 20.00%
   * ผู้ที่ไม่สูบบุหรี่และเป็นมะเร็งปอด: 21.30% และไม่เป็น: 30.00%
   * สรุป: การสูบบุหรี่มีความสัมพันธ์กับโอกาสเป็นมะเร็งปอดสูงขึ้น
3. **Allergy (ภูมิแพ้)**:
   * ผู้ที่มีภูมิแพ้และเป็นมะเร็งปอด: 30.93% และไม่เป็น: 0.93%
   * ผู้ที่ไม่มีภูมิแพ้และเป็นมะเร็งปอด: 19.07% และไม่เป็น: 49.07%
   * สรุป: ผู้ที่มีภูมิแพ้มีความสัมพันธ์กับการเกิดมะเร็งปอดมากกว่า
4. **Wheezing (หายใจหวีด)**:
   * ผู้ที่มีอาการหายใจหวีดและเป็นมะเร็งปอด: 30.19% และไม่เป็น: 1.30%
   * ผู้ที่ไม่มีอาการหายใจหวีดและเป็นมะเร็งปอด: 19.81% และไม่เป็น: 47.41%
   * สรุป: อาการหายใจหวีดมีความสัมพันธ์กับการเป็นมะเร็งปอดอย่างมาก
5. **Alcohol (การดื่มแอลกอฮอล์)**:
   * ผู้ที่ดื่มแอลกอฮอล์และเป็นมะเร็งปอด: 30.19% และไม่เป็น: 20.19%
   * ผู้ที่ไม่ดื่มแอลกอฮอล์และเป็นมะเร็งปอด: 19.44% และไม่เป็น: 48.70%
   * สรุป: การดื่มแอลกอฮอล์มีความสัมพันธ์กับการเกิดมะเร็งปอด
6. **Coughing (อาการไอ)**:
   * ผู้ที่มีอาการไอและเป็นมะเร็งปอด: 31.30% และไม่เป็น: 5.00%
   * ผู้ที่ไม่มีอาการไอและเป็นมะเร็งปอด: 18.70% และไม่เป็น: 45.00%
   * สรุป: อาการไอมีความสัมพันธ์กับการเกิดมะเร็งปอดสูง
7. **Shortness of Breath (หายใจลำบาก)**:
   * ผู้ที่มีอาการหายใจลำบากและเป็นมะเร็งปอด: 32.59% และไม่เป็น: 11.67%
   * ผู้ที่ไม่มีอาการหายใจลำบากและเป็นมะเร็งปอด: 17.41% และไม่เป็น: 38.33%
   * สรุป: อาการหายใจลำบากเป็นสัญญาณที่เกี่ยวข้องกับมะเร็งปอด
8. **Chest Pain (เจ็บหน้าอก)**:
   * ผู้ที่มีอาการเจ็บหน้าอกและเป็นมะเร็งปอด: 29.63% และไม่เป็น: 17.41%
   * ผู้ที่ไม่มีอาการเจ็บหน้าอกและเป็นมะเร็งปอด: 20.37% และไม่เป็น: 32.59%
   * สรุป: อาการเจ็บหน้าอกมีความสัมพันธ์กับการเกิดมะเร็งปอด
9. **Chronic Disease (โรคเรื้อรัง)**:
   * ผู้ที่มีโรคเรื้อรังและเป็นมะเร็งปอด: 26.30% และไม่เป็น: 8.15%
   * ผู้ที่ไม่มีโรคเรื้อรังและเป็นมะเร็งปอด: 23.70% และไม่เป็น: 41.85%
   * สรุป: โรคเรื้อรังมีความสัมพันธ์กับการเกิดมะเร็งปอด

**สรุป:**

แต่ละคุณลักษณะมีความสัมพันธ์กับความเป็นไปได้ในการเกิดมะเร็งปอด ตัวอย่างเช่น อาการไอ, หายใจลำบาก, และการสูบบุหรี่เป็นปัจจัยเสี่ยงที่พบได้บ่อยในผู้ป่วยมะเร็งปอด ข้อมูลเหล่านี้สามารถใช้ในการตรวจสอบและป้องกันความเสี่ยงการเกิดมะเร็งปอดได้

**3.5.Evaluation Metrics**

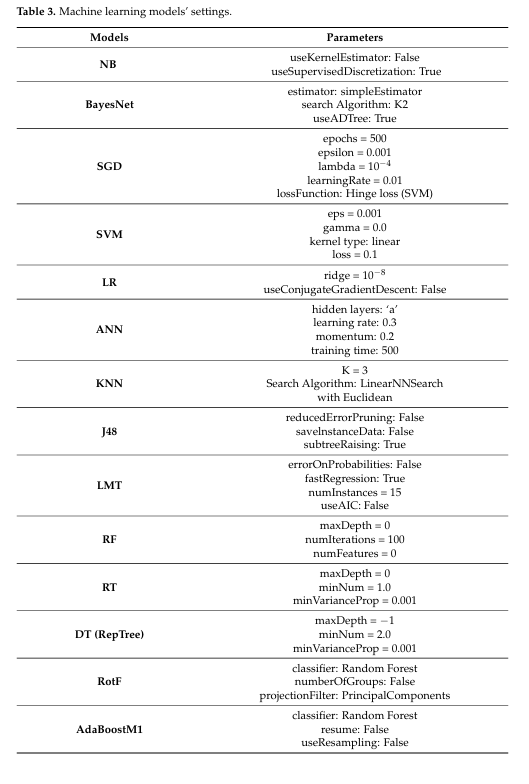
****

**ความแม่นยำ (Accuracy)** สรุปประสิทธิภาพของการทำงานในงานการจำแนกประเภท และวัดจำนวนตัวอย่างที่ทำนายถูกต้องจากข้อมูลทั้งหมด นอกจากนี้ เรายังได้ตรวจสอบ **การเรียกกลับ (Recall)** ซึ่งแสดงอัตราบวกจริงหรือความไวของโมเดลในการระบุผู้เข้าร่วมที่เป็นมะเร็งปอดจริงและถูกพิจารณาเป็นบวกเมื่อเทียบกับผู้เข้าร่วมที่เป็นบวกทั้งหมด **ความแม่นยำ (Precision)** เป็นการวัดคุณภาพ ในขณะที่การเรียกกลับเป็นการวัดปริมาณ **ค่า F-Measure** เป็นค่าเฉลี่ยเชิงฮาร์โมนิกของความแม่นยำและการเรียกกลับ ซึ่งช่วยให้โมเดลได้รับการประเมินด้วยคะแนนเดียว สุดท้าย **AUC** อยู่ในช่วงระหว่างศูนย์ถึงหนึ่ง ใช้เพื่อกำหนดโมเดล ML ที่มีประสิทธิภาพดีที่สุดในการแยกแยะระหว่างกรณีมะเร็งปอดและไม่เป็นมะเร็งปอด AUC เป็นตัวชี้วัดการแยก หาก AUC ถึงหนึ่ง หมายความว่าโมเดลนั้นมีความสามารถในการแยกสองกลุ่มอย่างสมบูรณ์แบบ

**4.Results and Discussion**

4.1.Experiments Setup

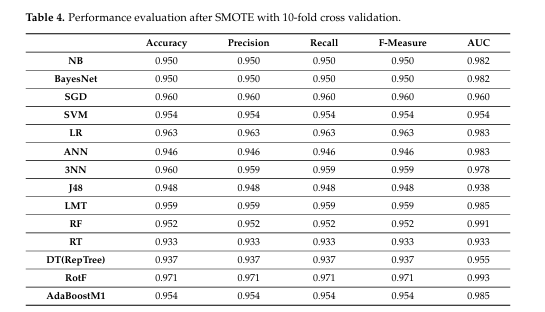
ประสิทธิภาพของโมเดลการเรียนรู้ของเครื่อง (ML) ได้รับการประเมินในสภาพแวดล้อมของ Weka [76] ซึ่งมีไลบรารีหลากหลายสำหรับการประมวลผลข้อมูลเบื้องต้น, การจำแนกประเภท, การจัดกลุ่ม, การทำนาย และการแสดงผล นอกจากนี้ การทดลองได้ดำเนินการบนระบบคอมพิวเตอร์ที่มีสเปคดังนี้: Intel(R) Core(TM) i7-1165G7 รุ่นที่ 11 @ 2.80 GHz, RAM 16GB, ระบบปฏิบัติการ Windows 11 Home แบบ 64 บิต และโปรเซสเซอร์แบบ x64 เราได้ใช้การตรวจสอบข้าม 10 ส่วน (10-fold cross validation) และเทคนิค SMOTE เพื่อวัดประสิทธิภาพของโมเดลบนชุดข้อมูลที่สมดุลจำนวน 540 ตัวอย่าง สุดท้ายในตารางที่ 3 เราได้แสดงการตั้งค่าพารามิเตอร์ที่เหมาะสมที่สุดของโมเดล ML ที่เสนอ



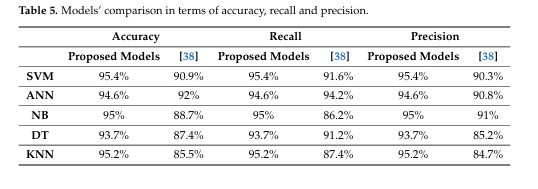
**4.2.Evaluation**

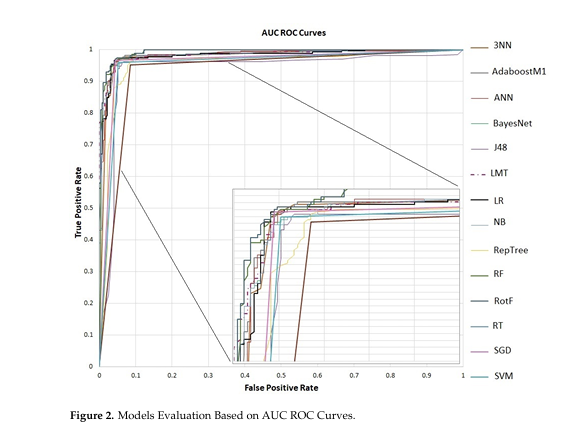
ในบริบทของงานวิจัยนี้ โมเดลการเรียนรู้ของเครื่องหลายชนิด เช่น NB, BayesNet, SGD, SVM, LR, ANN, KNN, J48, LMT, RF, RT, RepTree, RotF และ AdaBoostM1 ถูกประเมินในแง่ของความแม่นยำ (Accuracy), ความแม่นยำของการจำแนก (Precision), การเรียกกลับ (Recall), ค่า F-Measure และ AUC เพื่อกำหนดว่าโมเดลใดมีประสิทธิภาพในการทำนายที่ดีที่สุด

โดยเฉพาะในตารางที่ 4 เราแสดงผลการประเมินประสิทธิภาพของโมเดลหลังจากใช้ SMOTE และการตรวจสอบข้าม 10 ส่วน (10-fold cross-validation) โมเดลที่เรานำเสนอทั้งหมดมีเปอร์เซ็นต์สูงกว่า 93.3% (RT) ผลลัพธ์ที่ดีที่สุดเกิดจากโมเดล RotF ซึ่งใช้ RF เป็นฐานการจำแนก โดยมีค่าความแม่นยำ, ความแม่นยำของการจำแนก, การเรียกกลับ และค่า F-Measure เท่ากับ 97.1% และ AUC เท่ากับ 99.3% นอกจากนี้ ควรสังเกตว่าค่า AUC สูง ๆ ยังทำได้โดย RF ที่มีค่า 99.1% และ AdaBoostM1 ที่มีค่า 98.5% ซึ่งใช้ RF เป็นฐานการจำแนก สุดท้าย ในรูปที่ 2 เราได้ทำกราฟโค้ง AUC-ROC ของโมเดลการเรียนรู้ของเครื่องที่เราเสนอ ซึ่งแสดงให้เห็นถึงประสิทธิภาพที่เหนือกว่าของ RotF



นอกจากนี้ ในตารางที่ 5 มีการเปรียบเทียบโมเดลในแง่ของความแม่นยำ, การเรียกกลับ และความแม่นยำของการจำแนก ผู้เขียนในงานวิจัย [38] ได้ใช้ชุดข้อมูล [39] ที่มีจำนวนคุณลักษณะเท่ากับของเรา ผลลัพธ์ของโมเดลของพวกเขาถูกวัดหลังจากใช้การตรวจสอบข้าม 10 ส่วน (10-fold cross-validation) โมเดลที่เราเสนอมีผลลัพธ์ที่ดีกว่าในทั้งสามตัวชี้วัดเมื่อเปรียบเทียบกับโมเดลในงานวิจัยที่กล่าวถึงข้างต้น โดยเฉพาะอย่างยิ่ง ประสิทธิภาพที่ดีที่สุดของโมเดลที่เรานำเสนอในแง่ของความแม่นยำ, การเรียกกลับ และความแม่นยำของการจำแนกคือโมเดล SVM ที่มีค่า 95.4% ในขณะที่ใน [38] ผลลัพธ์ที่ดีที่สุดในตัวชี้วัดเดียวกันได้มาจากโมเดล ANN โดยมีค่าความแม่นยำ 92%, การเรียกกลับ 94.2%, และความแม่นยำของการจำแนก 90.8% ในทั้งสามตัวชี้วัด โมเดลที่เราเสนอนั้นทำได้ดีกว่า





**4.3. Discussion**

วิธีการที่นำเสนอในงานวิจัยนี้อ้างอิงจากชุดข้อมูลที่ประกอบด้วยคุณลักษณะที่สะท้อนถึงพฤติกรรมของมนุษย์ (เช่น การสูบบุหรี่และการดื่มแอลกอฮอล์) และอาการ/อาการแสดงที่เป็นปัจจัยเสี่ยงที่ผู้ป่วยมะเร็งปอดมักจะเผชิญ อย่างไรก็ตาม อาการเหล่านี้ไม่จำเป็นต้องเกี่ยวข้องกับโรคมะเร็งปอดเสมอไป ดังที่เราได้สังเกตจากการวิเคราะห์คุณลักษณะในส่วนที่ 3.3 ของวัสดุและวิธีการ ไม่เหมือนกับมะเร็งชนิดอื่น ๆ มะเร็งปอดไม่สามารถมองเห็นได้ด้วยตาเปล่า และอาการของมันมักจะเกิดร่วมกับอาการของโรคอื่น อาการที่พบบ่อยที่สุด ได้แก่ ภูมิแพ้ หอบหืด หายใจไม่ออก และไอ [33] ในงานวิจัยนี้ เราได้เลือกฝึกตัวจำแนกหลายตัวบนปัจจัยเสี่ยงต่าง ๆ ที่เกี่ยวข้องกับอาการเหล่านี้เพื่อให้สามารถระบุป้ายชื่อคลาส (มะเร็งปอดหรือไม่ใช่มะเร็งปอด) ของตัวอย่างที่ไม่รู้จักได้อย่างถูกต้อง และเพื่อประเมินความเสี่ยงที่เกี่ยวข้อง แม้ว่าจะยังไม่แสดงอาการก็ตาม การเฝ้าติดตามปัจจัยเสี่ยงและการตรวจสุขภาพติดตามผลเป็นวิธีปฏิบัติที่เหมาะสมสำหรับการจัดการมะเร็งปอด ซึ่งอาจช่วยป้องกันหรือจำกัดผลกระทบที่ไม่พึงประสงค์ของโรคผ่านการวินิจฉัยในระยะเริ่มต้น การตรวจและระบุโรคมะเร็งปอดมักเกิดขึ้นเมื่อทำการถ่ายภาพเอกซเรย์, CT, PET-CT และ MRI ของทรวงอกผู้ป่วย [77] ดังนั้น ชุดข้อมูลที่พิจารณาร่วมกับคุณลักษณะที่ได้จากภาพปอดจะเป็นประโยชน์อย่างมากต่อการวินิจฉัยโรคและระยะของโรคตั้งแต่เนิ่น ๆ ขอให้เราทบทวนว่าการศึกษานี้มีเป้าหมายเพื่อระบุว่ามีการเกิดมะเร็งปอดหรือไม่ ดังนั้นจึงเป็นปัญหาการจำแนกประเภทแบบสองค่าที่ได้ถูกศึกษา จากมุมมองการเรียนรู้ของเครื่อง การระบุระยะของมะเร็งสามารถแก้ได้ด้วยกลยุทธ์การจำแนกประเภทหลายคลาส เช่น วิธีหนึ่งต่อหนึ่ง (OVO) และหนึ่งต่อทั้งหมด (OVA) [25] อย่างไรก็ตาม ชุดข้อมูลที่พิจารณาไม่อนุญาตให้เราแก้ปัญหาในลักษณะดังกล่าว

ไม่ต้องสงสัยเลยว่า การเรียนรู้ของเครื่องได้กลายเป็นเครื่องมือสำคัญสำหรับผู้ดูแลสุขภาพและแพทย์ในการคัดกรองล่วงหน้า การทำนาย และ/หรือการคาดการณ์โรคหลายชนิด ความพยายามอย่างมากได้ถูกทำโดยนักวิจัยเพื่อเข้าถึงข้อมูลทางการแพทย์จากบันทึกสุขภาพของบุคคล รวบรวมข้อมูลผ่านแบบสอบถาม หรือสร้างชุดข้อมูลในห้องปฏิบัติการของตนเอง เพื่อสนับสนุนการวิเคราะห์ด้านสุขภาพผ่านการฝึกและทดสอบโมเดลที่เหมาะสม ซึ่งจะให้ข้อมูลเชิงลึกเกี่ยวกับการพัฒนาและการป้องกันโรคในอนาคต ตัวอย่างเช่น ในการศึกษาล่าสุดของเรา [32] ได้ฝึกตัวจำแนกหลายตัวเกี่ยวกับการทำนายโรคไตเรื้อรัง (CKD) ในขณะที่ในการศึกษานี้ มะเร็งปอดเป็นสภาวะสุขภาพที่เรามุ่งเน้น สองกรณีนี้แสดงให้เห็นถึงความยืดหยุ่นและความหลากหลายในแง่ของการนำการเรียนรู้ของเครื่องไปประยุกต์ใช้ในด้านสุขภาพ ไม่ว่าในกรณีของข้อมูลและโรคที่เกี่ยวข้อง เมื่อปรับสมดุลข้อมูลด้วย SMOTE โมเดลทั้งหมดแสดงประสิทธิภาพที่สูงในทุกตัวชี้วัด นอกจากนี้ ผลลัพธ์ที่มีความหวังยังได้มาจากการใช้โมเดลการรวมกันแบบ stacking และ voting ดังที่แสดงใน [32] ซึ่งไม่ได้ถูกนำมาศึกษาในที่นี้ จากโมเดลต้นไม้ การมีอยู่ของตัวจำแนกแบบ Rotation Forest ได้รับการยืนยันทั้งในกรณีของมะเร็งปอดและ CKD

สรุปผลและการอภิปราย เราต้องชี้ให้เห็นถึงข้อจำกัดของบทความนี้ งานวิจัยนี้อ้างอิงจากชุดข้อมูลที่เปิดเผยต่อสาธารณะ [39] และไม่ได้มาจากหน่วยงานของโรงพยาบาลหรือสถาบันใด ๆ ซึ่งอาจให้ข้อมูลที่หลากหลายมากกว่านี้ นอกจากนี้ การเข้าถึงข้อมูลทางการแพทย์ที่เป็นความลับนั้นยากเนื่องจากเหตุผลด้านความเป็นส่วนตัว อย่างไรก็ตาม ชุดข้อมูลที่เราใช้มีคุณลักษณะที่เป็นประโยชน์ ซึ่งนำไปสู่การสรุปผลการวิจัยที่เชื่อถือได้และแม่นยำ

**5.Conclusions**

ปอดเป็นอวัยวะหลักในการหายใจ มนุษย์จะไม่หยุดหายใจจนกว่าจะเสียชีวิต เพราะปอดทำหน้าที่ส่งออกซิเจนไปยังเลือด ซึ่งมีความสำคัญต่อการมีชีวิต มะเร็งปอดเป็นสาเหตุการเสียชีวิตจากมะเร็งที่พบมากที่สุดในทั้งสองเพศ ช่วงชีวิตของผู้ป่วยถูกกำหนดโดยระยะที่ลุกลามของมะเร็ง ยิ่งได้รับการวินิจฉัยเร็วเท่าใด ยิ่งมีอายุคาดหวังที่ยาวนานขึ้นเท่านั้น

ในงานวิจัยนี้ เราใช้การเรียนรู้แบบมีผู้สอนเพื่อพัฒนาโมเดลสำหรับการระบุบุคคลที่แสดงอาการของมะเร็งปอดตามคุณลักษณะและอาการต่าง ๆ โมเดลการเรียนรู้ของเครื่องหลากหลายชนิด รวมถึง NB, BayesNet, SGD, SVM, LR, ANN, KNN, J48, LMT, RF, RT, RepTree, RotF และ AdaBoostM1 ได้รับการประเมินในแง่ของความแม่นยำ, ความแม่นยำของการจำแนก, การเรียกกลับ, ค่า F-Measure และ AUC จากผลการทดลองและหลังจากใช้ SMOTE พร้อมกับการตรวจสอบข้าม 10 ส่วน (10-fold cross-validation) โมเดล RotF มีประสิทธิภาพดีกว่าโมเดลอื่น ๆ ด้วยความแม่นยำ, ความแม่นยำของการจำแนก, การเรียกกลับ และค่า F-Measure เท่ากับ 97.1% และ AUC เท่ากับ 99.3% นอกจากนี้ โมเดลที่เรานำเสนอมีผลการทำงานที่ดีกว่าโมเดลอ้างอิง [38] ตามที่แสดงในตารางที่ 5

ในงานวิจัยในอนาคต เรามีเป้าหมายที่จะขยายการศึกษาปัจจุบันออกเป็นสองแนวทาง แนวทางแรกคือการเสริมสร้างกรอบการทำงานของการเรียนรู้ของเครื่องด้วยการใช้วิธีการเรียนรู้เชิงลึก โดยเฉพาะอย่างยิ่งการใช้ Long Short-Term Memory (LSTM) และ Convolutional Neural Networks (CNN) และเปรียบเทียบผลลัพธ์ในแง่ของความแม่นยำกับงานวิจัยในขอบเขตเดียวกัน แนวทางที่สองคือการประเมินโมเดลการจำแนกประเภทในชุดข้อมูลเดียวกัน โดยสมมติว่ามีการใช้กระบวนการบูตสแตรปปิ้ง [78] นอกเหนือจากการตรวจสอบข้าม 10 ส่วนที่มีอยู่ ซึ่งเป็นวิธีการแบ่งข้อมูลแบบทางเลือกสำหรับการตรวจสอบความถูกต้องของโมเดล ซึ่งจะใช้การสุ่มตัวอย่างแบบทดแทนในข้อมูลต้นฉบับ