นิพนธ์ต้นฉบับ
Original Article

การคัดกรองมะเร็งปากมดลูกด้วยเทคนิคเหมืองข้อมูล

สฤษฏ์ชัย ปรีดาวัลย์, ปิ่นกมล สมพีร์วงศ์ ภ.ม.***

บทคัดย่อ

มะเร็งปากมดลูกเป็นโรคมะเร็งที่สำคัญชนิดหนึ่งในมะเร็งชนิดอื่น ๆ ของสุภาพสตรีวันนี้ การตรวจคัดกรองมะเร็งปากมดลูกในอดีตมีวิธีการหลายชนิด เช่นประวัติทางการแพทย์ การตรวจหาไวรัสเอชพีวี (Human Papilloma Virus : HPV) ชนิดความเสี่ยงสูง สารคัดหลั่งของร่างกาย การตรวจแป๊ปเสมียร์ (PAP Smear) และการตัดชิ้นเนื้อเล็ก ๆ ในการวิจัยครั้งนี้ ผู้วิจัยขอเสนอวิธีการตรวจคัดกรองมะเร็งปากมดลูกโดยใช้วิธีการทำเหมืองข้อมูลด้วยแอนท์ไมเนอร์อัลกอริทึม (Ant-Miner Algorithm) มีวัตถุประสงค์เพื่อค้นหาเทคนิคการทำเหมืองข้อมูลเพื่อสร้างแบบจำลองการตรวจคัดกรองมะเร็งปากมดลูกที่มีประสิทธิภาพในการจำแนก และการคัดเลือกคุณลักษณะด้วยค่าสหสัมพันธ์ของคุณลักษณะ (Correlation-based Feature Selection : CFS- เซตของคุณลักษณะที่ดีจะบรรจุคุณลักษณะที่มีความสัมพันธ์อย่างสูงกับคำตอบ) สำหรับการทำเหมืองข้อมูลการวิจัยนี้ใช้ชุดข้อมูลทางการแพทย์ (มีคุณลักษณะ 32 ค่า 4 คำตอบ จำนวน 858 ตัวอย่าง) พบว่า ค่าสหสัมพันธ์คุณลักษณะ (CFS) มีความผลต่อการเลือกคุณลักษณะทำให้ระบุคุณลักษณะได้อย่างรวดเร็ว กรองคุณลักษณะที่ไม่เกี่ยวข้อง คุณลักษณะที่ซ้ำซ้อนและคุณลักษณะที่ไม่สมบูรณ์ และความสัมพันธ์ของแต่ละคุณลักษณะไม่ขึ้นอยู่กับคุณลักษณะอื่น ๆ วิธีการคัดเลือกคุณลักษณะแบบ CFS ช่วยให้มีรายการคุณลักษณะมีขนาดเล็กลง แต่มีประสิทธิภาพในการคัดกรองมะเร็งปากมดลูกด้านความถูกต้องและความแม่นยำผลการวิจัย พบว่า อายุ (Age) จำนวนของคู่ทางเพศสัมพันธ์ (Number of Sexual Partners) อายุที่มีเพศสัมพันธ์ครั้งแรก (Age at 1st Sexual Coitus) จำนวนการตั้งครรภ์ (Number of Parturition) จำนวนปีการคุมกำเนิดด้วยฮอร์โมน (Hormonal Contraception) และจำนวนปีการใส่ห่วงคุมกำเนิด (IUDS) เป็นรายการคุณลักษณะหลักของแบบจำลองการทำนายผลการคัดกรองมะเร็งปากมดลูกด้วยแอนท์ไมเนอร์อัลกอริทึม มีความถูกต้องรวมทุกวิธีเฉลี่ยร้อยละ 94.68 และความแม่นยำเฉลี่ยร้อยละ 93.78 เมื่อพิจารณาเป็นรายวิธี พบว่า วิธี Hinselmann มีความถูกต้องร้อยละ 93.26 ความแม่นยำร้อยละ 90.00 วิธี Schiller มีความถูกต้องร้อยละ 90.86 ความแม่นยำร้อยละ 95.24 วิธี Cytology มีความถูกต้องร้อยละ 96.26 ความแม่นยำร้อยละ 92.10 และวิธี Biopsy มีความถูกต้องร้อยละ 98.35 ความแม่นยำร้อยละ 97.78 การทำเหมืองข้อมูลสำหรับการคัดกรองมะเร็งปากมดลูกด้วยแอนท์ไมเนอร์อัลกอริทึมเป็นเครื่องมือวินิจฉัยโรคชนิดหนึ่งที่มีประสิทธิภาพ

คำสำคัญ : มะเร็งปากมดลูก, ความสัมพันธ์ของข้อมูล, เหมืองข้อมูล

Cervical Cancer Screening using Data Mining Technique

Saritchai Predawan*, Pinkamon Sompeewong M.Pharm.**

Abstract

Cervical cancer is one of the most popular disease among other cancers in female these days. Previous screening diagnosis of the cervical cancer has been done by several methods, medical history, HPV high risk type testing, body fluids, PAP smear and tissue biopsy. In this paper, the authors have proposed a cervical cancer screening diagnostic method by using data mining with Ant-Miner Algorithm. The objective was to search the data mining techniques to create a cervical cancer screening model of efficiency in the classification and feature selection for data mining method through a correlation-based approach. This Experiments on medical datasets (There are 32 attributes, 4 classes with 858 samples) showed that Correlation based Feature Selection (CFS- good feature sets contain attributes that are highly correlated with the class) quickly identifies and screens irrelevant, redundant,

* *วิทยาจารย์ชำนาญการพิเศษ*
** *เภสัชกรชำนาญการ วิทยาลัยการสาธารณสุข สิรินธร จังหวัดชลบุรี*
* *Instructor, Senior Professional Level*
** *Pharmacist, Professional Level, Sirindhorn College of Public Health, Chonburi*
Received : July 19, 2020 Revised : Dec 26, 2020 Accepted : Dec 30, 2020

and noisy features, and identifies relevant features as long as their relevance does not strongly depend on other features. CFS help by providing a smaller number of features with high performance of cervical cancer screening by accuracy and precision. The results show that age, number of sexual partners, age at 1st sexual coitus, number of parturitions, hormonal contraception and IUDS are the main predictive features of cervical cancer screening model with average high accuracy with 94.68% and average precision with 93.78%. And when considering by type of class found that information, the accuracy of Hinselmann class was 93.26%, with a precision of 90.00%, the accuracy of Schiller class was 90.86%, with a precision of 95.24%, the accuracy of Cytology class was 96.26%, with a precision of 92.10% and the accuracy of Biopsy class was 98.35%, with a precision of 97.78% respectively. Data mining with Ant-Miner Algorithm is shown to be advantageous in handling a cervical cancer screening diagnostic assignment with excellent performance.

Keywords : Cervical Cancer, Correlation of Data, Data Mining

บทนำ

ปัญหาสาธารณสุขที่สำคัญของโลก[1] พบว่าโรคมะเร็งเป็นโรคหนึ่งที่เป็นสาเหตุสำคัญของการเสียชีวิต องค์กรระหว่างประเทศเพื่อการวิจัยโรคมะเร็ง หรือ IARC ขององค์การอนามัยโลกรายงานสถานการณ์มะเร็งทั่วโลกประจำปี 2018 ซึ่งครอบคลุมการสำรวจสถานการณ์ของโรคมะเร็ง 36 ประเภท ใน 185 ประเทศทั่วโลก พบผู้ป่วยที่ได้รับการวินิจฉัยว่าเป็นโรคมะเร็งรายใหม่เพิ่มขึ้นมากกว่า 18 ล้านราย และมีผู้เสียชีวิตจากโรคประมาณ 9.6 ล้านราย ซึ่งอาจกล่าวได้ว่า ทุก ๆ 6 รายที่เป็นโรคมะเร็งจะเสียชีวิต 1 ราย ที่น่าสนใจ คือ ตัวเลขของผู้ป่วยด้วยโรคมะเร็งรายใหม่ เกือบครึ่งหนึ่งและตัวเลขผู้เสียชีวิตจากโรคมะเร็งเกินครึ่งหนึ่งนั้นอยู่ในทวีปเอเชีย มีการประเมินว่าภายในปี 2040 หรืออีกประมาณ 20 ปีข้างหน้า จะมีผู้ป่วยโรคมะเร็งรายใหม่มากถึง 29.3 ล้านรายและอัตราการเสียชีวิตจะเพิ่มสูงขึ้นเป็น 16.3 ล้านราย นอกจากนี้ ยังพบว่า 1 ใน 5 ของเพศชาย และ 1 ใน 6 ของเพศหญิง จะเป็นโรคมะเร็งในช่วงชีวิตหนึ่ง

ในประเทศไทย ภาพรวมสถานการณ์ของโรคมะเร็งจากสถิติ[2] พบว่า โรคมะเร็งเป็นสาเหตุการเสียชีวิตอันดับ 1 คิดเป็นร้อยละ 16 ของเหตุการณ์เสียชีวิตทั้งหมด สูงกว่าอัตราการเสียชีวิตจาก อุบัติเหตุ และโรคหัวใจเฉลี่ย 2 ถึง 3 เท่าหรือมีผู้เสียชีวิตจากโรคมะเร็งเฉลี่ย 8 รายต่อชั่วโมง ในปี พ.ศ. 2561 พบว่ามีจำนวนผู้ป่วยรายใหม่วันละ 336 ราย หรือ 122,757 รายต่อปี และเสียชีวิตจากโรคมะเร็งวันละ 215 ราย หรือ 78,540 รายต่อปี และจากข้อมูลการเบิกจ่ายค่าบริการโรคมะเร็งในระบบหลักประกันสุขภาพแห่งชาติ ในช่วงเวลาระหว่างปี พ.ศ. 2559-2561 มีผู้ป่วยมะเร็งเข้าถึงการรักษาอย่างต่อเนื่อง จำนวน 4,117,504 ครั้ง ชดเชยค่ารักษาจำนวน 26,679 ล้านบาท ทั้งนี้เฉพาะในปี พ.ศ. 2561 มีผู้ป่วยเข้ารับการรักษา จำนวน 234,116 ราย รับการบริการรักษา จำนวน 1,431,759 ครั้ง ชดเชยค่ารักษาจำนวน 9,557 ล้านบาท ทำให้ประเทศต้องเสียค่าใช้จ่ายจำนวนมากเพื่อการรักษาสุขภาพของประชาชน และยังส่งผลกระทบต่อการบริหารประเทศในด้านอื่น ๆ ด้วย สำหรับ 5 อันดับแรกของมะเร็งที่พบบ่อยที่สุด[3] ได้แก่ มะเร็งปอด มะเร็งตับและท่อน้ำดี มะเร็งเต้านม มะเร็งลำไส้ใหญ่ และ มะเร็งปากมดลูก โดยโรคมะเร็งที่ทำให้เสียชีวิตมากที่สุด 5 อันดับแรก ได้แก่ มะเร็งตับและท่อน้ำดี มะเร็งปอด มะเร็งเต้านม มะเร็งลำไส้ใหญ่ และมะเร็งปากมดลูก ตามลำดับ

เหมืองข้อมูล (Data Mining) เป็นวิธีการสกัดข้อมูล (Data Extraction) ออกจากข้อมูลขนาดใหญ่ เพื่อให้ได้สารสนเทศ (Information) หรือความสัมพันธ์ของข้อมูล (Correlation of Data) ที่ยังไม่เคยค้นพบ ข้อมูลสารสนเทศที่ได้มีปัจจัยสำคัญที่มีอิทธิพลต่อการตัดสินใจเลือกคำตอบหรือแก้ไขปัญหา กระบวนทางเหมืองข้อมูลมีสำคัญในการทำ Knowledge Discovery in Database (KDD) แบบจำลองที่ได้จากเหมืองข้อมูล มี 5 ขั้นตอน คือ 1) การเข้าใจปัญหาหรือข้อมูล 2) การจัดเตรียมข้อมูล 3) การสร้างแบบจำลอง 4) การประเมินแบบจำลอง และ 5) การนำแบบจำลองไปใช้ และการตรวจสอบผลว่าบรรลุเป้าหมายที่กำหนดไว้เพียงใด วิธีการที่ส่งผลต่อประสิทธิภาพของแบบจำลองมากที่สุด คือ การคัดเลือกคุณลักษณะ (Feature Selection) อยู่ในขั้นตอนการเตรียมข้อมูลเพื่อช่วยเพิ่มประสิทธิภาพของอัลกอริทึมการเรียนรู้ให้มีความถูกต้องมากขึ้น โดยการนำเอาคุณสมบัติที่ไม่เกี่ยวข้องและซ้ำซ้อนออก ปรับปรุงการเลือกคุณลักษณะ ลดขนาดมิติ แล้วยังมีผลให้การสร้างแบบจำลองทำได้รวดเร็วขึ้น

การคัดกรองมะเร็งที่มีความสำคัญในระยะเริ่มต้นการคัดกรองมะเร็งเพื่อตรวจหารอยโรคก่อนมะเร็ง และมะเร็งตั้งแต่ระยะเริ่มแรก เพื่อลดความเสี่ยงต่อการเกิดโรคมะเร็งและสามารถให้การรักษาให้หายขาดได้ มีหลักฐานการศึกษาที่ชัดเจนว่าการตรวจคัดกรองสามารถลดอัตราการเสียชีวิตจากโรคมะเร็งที่สำคัญได้โดยเฉพาะมะเร็งปากมดลูก มะเร็งเต้านม มะเร็งลำไส้ใหญ่ และมะเร็งปอด ซึ่งเป็นมะเร็งที่พบได้บ่อยและมีความสำคัญในประเทศไทย อุปสรรคสำคัญของการตรวจคัดกรองโรคมะเร็งในประเทศไทยมีหลายประการ เช่น การเข้าถึงหน่วยงานบริการ ความรู้ความเข้าใจและความยอมรับในการตรวจคัดกรอง ความต่อเนื่องในการตรวจคัดกรอง เทคนิคการตรวจที่เหมาะสม ค่าใช้จ่ายราคาสูง การขาดแคลนบุคลากรที่ชำนาญใน การตรวจคัดกรอง ปัญหาเหล่านี้จะแก้ไขได้ต้องได้รับความร่วมมือจากหลายภาคส่วนตั้งแต่ภาคประชาชน ภาควิชาการ ภาครัฐ จึงจะสัมฤทธิ์ผล

ปัจจุบัน เทคโนโลยีสารสนเทศด้านข้อมูลมีการพัฒนา

ไปอย่างมาก ข้อมูลแต่ละประเภทมีเอกลักษณะเฉพาะตัวที่ไม่เหมือนกัน เช่น ข้อมูลทางการแพทย์ ข้อมูลทางกฎหมาย หรือข้อมูลทางธุรกิจ เป็นต้น เทคนิคทางด้านข้อมูลจึงจำเป็นต้องมีการวิจัยหาอัลกอริทึมที่เหมาะสมเฉพาะสำหรับข้อมูลแต่ละประเภท การนำอัลกอริทึมของงานวิจัยก่อนหน้ามาปรับใช้กับข้อมูลชุดใหม่ ๆ อาจไม่สามารถทำได้ทันเวลา จากข้อจำกัดดังกล่าว ผู้วิจัยจึงมีแนวคิดและความสนใจนำวิธีการทำเหมืองข้อมูลด้วย Ant-Miner Algorithm มาวิจัยกับข้อมูลด้านทางการแพทย์ ด้วยวิธีการปรับปรุงขั้นตอนและเทคนิคในการคัดเลือกคุณลักษณะข้อมูล (Feature Selection) อันเป็นการเพิ่มประสิทธิภาพและความสามารถในการสร้างแบบจำลองการวิเคราะห์ข้อมูลในการวินิจฉัยโรคที่เหมาะสมและความถูกต้องมากยิ่งขึ้น เพื่อวินิจฉัยผลการคัดกรองมะเร็งปากมดลูกจากฐานข้อมูลสากลและการจำแนกผลการคัดกรองมะเร็งปากมดลูกให้มีประสิทธิภาพเพิ่มสูงขึ้น

วัตถุประสงค์

1. ประยุกต์ใช้เทคนิคเหมืองข้อมูลสำหรับชุดข้อมูลทางการแพทย์แบบหลายคำตอบ (Multi Label)

2. เพื่อศึกษาวิธีปรับปรุงกระบวนการทำเหมืองข้อมูลด้วยการคัดเลือกคุณลักษณะ เพื่อเพิ่มประสิทธิภาพความถูกต้องและความแม่นยำในการวินิจฉัยคัดกรองมะเร็งปากมดลูกเบื้องต้น (Cervical Cancer Screening)

วิธีดำเนินการวิจัย

งานวิจัยนี้เป็นการวิจัยเชิงทดลอง (Experimental Research) ใช้ฐานข้อมูลทางการแพทย์สากลแบบเปิด ด้านปัจจัยเสี่ยงมะเร็งปากมดลูกของมหาวิทยาลัยมลรัฐแคลิฟอร์เนียเมืองเออร์ไวน์[4] แบบหลายคำตอบ (Multi Label) กลุ่มตัวอย่างจำนวน 858 คน มีคุณลักษณะ (Attribute) สำคัญ 32 ค่า และตัวแปรสำหรับคำตอบหรือผลการวินิจฉัย (Label/Target) จำนวน 4 ผลวินิจฉัยหรือคำตอบของแบบจำลอง คือ 1) Hinselmann, 2) Schiller, 3) Cytology และ 4) Biopsy[5] ตามตารางที่ 1 เป็นคุณลักษณะข้อมูลต้นแบบสำหรับการวิจัยนี้ เพื่อสร้างแบบจำลอง (Model) การคัดกรองมะเร็งปากมดลูกด้วย Ant-Miner Algorithm

การจัดเตรียมข้อมูล (Data Preparation) ประกอบด้วย Data Cleaning, Data Transformation และ Data Reduction เพื่อใช้ในการฝึกสอนระบบเหมืองข้อมูล (Data Training) และสำหรับการทดสอบ (Data Testing) แบบจำลองที่สร้างขึ้นจากวิธีระบบเหมืองข้อมูล ด้วยอัลกอริทึมแอนท์-ไมเนอร์ (Ant-Miner Algorithm)

ตารางที่ 1 ข้อมูลคุณลักษณะ[6]

ลำดับ	ชื่อคุณลักษณะ (Attribute Name)	ลำดับ	ชื่อคุณลักษณะ (Attribute Name)
1	Age	21	STDs: Molluscum Contagiosum
2	Number of Sexual Partners	22	STDs: AIDs
3	Age at 1st Sexual Coitus (Age)	23	STDs: HIV
4	Number of Parturition	24	STDs: Hepatitis B
5	Smokes	25	STDs: HPV
6	Smokes (Year)	26	STDs: Number of Diagnosis
7	Smokes (Packs/Year)	23	STDs: HIV
8	Hormonal Contraception	24	STDs: Hepatitis B
9	Hormonal Contraception (Years)	25	STDs: HPV
10	Intrauterine Devices (IUDs)	26	STDs: Number of Diagnosis
11	IUDS (Years)	27	STDs: Time Since First Diagnosis
12	Sexually Transmitted Diseases (STDs)	28	STDs: Time Since Last Diagnosis
13	STDs (Number)	29	Dx: Cancer
14	STDs: Condylomatosis	30	Dx: CIN
15	STDs: Cervical Condylomatosis	31	Dx: HPV
16	STDs: Vaginal Condylomatosis	32	Dx
17	STDs: Vulvo-Perineal Condylomatosis	33*	Hinselmann
18	STDs: Syphilis	34*	Schiller
19	STDs: Pelvic Inflammatory Disease	35*	Cytology
20	STDs: Genital Herpes	36*	Biopsy

การคัดเลือกคุณสมบัติ (Feature Selection) หรือรูปแบบ (Format) ของข้อมูลด้วยการความสัมพันธ์คุณลักษณะแบบเหมาะสมใช้วิธี Correlation-based Feature Subset Selection (CFS)[7] เป็นการเลือกคุณสมบัติแบบวิธีฟิลเตอร์ (Filter) โดยทำการจัดเรียงลำดับของคุณลักษณะ (Feature Ranking Technique) ด้วยการคัดเลือกคุณลักษณะที่เหมาะสมสัมพันธ์ ข้อดีของวิธีเลือกคุณลักษณะแบบฟิลเตอร์นี้ คือ เทคนิคที่คำนวณได้ง่าย รวดเร็ว และหลีกเลี่ยงการเกิดโอเวอร์ฟิตติ้ง (Overfitting) ซึ่งเป็นความสัมพันธ์ระหว่างคุณลักษณะแต่ละค่ากับผลลัพธ์หรือคำตอบเท่านั้น ไม่ได้คำนึงถึงความสัมพันธ์ระหว่างคุณลักษณะกันเอง ตามสมการที่ (1)

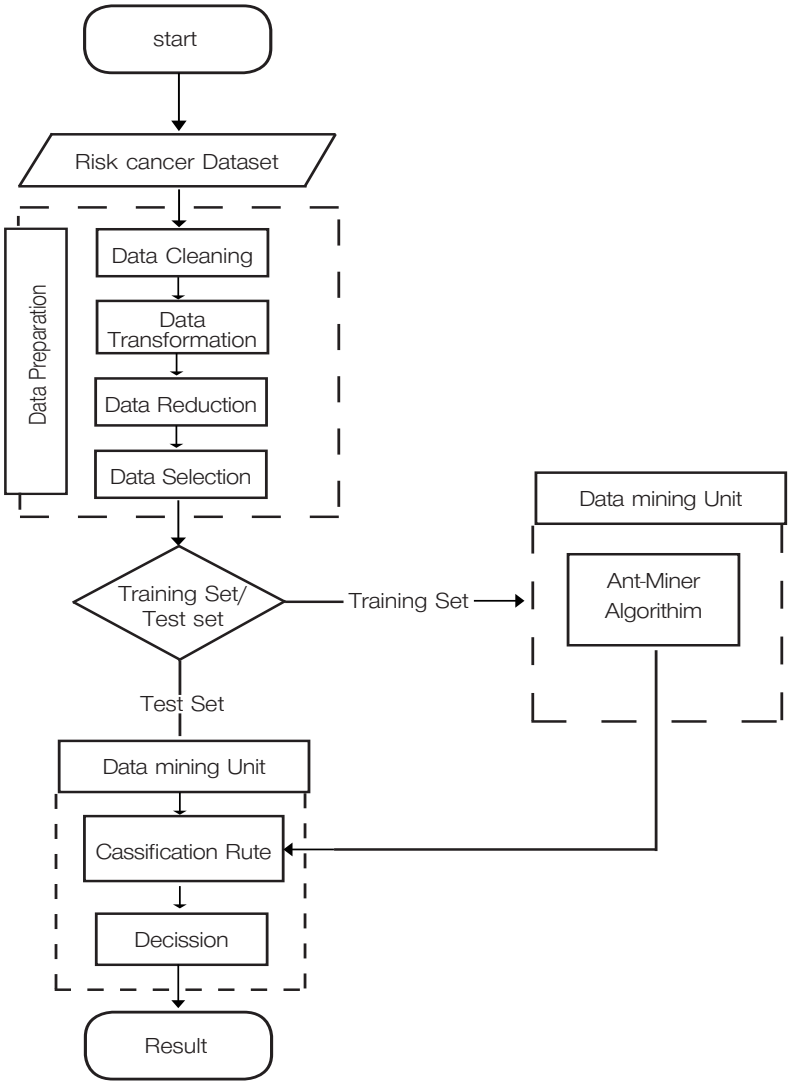
$$Merit_s = \frac{kr_{kc}}{\sqrt{k+(k-1)r_{kk}}} \quad \text{.....(1)}$$

เมื่อ $Merit_s$ คือ เซตของคุณลักษณะที่สัมพันธ์

r_{kc} คือ ค่าสัมประสิทธิ์เฉลี่ยความสัมพันธ์ของคุณลักษณะกับผลลัพธ์

r_{kk} คือ ค่าสัมประสิทธิ์เฉลี่ยความแตกต่างของคุณลักษณะกับผลลัพธ์

การสร้างแบบจำลองการคัดกรองมะเร็งปากมดลูกด้วยวิธีการทำเหมืองข้อมูล (Data Mining Technique) ด้วยแอนท์-ไมเนอร์อัลกอริทึม (Ant-Miner Algorithm) แบบหลายคำตอบ (Multi Label) การวิจัยครั้งนี้ ผู้วิจัยมีกรอบแนวคิดและขั้นตอนการดำเนินการวิจัย การสร้างแบบจำลองการคัดกรองมะเร็งปากมดลูกด้วยแอนท์ไมเนอร์อัลกอริทึม ดังภาพที่ 1



ภาพที่ 1 กรอบแนวคิดและขั้นตอนการดำเนินการวิจัย

จากภาพที่ 1 อธิบายรายละเอียดขั้นตอนการทำงานของการดำเนินการวิจัย ดังนี้

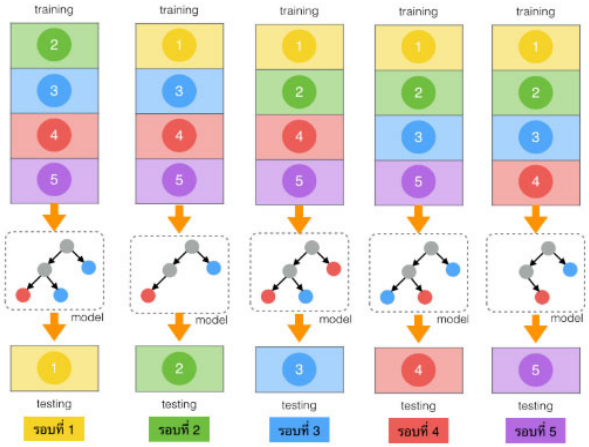
1. การจัดเตรียมข้อมูลเบื้องต้น เพื่อทำคัดแยกข้อมูลที่ไม่สมบูรณ์ ออกจากข้อมูล ทำการแปลงค่าข้อมูล ลดขนาดข้อมูล และการคัดเลือกคุณลักษณะ (CFS) เพื่อกำหนดรูปแบบคุณลักษณะ (Format) ของข้อมูล จัดแบ่งข้อมูลเป็น 2 ส่วน ชุดข้อมูลสำหรับฝึกสอนระบบ (Training Set) และชุดข้อมูลสำหรับทดสอบระบบ (Test Set)

2. การสร้างแบบจำลองการวินิจโรคมะเร็ง ในการวิจัยครั้งนี้ ผู้วิจัยใช้วิธีการสร้างแบบจำลองด้วยวิธี Ant-Miner Algorithm โดย Parpinelli, Lopes และFreitas[8] ได้พัฒนาอัลกอริทึมนี้เพื่อค้นหาคุณลักษณะเด่นหรือคุณลักษณะที่สำคัญทั้งหมดในการคัดแยกสิ่งต่าง ๆ ตามกฎ ในกรณีการฝึกสอนอัลกอริทึมนี้จะสร้างเซตของกฎ IF- THEN ในรูปของ IF < term1 ^ term2 ^ ... > and THEN < Class > การสร้างกฎของการคัดแยกจากข้อมูลด้วยคุณลักษณะเด่นหรือคุณลักษณะสำคัญของสิ่งต่าง ๆ แสดงดังภาพที่ 2

```
Training set = all training cases;
WHILE (No. of uncovered cases in the Training set >
max_uncovered_cases)
  i=0;
  REPEAT
    i=i+1;
    Anti incrementally constructs a classification rule;
    Prune the just constructed rule;
    Update the pheromone of the trail followed by Anti;
  UNTIL (i ≥ No_of_Ants) or (Anti constructed the
  same rule as the previous No_Rules_Converg-1 Ants)
    Select the best rule among all constructed rules;
    Remove the cases correctly covered by the selected
    rule from the training set;
END WHILE
```

ภาพที่ 2 Psudocode การทำเหมืองข้อมูลด้วย Ant-Miner Algorithm[8]

3. การตรวจสอบแบบไขว้ (Cross Validation) เป็นวิธีการฝึกสอนแบบจำลองเพื่อเพิ่มประสิทธิภาพในการทำนายผลลัพธ์ด้วยการคาดการณ์ค่าความผิดพลาดของแบบจำลอง (Model) ในส่วนการเรียนรู้ของวิธีเหมืองข้อมูลจากชุดข้อมูลการฝึกสอนแบบจำลอง พื้นฐานวิธี Cross Validation[9] เป็นการสุ่มตัวอย่าง (Re-sampling) เริ่มจากแบ่งชุดข้อมูลออกเป็นส่วน ๆ และนำบางส่วนของกลุ่มตัวอย่างชุดข้อมูลที่แบ่งได้มาใช้ตรวจสอบ คำตอบของการ Cross Validation จะถูกใช้เป็นตัวเลือกในการกำหนดรูปแบบจำลอง เช่น สถาปัตยกรรมเครือข่ายการสื่อสาร (Network Architecture) หรือ แบบจำลองการคัดแยกประเภท (Classification Model) โดยใช้เทคนิคด้าน Neural Network หรือ Decision Tree ของการทำเหมืองข้อมูล ซึ่งจะต้องแบ่งข้อมูลออกเป็นสองชุดคือ ชุดฝึกสอน (Training Set) และชุดตรวจสอบ (Validation Set) การเลือกชุดข้อมูลที่ง่าย และไม่เหมาะสมมาใช้เป็นชุดฝึกสอน หรือชุดทดสอบ ก็จะทำให้ได้แบบจำลองที่นำไปใช้ประโยชน์ได้ไม่ดี จึงนำวิธีการ k-Fold Cross Validation มาใช้แก้ปัญหา ในการฝึกสอนและทดสอบแบบจำลอง โดยมีการสลับสับเปลี่ยนข้อมูลไปมาแบ่งข้อมูลออกเป็นกลุ่ม k กลุ่ม (ในที่นี้ k แทนจำนวนกลุ่ม) ทำการทดลองฝึก k ครั้ง โดยครั้งแรกให้ข้อมูลกลุ่มแรกเป็นข้อมูลตรวจสอบ ที่เหลือเป็นข้อมูลฝึกฝน ครั้งต่อมาให้กลุ่มที่ 2 เป็นข้อมูลตรวจสอบ แล้วก็ไล่ไปเรื่อย ๆ จนทุกกลุ่มถูกใช้เป็นข้อมูลทดสอบทั้งหมด ดังภาพที่ 3 สุดท้ายก็นำข้อมูลผลลัพธ์ที่ได้จากการฝึก k ครั้งหาค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน เป็นข้อมูลในการประเมินประสิทธิภาพความถูกต้องและแม่นยำของแบบจำลองด้วยวิธี Confusion Matrix ดังตารางที่ 2 เป็นเครื่องมือสำคัญในการประเมินผลลัพธ์ของแบบจำลองการวินิจฉัยโรค (Decision) ที่มาสร้างจากเหมืองข้อมูล โดยประสิทธิภาพของผลการประเมินเป็นค่าความถูกต้อง (Accuracy) ตามสมการที่ (2) และค่าความแม่นยำ (Precession) ตามสมการที่ (3) ของผลการวินิจฉัยจากข้อมูลทดสอบ (Test Data)



ภาพที่ 3 ตัวอย่าง 5-fold Cross Validation[10]

ตารางที่ 2 Confusion Matrix

	Actually Positive	**Actually Negative**
Predicted Positive	True Positive (TPs)	False Positive (FPs)
Predicted Negative	False Negative (FNs)	True Negative (TNs)

ความถูกต้อง (Accuracy)

$$Accuracy = \frac{TPs+TNs}{TPs+TNs+FPs+FNs} \quad \text{.....(2)}$$

ความแม่นยำ (Precession)

$$Precession = \frac{TPs}{TPs+FPs} \quad \text{.....(3)}$$

ผลการวิจัย

การวิจัยนี้ ใช้คอมพิวเตอร์ส่วนบุคคลที่ติดตั้งระบบปฏิบัติการ Windows 10 ระบบประมวลผลเป็น Intel ® Core TM i5 Duo CPU @ 2.66GHz และมีค่าพารามิเตอร์ในการสร้างแบบจำลองการคัดกรองมะเร็งปากมดลูกด้วยวิธี Ant-Miner Algorithm ดังนี้

- Cross Validation = 10
- Number of Ants = 100
- Min. Case per Rule = 5
- Max. Uncovered Cases = 10
- Rules for Convergence = 10
- Number of Iterations = 50

การประมวลผลข้อมูลเบื้องต้น พบว่า ข้อมูล (Raw Data) จำนวนทั้งหมด 858 ตัวอย่าง มีคุณลักษณะทั้งหมด 36 ค่า รวมค่าผลลัพธ์หรือคำตอบ เมื่อนำมาประมวลผลการจัดเตรียมข้อมูลเบื้องต้น ด้วยวิธีการทำความสะอาดข้อมูล (Data Cleaning) เพื่อตรวจสอบและการแก้ไข (หรือลบข้อมูล) ในค่าคุณลักษณะที่ผิดพลาด และค่าคุณลักษณะที่ไม่มีผลต่อการทำนายผลลัพธ์จากแบบจำลองการคัดกรองมะเร็งปากมดลูก ออกไปจากชุดข้อมูลนำเข้า ระบบสร้างแบบจำลอง ซึ่งข้อมูลนำเข้าสำหรับใช้ในขั้นตอนการประมวลผลลำดับถัดไป มีจำนวนคงเหลือทั้งหมด 668 ตัวอย่าง และมี 8 รายการคุณลักษณะจะถูกแปลงข้อมูล (Data Transformation) ด้วยวิธีการทำนอร์มอลไลซ์ (Normalization) คือ Min-Max Normalization โดยแปลงค่าข้อมูลให้อยู่ในช่วงสั้น ๆ ที่อัลกอริทึมการทำเหมือง ข้อมูลสามารถนำไปใช้ประมวลผลได้ ผลการทำนอร์มอลไลซ์ ค่าคุณลักษณะ ตามตารางที่ 3

ตารางที่ 3 ผลการทำนอร์มอลไลซ์ (Normalization)

Attribute Name	Before Normalization	After Normalization
Age	43	5
Number of Sex Partner	13	5
Age at 1st Sexual Coitus	22	5
Number of Parturition	12	5
Smokes Years	31	5
Smokes Packs Years	63	5
Hormonal Contraception (Year)	41	5
IUD Years	27	5

การคัดเลือกคุณลักษณะด้วยวิธี CFS หลังจากกระบวนประมวลผลข้อมูลเบื้องต้น พบว่า รายการคุณลักษณะที่ได้มีค่าความสัมพันธ์สูงในการทำนายคำตอบของแบบจำลองการคัดกรองมะเร็งปากมดลูกด้วยแอนท์ไมเนอร์อัลกอริทึม สามารถจัดเรียงลำดับรายการคุณลักษณะตามลำดับความสัมพันธ์จากมากไปน้อยตามตารางที่ 4

ตารางที่ 4 ผลการคัดเลือกคุณลักษณะด้วยวิธี CFS

วิธีการ	คุณลักษณะ (Attributes)
Hinselmann	2,3,4,11,9,1,23,26,31,13,6,29,7,8,5
Schiller	11,9,4,1,2,3,7,6,8,13,31,26,29,23,5
Cytology	9,11,4,2,1,3,8,7,32,13,23,31,6,5,29
Biopsy	3,4,9,11,2,1,31,7,6,29,8,32,13,23

จากตารางที่ 4 พบว่า ลำดับรายการคุณลักษณะของปัจจัยเสี่ยงการคัดกรองมะเร็งปากมดลูก 6 รายการแรก ประกอบไปด้วย อายุ (Age) จำนวนของคู่เพศสัมพันธ์ (Number of Sexual Partners) อายุที่มีเพศสัมพันธ์ครั้งแรก (Age at 1st Sexual Coitus [Age]) จำนวนของการตั้งครรภ์ (Number of Parturition) จำนวนปีการคุมกำเนิดด้วยฮอร์โมน (Hormonal Contraception [Years]) และ จำนวนปีการใส่ห่วงคุมกำเนิด (IUD [Years]) โดยลำดับของคุณลักษณะปัจจัยเสี่ยงมีอิทธิพลต่อความสัมพันธ์ของคุณลักษณะกับผลลัพธ์ของแบบจำลองการคัดกรองมะเร็งปากมดลูก โดยคุณลักษณะ (Attribute) ที่มีสัมพันธ์เหมาะสมกับแบบจำลองทั้งสิ้น 16 รายการ ในการสร้างแบบจำลองการคัดกรองมะเร็งปากมดลูกด้วยวิธี Ant-Miner Algorithm และสามารถจัดแบ่งข้อมูลได้ตามวิธีการต่าง ๆ ในการคัดกรองมะเร็งปากมดลูกตามตารางที่ 5

ตารางที่ 5 ผลลัพธ์การคัดกรองมะเร็งปากมดลูกด้วยวิธีการต่าง ๆ

Method	Benign	Malignant
Hinselmann	638	30
Schiller	625	63
Cytology	630	38
Biopsy	623	45

จากตารางที่ 5 ชุดข้อมูลที่ใช้ในการวิจัย พบว่า ผลการคัดกรองมะเร็งปากมดลูกเป็นเนื้องอก (Benign) มากที่สุดด้วยวิธี Hinselmann จำนวน 638 ราย รองลงมา คือ วิธี Cytology มีผลการตรวจเป็นเนื้องอก จำนวน 630 ราย อันดับที่สาม คือ วิธี Schiller มีผลการตรวจเป็นเนื้องอก จำนวน 625 ราย โดยสุดท้ายเป็นวิธี Biopsy มีผลการตรวจเป็นเนื้องอก 623 ราย ตามลำดับ และผลการตรวจคัดกรองมะเร็งปากมดลูกเป็นมะเร็ง (Malignant) มากเป็นอันดับแรก คือ วิธี Schiller จำนวน 63 ราย อันดับสอง คือ วิธี Biopsy มีผลการตรวจเป็นมะเร็ง 45 ราย อันดับที่สาม คือ วิธี Cytology ผลการตรวจเป็นมะเร็ง 38 ราย อันดับที่สี่ คือ วิธี Hinselmann ผลการตรวจเป็นมะเร็ง 30 ราย ตามลำดับ

ตารางที่ 6 แสดงผลความถูกต้อง (ร้อยละ) จำนวนของกฎและจำนวนของเทอมในแต่ละกฎ จำแนกตามวิธีต่าง ๆ จากการทดสอบแบบจำลองการคัดกรองมะเร็งปากมดลูกด้วยวิธีแอนท์ไมเนอร์อัลกอริทึม

วิธีการ	Accuracy (%) (Average ± SD)	Number of Rules (Average ± SD)	Number of Terms (Average ± SD)
Hinselmann	93.26±0.91	6.8±0.7	13.1±1.12
Schiller	90.86±1.21	6.4±0.3	14.7±1.02
Cytology	96.26±0.84	7.1±0.4	15.1±1.07
Biopsy	98.35±1.08	6.1±0.3	16.7±1.2

จากตารางที่ 6 พบว่า ความถูกต้องของการทดสอบแบบจำลองการคัดกรองมะเร็งปากมดลูกในรูปแบบกฎ IF < term1 ^ term2 ^ ... > and THEN < Class > ด้วยวิธีการ Biopsy จากแบบจำลองการวินิจจัยมะเร็งปากมดลูก Ant-Miner Algorithm มีความถูกต้องเฉลี่ยสูงสุด คือร้อยละ 98.35 จำนวนกฎเฉลี่ย 6.1 กฎ และเทอมเฉลี่ยต่อกฎเท่ากับ 16.7 เทอม อันดับรองลงมา คือ วิธีการ Cytology มีความถูกต้องเฉลี่ยร้อยละ 96.26 มีกฎเฉลี่ย 7.1 กฎ และเทอมเฉลี่ยต่อกฎท่ากับ 15.1 เทอม อันดับที่สาม คือ วิธี Hinselmann มีความถูกต้องเฉลี่ยร้อยละ 93.26 จำนวนกฎเฉลี่ยที่ใช้ 6.8 กฎ และจำนวนของเทอมเฉลี่ยต่อกฎเท่ากับ 13.1 เทอม และอันดับสุดท้าย คือ วิธี Schiller มีความถูกต้องเฉลี่ยร้อยละ 90.86 มีกฎเฉลี่ยจำนวน 6.4 กฎ จำนวนเทอมเฉลี่ยต่อกฎ คือ 14.7 เทอม ตามลำดับ

ตารางที่ 7 เปรียบเทียบผลความถูกต้อง (ร้อยละ) ของการทดสอบแบบจำลองการคัดกรองมะเร็งปากมดลูกระหว่างวิธีใช้คุณลักษณะทั้งหมด (Full Attributes) กับวิธีการคัดเลือกคุณลักษณะด้วยวิธี CFS

Method	Accuracy (%)	
	Full Attributes	CFS
Hinselmann	87.97	93.26
Schiller	84.12	90.86
Cytology	85.23	96.25
Biopsy	90.47	98.35

จากตารางที่ 7 พบว่า การทดสอบแบบจำลองสำหรับการคัดกรองมะเร็งปากมดลูกด้วยวิธี Ant-Miner Algorithm ที่ใช้การคัดเลือกคุณลักษณะที่เหมาะสมด้วยวิธี CFS มีความถูกต้องของผลลัพธ์แบบจำลองสูงกว่าวิธีที่ใช้คุณลักษณะทั้งหมดสร้างแบบจำลอง โดยการตรวจด้วยวิธี Biopsy ให้ผลความถูกต้องสูงสุดทั้งวิธีใช้คุณลักษณะทั้งหมด คือร้อยละ 90.46 และวิธีการคัดเลือกคุณลักษณะด้วย CFS คือร้อยละ 98.35 อันดับที่สอง คือ วิธี Cytology มีความถูกต้องร้อยละ 96.26 ด้วยวิธีการคัดเลือกคุณลักษณะ (CFS)และมีความถูกต้องร้อยละ 85.23 ด้วยการใช้คุณลักษณะทั้งหมด อันดับที่สาม คือ วิธีการ Hinselmann แบบจำลองการคัดกรองมะเร็งปากมดลูกที่สร้างจากการคัดเลือกคุณลักษณะด้วยวิธี CFS มีความถูกต้องร้อยละ 93.26 การใช้คุณลักษณะทั้งหมด (Full Attributes) มีความถูกต้องร้อยละ 87.97 วิธี Schiller มีความถูกต้องน้อยที่สุดร้อยละ 90.86 สำหรับวิธีการคัดเลือกลักษณะด้วย CFS และมีความถูกต้องร้อยละ 84.12 ในการใช้คุณลักษณะทั้งหมดตามลำดับ

การประเมินประสิทธิภาพของแบบจำลองการคัดกรองมะเร็งปากมดลูกด้วยวิธี Ant-Miner Algorithm จากข้อมูลชุดทดสอบ (Test Set) ในด้านความถูกต้อง (Accuracy) ตามสมการที่ (2) และด้านความแม่นยำ (Precession) ตามสมการที่ (3) ได้ผลลัพธ์ประสิทธิภาพตามตารางที่ 8

ตารางที่ 8 ค่าร้อยละประสิทธิภาพของแบบจำลองด้านความถูกต้อง (Accuracy) และด้านความแม่นยำ (Precession) จากข้อมูลชุดทดสอบ (Test Set)

Method	Accuracy (%)	Precession (%)
Hinselmann	93.26	90.00
Schiller	90.86	95.23
Cytology	96.25	92.10
Biopsy	98.35	97.78

จากตารางที่ 8 พบว่า แบบจำลองการตรวจคัดกรองมะเร็งปากมดลูกด้วยวิธี Ant-Miner Algorithm ในข้อมูลชุดทดสอบ (Test Set) วิธี Biopsy มีประสิทธิภาพสูงที่สุดเฉลี่ยร้อยละ 98.06 เป็นประสิทธิภาพด้านความถูกต้อง (Accuracy) และประสิทธิภาพด้านความแม่นยำ (Precession) คือร้อยละ 98.35 และร้อยละ 97.79 ตามลำดับ รองลงมา คือวิธี Cytology ประสิทธิภาพโดยเฉลี่ยทั้ง 2 ด้านร้อยละ 94.18 โดยด้านความถูกต้องร้อยละ 96.26 และความแม่นยำร้อยละ 92.10 อันดับที่สาม วิธี Schiller มีประสิทธิภาพโดยเฉลี่ยร้อยละ 93.05 ในด้านความถูกต้องร้ อยละ 90.86 ความแม่นยำร้อยละ 95.23 และสุดท้ายมีประสิทธิภาพโดยเฉลี่ย 91.63 คือ วิธี Hinselmann มีความถูกต้องร้อยละ 93.26 และความแม่นยำร้อยละ 90.00 ตามลำดับ

วิจารณ์

Ant-Miner Algorithm[8] ที่ใช้ในการวิจัยครั้งนี้ เป็นอันกอริทึมที่ผู้วิจัยพัฒนาเพิ่มเติมในส่วนการคัดเลือกตัวแปรคุณลักษณะ และพัฒนาขั้นตอนการคัดเลือกคุณลักษณะด้วยวิธี Correlation-based Feature Selection (CFS) ที่เหมาะสมกับชุดข้อมูลทางด้านการแพทย์ ทำให้การวินิจัยทางการแพทย์มีประสิทธิภาพทั้งด้านความถูกต้องและความแม่นยำเพิ่มมากขึ้น

แบบจำลองการตรวจคัดกรองมะเร็งปากมดลูกด้วยวิธี Ant-Miner Algorithm เป็นเครื่องมือหนึ่งที่มีประสิทธิภาพในการตรวจวินิจฉัยมะเร็งปากมดลูกเบื้องต้นโดยวิเคราะห์จากข้อมูลทางการแพทย์ประเภทปัจจัยเสี่ยงต่าง ๆ ที่มีผลต่อการเป็นมะเร็งปากมดลูก กระบวนการทำเหมืองข้อมูลด้วยการค้นหาความสัมพันธ์ที่ซ่อนอยู่ในข้อมูลด้วยการคัดเลือกคุณลักษณะวิธี CFS มีการตรวจสอบผลลัพธ์โดยคำนวณค่าความผิดพลาดของคำตอบด้วยวิธีการตรวจสอบไขว้ (Cross Validation) และตรวจสอบผลจากข้อมูลชุดทดสอบด้วยการประเมินประสิทธิภาพของผลลัพธ์ด้วยวิธี Confusion Matrix เป็นการเพิ่มประสิทธิภาพของแบบจำลองที่สร้างขึ้น ผลการประเมินประสิทธิภาพของแบบจำลองของงานวิจัยนี้

มีความถูกต้องเฉลี่ยรวมร้อยละ 94.68 และความแม่นยำเฉลี่ยรวมร้อยละ 93.78 สอดคล้องกับการศึกษาของ Al-Wesabi, Y.M.S. และคณะ[11] ที่พบว่า ประสิทธิภาพด้านความถูกต้องเฉลี่ยด้วยวิธีการต้นไม้ตัดสินใจ (Decision Tree) แบบพื้นฐาน (Basic Classification) เท่ากับร้อยละ 90.12 และเมื่อใช้วิธีการคัดเลือกคุณลักษณะ (Feature Selection) แบบ CFS เช่นเดียวกับผู้วิจัย ประสิทธิภาพด้านความถูกต้องเฉลี่ยของแบบจำลองด้วยวิธีการต้นไม้ตัดสินใจ (Decision Tree) เพิ่มขึ้นเป็นร้อยละ 95.00 เมื่อเปรียบเทียบการศึกษาของ Muhammed Fahri Unlersen และคณะ[12] ที่สร้างแบบจำลองการตรวจคัดกรองมะเร็งปากมดลูกด้วยวิธี Bayes Net กับแบบจำลองการตรวจคัดกรองมะเร็งปากมดลูกที่สร้างด้วย Ant-Miner Algorithm ของผู้วิจัย พบว่า มีประสิทธิภาพความถูกต้องสูงกว่า คือ ร้อยละ 97.26 และร้อยละ 98.35 ตามลำดับ

ข้อเสนอแนะ

การสร้างแบบจำลองเพื่อการคัดกรองมะเร็งปากมดลูกด้วยวิธี Ant-Miner Algorithm ด้วยการคัดเลือกคุณลักษณะ (Feature Selection) แบบ Correlation-based Feature Selection (CFS) สำหรับข้อมูลทางการแพทย์ในการวิจัยนี้เป็นวิธีการเพิ่มประสิทธิภาพของอัลกอริทึมที่ใช้ในการทำเหมืองข้อมูลด้วยวิธี Ant-Miner Algorithm แบบใหม่ โดยนำวิธีการคัดเลือกคุณลักษณะแบบ CFS มาประยุกต์ใช้ในการจัดเตรียมข้อมูล เป็นวิธีการที่สามารถทำงานได้อย่างเหมาะสมสอดคล้องกับกลุ่มข้อมูลที่ใช้ในการศึกษา สามารถคัดเลือกคุณลักษณะที่มีความสัมพันธ์กับกลุ่มข้อมูลในแต่ละคำตอบของวิธีการวินิจฉัย และตัดคุณลักษณะที่ไม่มีความสัมพันธ์ออกได้ ทำให้ชุดคุณลักษณะของกลุ่มข้อมูล มีขนาดและจำนวนลดลง และสามารถสร้างแบบจำลองการคัดกรองมะเร็งปากมดลูกได้เร็วขึ้น ดังนั้นวิธีการคัดเลือกคุณลักษณะในรูปแบบอื่น ๆ ที่แตกต่างจากงานวิจัยนี้ อาจทำให้มิติของคุณลักษณะมีความสัมพันธ์ข้อมูลกันมากขึ้น ขนาดลดลง และทำให้ประสิทธิภาพการทำเหมืองข้อมูลกับข้อมูลทางการแพทย์เพิ่มสูงยิ่งขึ้นต่อไป

เอกสารอ้างอิง

1. World Health Organization. Cervix uteri [Internet]. 2019 [cited 2020 February 10]. Available from: https://gco.iarc.fr/today/data/factsheets/cancers/23-Cervix-uteri-fact-sheet.pdf
2. โพสต์ทูเดย์. คณะแพทยศาสตร์ 4 สถาบันเสนอนโยบายสู้มะเร็งที่ถูกต้องลดการเสียชีวิต [อินเทอร์เน็ต]. 2562 [เข้าถึงเมื่อ 2 มีนาคม 2563]. เข้าถึงได้จาก: https://www.posttoday.com/pr/597295
3. กองบรรณาธิการ. มะเร็งปากมดลูก : สาเหตุ อาการ การวินิจฉัย การรักษา และ วัคซีนป้องกัน [อินเทอร์เน็ต]. 2561 [เข้าถึงเมื่อ 19 กุมภาพันธ์ 2563]. เข้าถึงได้จาก: https://www.honestdocs.co/cervical-cancer-symptoms-treatment-prevention
4. UCI Machine Learning Repository. Cervical cancer (risk factors) data set [Internet]. 2017 [cited 2020 March 3]. Available from: https://archive.ics.uci.edu/ml/datasets/ Cervical+cancer+%28Risk+Factors%29
5. Fernandes K, Cardoso SJ, Fenandes J. Transfer learning with partial observability applied to cervical cancer screening. IbPRIA 2017;243-50.
6. Akyol K. A study on test variable selection and balanced data for cervical cancer disease. Int J of Information Engineering and Electronic Business 2018;(10):1-7.
7. Hall MA. Correlaton-based feature selection for discrete and numeric class machine learning. In Proceeding of the 17th International Conference on Machine Learning 2000;359-66.
8. Parpinelli R, Lopes H, Freitas A. Data mining with an Ant Colony Optimization algorithm. Evaolutionary Computation, IEEE Transaction on 2002 Sep;(6):321-32.

9. Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. Wisconsin: University of Wisconsin–Madison; 2018.
10. เอกสิทธิ์ พัชรวงศ์ศักดา. การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้าไมน์นิง เบื้องต้น. กรุงเทพมหานคร: เอเชีย ดิจิตอล การพิมพ์; 2557.
11. Al-Wesabi YMS, Choudhury A, Won D. Classification of cervical cancer dataset. In Proceedings of the 2018 IISE Annual Conference 2018;1456-61.
12. Unlersen MF, Sabanci K, Ozcan M. Dertemining cervical cancer possibility by using machine learning methods. International Journal of Latest Research in Engineering and Technology 2017 December;03(12):65-71.