

BAB II

LANDASAN TEORI

Bab ini membahas mengenai teori - teori dari topik penelitian, metode pengolahan data yang akan dipakai dalam penelitian, dan daftar penelitian terdahulu dengan hasil kesimpulannya.

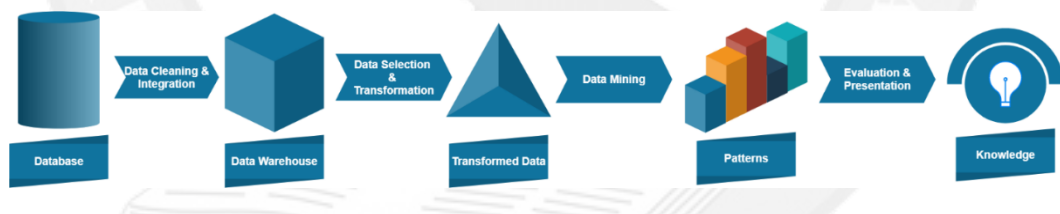
2.1 Data Mining

Data mining (DM) adalah sebuah proses menemukan pengetahuan dan pola yang terbentuk dari sekumpulan data [5]. Data tersebut dapat berupa *web*, teks, *database*, atau bentuk – bentuk lain. *Data mining* menggabungkan beberapa disiplin ilmu seperti Statistika, *Machine Learning*, *Artificial Intelligence*, dan lain - lain. Menurut Han, Kamber, dan Pei [5], proses untuk melakukan *data mining* terdiri dari beberapa tahapan yang bersifat iteratif, antara lain *data cleaning*, *data integration*, *data selection*, *data transformation*, *data mining*, *pattern evaluation*, dan *knowledge presentation*. Gambar 2.1 menjelaskan proses dalam melakukan *data mining* sehingga menghasilkan pengetahuan. Penjelasan untuk setiap tahapan adalah sebagai berikut:

1. *Data cleaning*, yaitu tahapan untuk menghilangkan data *noise* dan data tidak konsisten.
2. *Data integration*, yaitu tahapan menggabungkan data dari berbagai sumber data.
3. *Data selection*, yaitu tahapan memilih data – data yang relevan dari *database* untuk proses analisis.
4. *Data transformation*, yaitu tahapan mengubah data menjadi bentuk yang tepat untuk dilakukannya analisis.
5. *Data mining*, yaitu tahapan menggunakan metode – metode yang dipilih untuk menghasilkan pola dari suatu data.

6. *Pattern evaluation*, yaitu tahapan untuk mengidentifikasi dan mengevaluasi pola yang ditemukan berdasarkan tingkat kesesuaian dengan pengetahuan yang diinginkan.
7. *Knowledge presentation*, yaitu tahapan menggunakan visualisasi dan teknik – teknik untuk menampilkan pengetahuan yang dihasilkan kepada pengguna.

Dalam melakukan *data mining*, seseorang berusaha untuk mencari pola – pola yang terbentuk dalam sekumpulan data. Beberapa cara untuk mendapatkan pola tersebut antara lain adalah karakterisasi dan pembedaan (*characterization & discrimination*), melihat pola yang sering memiliki hubungan dan korelasi (*frequent pattern, association, and correlation*), klasifikasi dan regresi (*classification and regression*), analisis data – data berkelompok (*clustering analysis*), dan analisis data – data *outlier* (*outlier analysis*) [5]. Setiap cara yang dipakai memiliki tujuan yang berbeda – beda. Secara umum, tujuan tersebut terbagi menjadi dua yaitu deskriptif dan prediktif. Deskriptif berarti berusaha untuk menggambarkan sifat atau natur yang dimiliki oleh data dalam sebuah data set, sedangkan prediktif berarti berusaha menyimpulkan data yang ada untuk memprediksi sesuatu.



Gambar 2.1 Proses Data Mining
Sumber : Han, Kamber, dan Pei [5]

Penelitian ini menggunakan klasifikasi (*classification*) dalam mencari pola – pola yang ada dalam dataset yang akan diteliti. Teknik klasifikasi dipilih berdasarkan banyaknya penelitian terdahulu yang ditemukan memakai cara ini dalam melakukan penelitian sejenis.

2.2 Klasifikasi (Classification)

Klasifikasi (*classification*) merupakan salah satu teknik dari *data mining* yang menghasilkan model untuk menjelaskan dan memprediksi label data yang penting

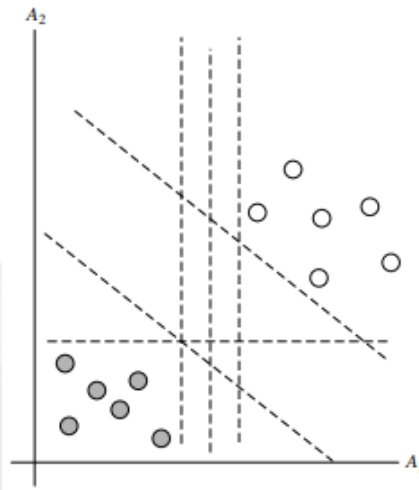
[5]. Klasifikasi terdiri dari dua proses, yaitu proses pembelajaran (*learning step*) dan proses klasifikasi (*classification step*). Proses pembelajaran merupakan proses dimana sebuah algoritma klasifikasi membentuk sebuah model dengan cara menganalisis *training data*. *Training data* yang dianalisis berisi data yang berasal dari *database* beserta label yang dimiliki oleh tiap baris data. Proses pembelajaran dari klasifikasi termasuk ke dalam kategori *supervised learning*, dimana setiap *training data* yang dianalisis memiliki label sehingga model yang dihasilkan dapat mempelajari *training data* berdasarkan label yang diberikan. Proses yang kedua adalah proses klasifikasi, yaitu proses yang dilakukan untuk mengukur keakuratan pengukuran yang dilakukan oleh model. Proses pengukuran tidak menggunakan lagi *training data*, karena estimasi keakuratan yang dihasilkan hampir dipastikan akan bernilai baik. Pengukuran harus menggunakan data lain diluar *training data* sehingga performa dari model dapat benar – benar terlihat. Proses ini menggunakan kumpulan data baru yang disebut *test data*. *Test data* memiliki label yang berguna untuk dicocokkan dengan label yang dihasilkan oleh model. Tingkat akurasi dari model akan dihitung berdasarkan ketepatan model dalam memprediksi label yang benar dari *test data*. Jika akurasi yang dihasilkan memuaskan, model dapat digunakan untuk mengukur data – data lain yang sebelumnya tidak memiliki label.

Penelitian ini akan membandingkan tiga algoritma klasifikasi. Algoritma yang dipilih merupakan algoritma yang sudah banyak dipakai oleh penelitian lain dalam melakukan penelitian sejenis. Tiga algoritma tersebut adalah *Support Vector Machine*, *Logistic Regression*, dan *Naïve Bayes*.

2.2.1 Support Vector Machine

Support Vector Machine (SVM) merupakan salah satu algoritma untuk mengklasifikasi data linear dan non linear [5]. Menurut Cortes dan Vapnik, dan Maglogiannis et al., yang juga dikutip oleh Malik et al., saat ini SVM banyak digunakan dalam aplikasi klinis [15] [16] [7]. SVM pertama kali diperkenalkan oleh Boser, Guyon, dan Vapnik [17]. SVM menggunakan pemetaan secara non linear untuk mengubah dimensi data. Dalam dimensi data yang baru, SVM mencari *hyperplane*, yaitu pembatas yang membedakan antara satu kelas dengan kelas lain.

SVM menggunakan konsep *maximum marginal hyperplane* (MMH) untuk mencari *hyperplane* yang paling baik [5] [15].



Gambar 2.2 Proses Menentukan *Hyperplane*
Sumber : Han, Kamber, dan Pei [5]

Konsep MMH membutuhkan *support vector*, yaitu data dari masing - masing kelas yang paling dekat dengan *hyperplane* yang dibuat, dan *margin*, yaitu jarak antara kedua *support vector* yang ditentukan oleh *hyperplane*.

Gambar 2.2 menunjukkan *hyperplane* yang mungkin dipakai untuk membedakan satu kelas dengan kelas lain. Konsep MMH berusaha untuk mencari *hyperplane* dengan *margin* yang paling besar sehingga proses klasifikasi memiliki tingkat akurasi yang tinggi. Untuk menentukan MMH, rumus 2.1 digunakan pada *training data* untuk membuat sebuah *trained support vector machine*.

$$d(\mathbf{X}^T) = \sum_{i=1}^l y_i \alpha_i \mathbf{X}_i \mathbf{X}^T + b_0,$$

Rumus 2.1 Rumus *Support Vector Machine*
Sumber: Han, Kamber, dan Pei [5]

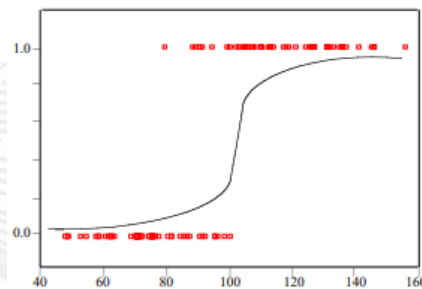
Rumus 2.1 menjelaskan cara SVM mengklasifikasi kelas dalam sebuah *data set* berdasarkan formula Lagrangian. Keterangan dari rumus 2.1 adalah sebagai berikut:

- $d(\mathbf{X}^T)$ = *Maximum Marginal Hyperplane*

- y_i = label kelas dari *support vector*
- α_i = Lagrangian *multiplier*
- X_i = *support vector*
- X^T = *test data* yang dicoba untuk diklasifikasikan
- b_0 = parameter yang telah ditentukan

2.2.2 Logistic Regression

Logistic Regression (LR) merupakan salah satu tipe dari *linear regression*. Secara umum LR cocok digunakan untuk menunjukkan adanya hubungan antara sebuah variabel hasil yang bertipe kategori dengan satu atau lebih variabel prediktor bertipe kategori atau bilangan. LR dapat memprediksi nilai variabel yang hanya memiliki dua nilai (dikotomis) dan variabel yang memiliki lebih dari dua nilai [18]. LR dirancang untuk menyelesaikan masalah yang muncul saat seseorang bekerja menggunakan *linear regression* dalam upaya menunjukkan hubungan antara variabel dimaksud. Gambar 2.3 menunjukkan kurva berbentuk S yang sulit dijelaskan oleh *linear regression*. Dalam hal ini, LR menyelesaikan masalah tersebut dengan mengaplikasikan logaritma natural pada variabel hasil.



Gambar 2.3 Contoh Kurva Hubungan Antara Variabel Hasil Bersifat Dikotomis dan Variabel Prediktor Bersifat Continuous
Sumber : Peng, Lee, dan Ingersoll [18]

Rumus 2.2 menjelaskan perhitungan probabilitas atau peluang munculnya hasil atau kejadian yang diinginkan menurut algoritma *logistic regression*.

$$\pi = \text{Probability}(Y = \text{outcome of interest} | X = x, \\ \text{a specific value of } x) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

Rumus 2.2 Rumus *Logistic Regression*
Sumber: Han, Kamber, dan Pei [5]

dimana:

- π = peluang
- β = koefisien regresi
- α = nilai *intercept* dari Y
- e = basis dari logaritma natural (2,71828)

2.2.3 Naïve Bayes

Naïve Bayes merupakan salah satu algoritma klasifikasi yang berasal dari metode statistik. *Naïve Bayes* dibuat berdasarkan teori statistik yang dibuat oleh Thomas Bayes, seorang pendeta dari Inggris yang meneliti probabilitas dan teori keputusan pada abad ke-18 [5]. *Naïve Bayes* memiliki asumsi *class-conditional independence*, yang beranggapan bahwa atribut dalam sebuah kelas tidak saling bergantung satu sama lain [19]. Melalui hasil klasifikasinya, *Naïve Bayes* dikenal memiliki akurasi dan kecepatan yang tinggi dalam mengolah data asli dalam jumlah besar [5] [7]. Rumus 2.3 menunjukkan rumus dari *Naïve Bayes* dalam menentukan probabilitas suatu data terhadap label kelas yang akan diuji.

$$P(C_i|X) = \frac{P(X|C_i) * P(C_i)}{P(X)}$$

Rumus 2.3 Rumus *Naïve Bayes*
Sumber: Han, Kamber, dan Pei [5]

Keterangan dari rumus 2.3 adalah sebagai berikut :

- $P(C_i|X)$ = peluang terjadinya C_i dengan syarat X

- $P(X|C_i)$ = peluang terjadinya X dengan syarat C_i
- $P(C_i)$ = peluang terjadinya C_i
- $P(X)$ = peluang terjadinya X
- X = data yang diteliti
- C_i = label kelas yang diteliti

2.3 Evaluasi Algoritma

Menurut Han, Kamber dan Pei, terdapat beberapa aspek yang dapat diukur dalam mengevaluasi performa dari sebuah algoritma [5]. Aspek tersebut antara lain akurasi, sensitivitas, *specificity*, presisi, F_1 score, dan F_β . Tabel 2.1 menunjukkan sebuah tabel yang berisi rumus untuk mengukur aspek – aspek tersebut. Melalui Tabel 2.1 terlihat beberapa terminologi yang digunakan dalam pengukuran evaluasi, seperti TP (*True Positive*), TN (*True Negative*), FP (*False Positive*), FN (*False Negative*), P (*Positive*), dan N (*Negative*). TP, TN, FP, dan FN dapat ditampilkan dalam bentuk tabel yang disebut *confusion matrix*. Gambar 2.4 menunjukkan *confusion matrix*. Dalam sebuah penelitian, terdapat *positive tuple* (*tuple* yang menjadi ketertarikan utama) dan *negative tuple* (*tuple* dengan hasil lain). TP mengacu pada berapa banyak *positive tuple* yang berlabel benar setelah data diproses oleh algoritma. TN merupakan jumlah *negative tuple* yang juga berlabel benar setelah data diproses. FP merupakan jumlah data *negative tuple* yang memiliki label positif setelah diproses. FN merupakan *positive tuple* yang salah dilabelkan menjadi negatif.

		Predicted class		
		yes	no	Total
Actual class	yes	TP	FN	P
	no	FP	TN	N
Total		P'	N'	P + N

Gambar 2.4 *Confusion Matrix*
Sumber : Han, Kamber, dan Pei [5]

Tabel 2.1 Pengukuran Evaluasi

Pengukuran	Formula
Akurasi, <i>recognition rate</i>	$\frac{TP + TN}{P + N}$
Error rate, <i>misclassification rate</i>	$\frac{FP + FN}{P + N}$
Sensitivitas, <i>true positive rate, recall</i>	$\frac{TP}{P}$
Specificity, <i>true negative rate</i>	$\frac{TN}{N}$
Presisi (<i>precision</i>)	$\frac{TP}{TP + FP}$
F , $F1$, F -score	$\frac{2 \times precision \times recall}{precision + recall}$
F_β , dimana β adalah bilangan asli non negatif	$\frac{(1 + \beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$

Sumber: Han, Kamber, dan Pei [5]

Pada penelitian ini, aspek yang akan digunakan akan untuk mengevaluasi performa dari algoritma adalah akurasi, sensitivitas, *error rate*, presisi, dan F_1 score. Setiap algoritma akan dievaluasi berdasarkan performanya dalam menghasilkan diagnosis dari tiap penyakit. Setiap aspek evaluasi akan ditampilkan dalam bentuk persentase.

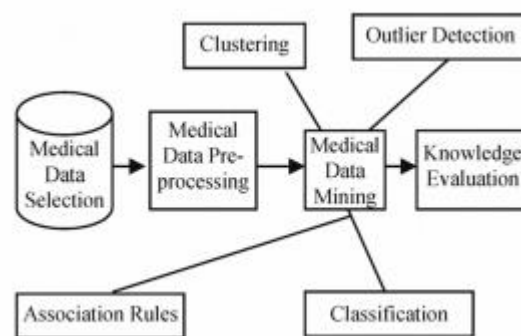
2.4 Medical Data Mining

Industri kesehatan merupakan salah satu industri yang terus menerus berkembang karena selalu memiliki data - data baru dalam bentuk rekam medis, data administrasi, dan tolak ukur penemuan lainnya [2] [20]. Penambahan data – data baru yang terus terjadi membuat industri kesehatan memiliki kesulitan untuk mengekstraksi data. Namun, sebagian besar dari data - data belum digunakan dengan baik [2]. Penerapan *data mining* memiliki potensi yang besar dalam membantu industri kesehatan. Sharda et al. menjelaskan bahwa dunia medis memiliki beberapa keuntungan dalam menerapkan *data mining*, yaitu penelitian terpublikasi yang terus meningkat, literatur dunia medis yang sudah terstandarisasi, dan penggunaan berbagai terminologi yang konstan dalam literatur [21]. *Data mining* juga sudah digunakan untuk mendiagnosis beberapa penyakit, seperti asma [22] [23], jantung [24] [10], diabetes [7] [25], dan penyakit – penyakit lain. *Data*

mining dapat memanfaatkan data yang ada untuk memprediksi berbagai penyakit dan membantu proses diagnosis, sehingga dokter dapat membuat keputusan yang lebih baik [2].

Zhao dan Wang merancang sebuah *framework* untuk membantu peneliti umum dalam mencari pola dalam bidang medis [26]. *Framework* dimaksud ditunjukkan dalam Gambar 2.5. Keempat proses yang didefinisikan oleh Zhao dan Wang dalam *framework* ini meliputi *medical data selection*, *medical data pre-processing*, *medical data mining*, dan *knowledge evaluation*. Penjelasan dari masing - masing proses diberikan sebagai berikut:

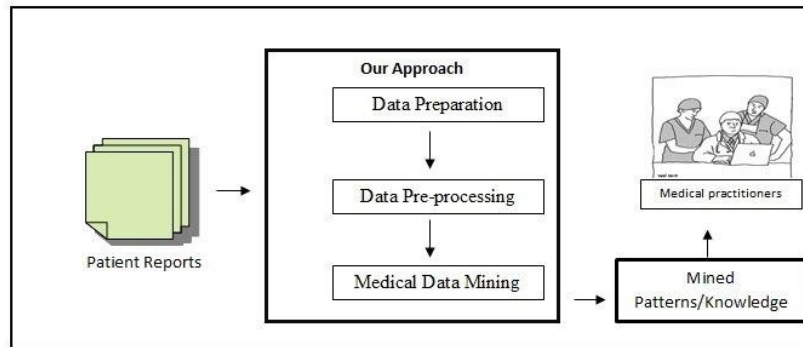
1. *Medical data selection*: dalam proses ini dilakukan pemilihan tipe data medis yang akan diteliti. Tipe data tersebut dapat berupa teks, gambar, hasil laboratorium, video, frekuensi sinyal, dan lain – lain.
2. *Medical data pre-processing*: proses ini berkaitan dengan pembersihan data dari *noise*, data yang tidak lengkap maupun data yang tidak konsisten.
3. *Medical data mining*: proses ini berkaitan dengan pengolahan data menggunakan teknik *data mining* untuk menghasilkan pengetahuan (*knowledge*).
4. *Knowledge evaluation*: proses ini mengevaluasi pengetahuan yang dihasilkan dari proses *medical data mining*.



Gambar 2.5 Framework Medical Data Mining
Sumber: Zhao dan Wang [26]

Choudhry et al. juga merancang sebuah *medical data mining framework* yang digunakan untuk mengembangkan *expert system* [27]. *Framework* yang dibuat

memiliki tiga proses, yaitu *data preparation*, *data pre-processing*, dan *medical data mining*. Gambar 2.6 menunjukkan *framework* yang dibuat.



Gambar 2.6 *Medical Data Mining Framework* Untuk Mengembangkan *Expert System*

Sumber: Choudhry et al. [27]

Itani et al. mencoba mencari tahu kebutuhan pengembangan *medical data mining* untuk membantu proses diagnosis [28]. Diagnosis adalah sebuah proses memberikan keterangan terhadap suatu kondisi berdasarkan set kategori yang telah ditetapkan sebelumnya [14]. Saat sebuah diagnosis dilakukan secara akurat, seorang pasien bisa mendapatkan penanganan yang tepat karena tindakan medis yang dilakukan sesuai dengan masalah yang benar – benar dialami oleh pasien [14] [27]. Dunia medis yang semakin kompleks, kemampuan kognitif yang terbatas, dan batasan ruang dan waktu membuat penerapan ilmu diagnosis menjadi semakin sulit [14]. Karena itu dunia medis butuh mengadopsi metode – metode analisis menggunakan teknologi dengan tetap mengutamakan akurasi dari proses analisis.

Model *data mining* yang akan dibuat dalam penelitian adalah *medical data mining* yang dapat menghasilkan diagnosis untuk membantu orang awam dalam membuat keputusan yang bersifat medis. Selain itu, model yang dihasilkan diharapkan juga dapat membantu praktisi dalam mendiagnosis penyakit. *Medical data mining* yang akan dibuat berfokus pada menggunakan teknik klasifikasi. Input yang dibutuhkan oleh model yang akan dibuat didapatkan melalui gejala – gejala yang tercatat dalam file dataset.

2.5 Klasifikasi Penyakit

Para pasien rumah sakit memiliki pengkodean jenis - jenis penyakit yang tercantum dalam rekam medis mereka. Dalam rekam medis tersebut, jenis - jenis penyakit dikodekan sesuai standar WHO, dimana kode tersebut dinamakan ICD (International Classification Of Diseases). Kode ini terus dikembangkan setiap waktu melihat perkembangan penyakit baru yang ada di dunia. Sampai saat ini, teknik pengkodean ICD yang dipakai adalah ICD-10. ICD-10 menjadi panduan seluruh penyelenggara pelayanan kesehatan (Rumah sakit, puskesmas, dll) dalam melakukan pengarsipan rekam medis pasien. Di dalam ICD-10 dilakukan pengelompokan dari penyakit menjadi 20 kelompok penyakit. Dalam ICD-10 pengelompokan ini didasarkan pada karakter dan jenis penyakit tersebut. Hal ini difungsikan untuk memudahkan dalam melakukan administrasi rekam medis sebab keterangan detail dari seorang pasien menjadi kesulitan tersendiri dalam melakukan pengarsipan, yang dapat berakibat pada hasil diagnosis yang tidak akurat [29].

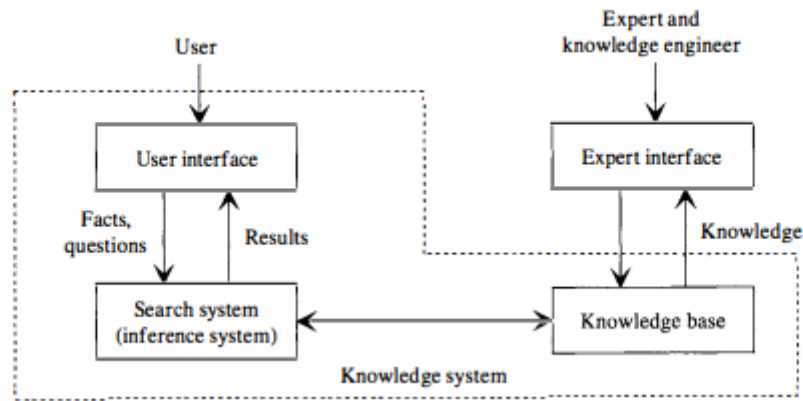
ICD-10 terdiri dari 3 volume, dimana klasifikasi penyakit terdapat pada volume 1 yang terbagi lagi menjadi 22 bab [30]. Menurut daftar tersebut, penyakit dibagi menjadi beberapa bagian, seperti penyakit pada mata, hidung, sistem pernapasan, sistem pencernaan, kulit, dan lain – lain.

Penelitian ini menggunakan dataset penyakit obesitas, jantung, tiroid, kanker payudara, dan COVID-19 sebagai objek penelitian. Penyakit yang dipilih untuk diteliti berdasarkan data yang tersedia. Performa dari tiap algoritma yang telah dipilih akan dibandingkan untuk melihat algoritma terbaik untuk mendiagnosis kategori penyakit berdasarkan data yang didapatkan.

2.6 Sistem Pakar (Expert System)

Sistem pakar (*expert system*) atau disebut juga *knowledge system* adalah sebuah sistem komputer yang mampu memberikan solusi layaknya seorang ahli manusia untuk memecahkan suatu masalah [31]. Layaknya seperti manusia yang membutuhkan pelatihan sebelum dapat dikatakan ahli dalam satu bidang, sebuah sistem pakar harus dilatih agar dapat menjadi ahli dalam memberikan sebuah solusi. Sistem pakar memiliki beberapa jenis, seperti sistem pakar dengan metode *forward*

chaining, backward chaining, case-based reasoning, model-based reasoning, dan lain – lain [32] [33].



Gambar 2.7 Diagram *Expert System*
Sumber: M. Stefik [32]

Dalam pembuatan sebuah sistem pakar, terdapat beberapa bagian, yaitu basis pengetahuan (*knowledge base*), *user interface*, *expert interface*, dan *inference system* [32]. Gambar 2.7 menjelaskan rangkaian dari sebuah sistem pakar. *User interface* berfungsi untuk menerima input dan mengubah input tersebut menjadi bentuk yang dapat diproses oleh *inference system*. *Inference system* mengolah input yang diberikan menggunakan algoritma yang dipakai dan menampilkan hasil dari data yang sudah diproses kepada pengguna. *Expert interface* adalah tampilan yang digunakan untuk memasukkan *knowledge* ke dalam sistem. *Knowledge base* menyimpan *knowledge* yang dimasukkan dalam bentuk *rule* yang dipakai oleh *inference system* dalam memberikan solusi.

Pada penelitian ini, hasil akhir yang diharapkan adalah sebuah sistem pakar yang mampu mendiagnosis penyakit. Metode klasifikasi *data mining* digunakan untuk membuat model yang menjadi *knowledge base* dari sistem pakar yang akan dibuat. Penelitian ini menggunakan dataset medis yang akan digunakan sebagai media pembelajaran atau *training* dari *knowledge base* sistem pakar. Beberapa penelitian terdahulu memakai metode ini untuk melatih sistem pakar yang akan dibuat.

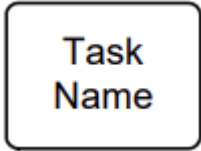
2.7 BPMN

BPMN atau *Business Process Modeling Notation* adalah sebuah notasi yang digunakan untuk menggambarkan proses dalam bisnis. BPMN dikembangkan dan distandarkan oleh *Object Management Group* (OMG). Dikutip dari dokumen *Business Process Model and Notation (BPMN)* tahun 2011, tujuan utama dari BPMN adalah “*to provide a notation that is readily understandable by all business users, from the business analysts that create the initial drafts of the processes, to the technical developers responsible for implementing the technology that will perform those processes, and finally, to the business people who will manage and monitor those processes*” [33, p. 1]. BPMN merupakan representasi abstrak dari proses sebenarnya di dunia nyata, dan merupakan suatu jawaban atas kebutuhan akan sebuah “bahasa” yang dapat melambangkan poin – poin penting dari sebuah proses bisnis. Adanya BPMN diharapkan dapat mampu mempermudah pemodelan proses bisnis, dan disaat yang sama mampu mengatasi kompleksitas dari proses bisnis sendiri.

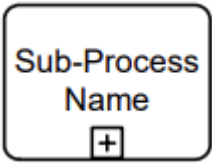
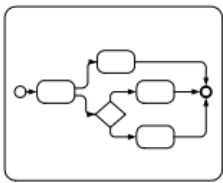



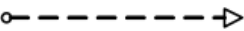

BPMN sudah banyak digunakan oleh berbagai perusahaan dan diajarkan dalam sistem pendidikan untuk menjelaskan proses bisnis. BPMN memiliki elemen desain yang cukup sederhana sehingga mudah dipahami oleh orang awam. Dalam BPMN terdapat *pool* dan *message flow* yang dapat menggambarkan proses pengiriman pesan antar pihak yang terkait dalam proses bisnis.

Pada penelitian ini BPMN digunakan sebagai alat untuk menggambarkan proses dari penelitian dan juga program sistem pakar. Tabel 2.2 menunjukkan elemen – elemen desain yang dipakai pada alur diagram penelitian dan alur sistem pakar.







Tabel 2.2 Elemen Desain BPMN yang Digunakan

Nama Elemen	Keterangan	Notasi
<i>Task</i>	Notasi yang menandakan sesuatu yang dikerjakan di dalam proses	

Tabel 2.2 Elemen Desain BPMN yang Digunakan (lanjutan)

Nama Elemen	Keterangan	Notasi
<i>Collapsed Sub-Process</i>	Notasi yang menandakan sebuah sub-proses dan memiliki tingkatan detail yang lebih dalam	
<i>Expanded Sub-Process</i>	Notasi yang menunjukkan detail dari sebuah sub-proses	
<i>Start Event</i>	Notasi yang menandakan dimulainya suatu proses	
<i>End Event</i>	Notasi yang menandakan berakhirnya suatu proses	
<i>Sequence Flow</i>	Notasi yang digunakan untuk menunjukkan urutan <i>activity</i> yang akan dilakukan	
Message Flow	Notasi yang digunakan untuk menunjukkan pengirim dan penerima pesan dalam proses	
Parallel Gateway	Notasi yang digunakan untuk menunjukkan <i>flow</i> yang dilakukan secara bersamaan dan menggabungkan 2 <i>flow</i> yang dilakukan bersamaan menjadi 1 <i>flow</i>	

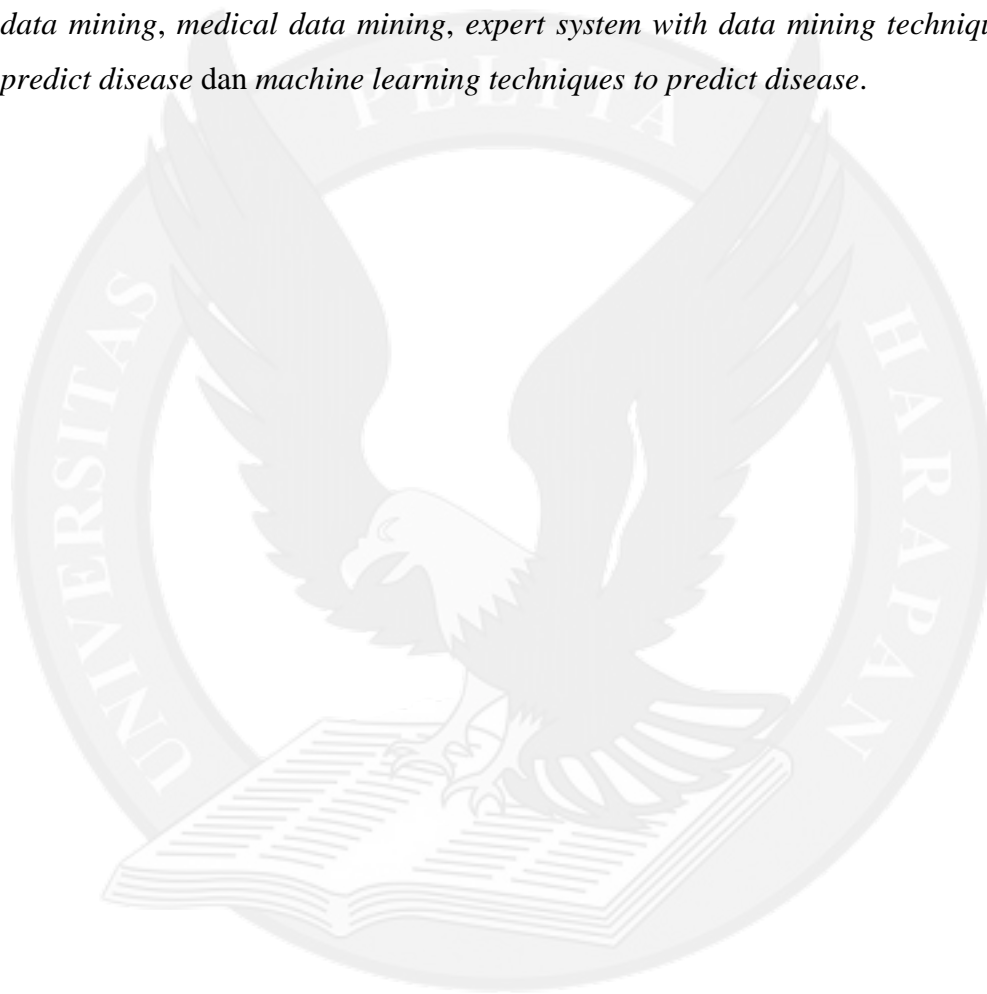
Tabel 2.2 Elemen Desain BPMN yang Digunakan (lanjutan)

Nama Elemen	Keterangan	Notasi
Exclusive Gateway	Notasi yang digunakan untuk membuat pilihan <i>flow</i> dalam proses. Hanya satu <i>flow</i> yang boleh diambil.	
Data Object	Notasi yang menunjukkan informasi atau data apa yang dibutuhkan sebuah <i>task</i> untuk dapat berjalan	
Message Throwing Event	Notasi yang digunakan untuk menunjukkan adanya pengiriman pesan	
Message Catching Event	Notasi yang digunakan untuk menunjukkan adanya penerimaan pesan	
Timer	Notasi yang digunakan untuk menunjukkan penundaan atau <i>delay</i> berdasarkan satuan waktu tertentu	
Pool	Notasi yang digunakan untuk menunjukkan entitas atau sesuatu yang melakukan proses	

2.8 Penelitian Terdahulu

Tabel 2.3 merupakan tabel perbandingan penelitian terdahulu yang telah dirangkum dan dinilai memiliki kesamaan dengan topik yang diambil dalam penelitian ini. Penelitian – penelitian tersebut adalah *Deep learning facilitates the diagnosis of adult asthma* oleh K. Tomita, R. Nagao, H. Touge, T. Ikeuchi, H. Sano, A. Yamasaki, dan Y. Tohda pada tahun 2017, *Different medical data mining approaches based prediction of ischemic stroke* oleh A. K. Arslan, C. Colak, dan M. E. Sarihan pada tahun 2016, *Comparing performances of logistic regression, decision trees , and neural networks for classifying heart disease patients* oleh A. Khemphila dan V. Boonjing pada tahun 2010, *A data mining approach for diagnosis of coronary artery disease* oleh R. Alizadehsani, J. Habibi, M. J.

Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, dan Z. A. Sani pada tahun 2013, dan *Developing a Prototype Knowledge-Based System for Diagnosis and Treatment of Diabetes Using Data Mining Techniques* oleh K.Eyasu, W. Jimma, dan T. Tadesse pada tahun 2020. Penelitian – penelitian ini diperoleh melalui sumber jurnal ilmiah yaitu *Science Direct* dan *National Center for Biotechnology Information* (NCBI) dengan menggunakan kata kunci seperti *data mining, medical data mining, expert system with data mining techniques to predict disease* dan *machine learning techniques to predict disease*.



Tabel 2.3 Penelitian Terdahulu Bagian I

Atribut	Jurnal 1	Jurnal 2	Jurnal 3	Jurnal 4	Jurnal 5
Penulis	K. Tomita, R. Nagao, H. Touge, T. Ikeuchi, H. Sano, A. Yamasaki, dan Y. Tohda	A. K. Arslan, C. Colak, dan M. E. Sarihan	A. Khemphila dan V. Boonjing	R. Alizadehsani, J. Habibi, M. J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, dan Z. A. Sani	Kedir Eyasu, Worku Jimma, dan Takele Tadesse
Judul	<i>Deep learning facilitates the diagnosis of adult asthma</i> [22]	<i>Different medical data mining approaches based prediction of ischemic stroke</i> [34]	<i>Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients</i> [24]	<i>A data mining approach for diagnosis of coronary artery disease</i> [10]	<i>Developing a Prototype Knowledge-Based System for Diagnosis and Treatment of Diabetes Using Data Mining Techniques</i> [35]
Teknik DM	<i>Classification</i>	<i>Classification</i>	<i>Classification</i>	<i>Classification</i>	<i>Classification</i>
Obyek Penelitian	Mendiagnosis penyakit asma	Memprediksi penyakit <i>ischemic stroke</i>	Memprediksi penyakit jantung pada pasien	Mendiagnosis penyakit jantung	Mengembangkan <i>prototype expert system</i> untuk mendiagnosis penyakit diabetes menggunakan metode data mining
Tipe Penyakit	Asma	Stroke	Jantung	Jantung Koroner	Diabetes

Tabel 2.3 Penelitian Terdahulu Bagian I (lanjutan)

Atribut	Jurnal 1	Jurnal 2	Jurnal 3	Jurnal 4	Jurnal 5
Perbandingan	Penelitian ini menggunakan teknik <i>Classification</i> dengan algoritma <i>Logistic Regression Analysis</i> , <i>Support Vector Machine</i> (SVM), dan <i>Deep Neural Network</i> (DNN) untuk mendiagnosis penyakit asma pada pasien dewasa	Penelitian ini menggunakan teknik <i>Classification</i> dengan algoritma <i>Support Vector Machine</i> (SVM), <i>Stochastic Gradient Boosting</i> (SGB), dan <i>Penalized Logistic Regression</i> (PLR) untuk memprediksi <i>ischemic stroke</i>	Penelitian ini menggunakan teknik <i>Classification</i> dengan algoritma <i>Logistic Regression</i> (LR), <i>Artificial Neural Network</i> (ANN), dan <i>Classification and Regression Trees</i> (CART) untuk memprediksi penyakit jantung	Penelitian ini menggunakan teknik <i>Classification</i> dengan algoritma <i>Sequential Minimal Optimization</i> (SMO), <i>Naïve Bayes</i> , <i>Bagging SMO</i> , dan <i>Artificial Neural Network</i> (ANN) untuk memprediksi penyakit jantung koroner	Penelitian ini menggunakan teknik <i>Classification</i> dengan algoritma <i>Decision Tree J48</i> , <i>PART</i> , dan <i>JRip</i> untuk membuat <i>knowledge-based system</i> yang dapat mendiagnosis penyakit diabetes
Keterbatasan	Penelitian ini dapat memberikan hasil yang berbeda tergantung pada situasi dan kondisi tertentu.	Penelitian ini membutuhkan simulasi lebih lanjut untuk mendapatkan hasil perbandingan algoritma yang lebih akurat dan terpercaya.	Penelitian ini memiliki kesulitan dalam memperkenalkan teknik yang dipakai karena (1) sifat <i>black box</i> dari <i>artificial neural network</i> (2) cara memvalidasi <i>neural network</i> yang sudah terlatih. Penelitian ini juga memiliki kekurangan data dari beberapa faktor yang mempengaruhi seseorang memiliki penyakit jantung.	Penelitian ini membutuhkan data, fitur, dan algoritma yang lebih banyak dan luas untuk mendapatkan hasil yang lebih baik.	Penelitian ini melakukan penyimpanan dataset diabetes secara manual sehingga membuat proses preprocessing menjadi sulit.

Tabel 2.3 Penelitian Terdahulu Bagian I (lanjutan)

Atribut	Jurnal 1	Jurnal 2	Jurnal 3	Jurnal 4	Jurnal 5
Hasil	<p>Penelitian ini bertujuan untuk mengembangkan sebuah sistem yang dapat mendiagnosis penyakit asma. Hasil dari penelitian ini menyatakan bahwa <i>Deep Neural Network</i> (DNN) memiliki tingkat akurasi yang paling baik yaitu 98%. DNN mampu mengklasifikasikan data dengan akurasi yang lebih baik dibandingkan dengan algoritma <i>Support Vector Machine</i> (SVM), yang lebih populer dan lebih banyak diteliti.</p>	<p>Penelitian ini bertujuan menggunakan beberapa algoritma <i>data mining</i> untuk memprediksi <i>ischemic stroke</i>. Algoritma – algoritma yang dipakai dibandingkan menggunakan beberapa matriks pengukuran performa, seperti akurasi, <i>area under receiver operating characteristic curve</i> (AUC), specificity, <i>positive predictive value</i>, dan <i>negative predictive value</i>. Hasil penelitian menunjukkan bahwa algoritma SVM memiliki performa paling baik dalam memprediksi dibandingkan algoritma lain.</p>	<p>Penelitian ini menggunakan tiga algoritma <i>data mining</i> dan membandingkannya secara performa. <i>Neural network</i> melakukan klasifikasi lebih akurat daripada metode lain. Keuntungan dari <i>neural network</i> adalah hubungan antara tanda – tanda penyakit dan diagnosis tidak bersifat satu arah saja. Penting bagi model <i>neural network</i> untuk dapat mempelajari data sebanyak mungkin. Keuntungan terpenting dari <i>neural network</i> adalah dapat menarik kesimpulan yang konsisten dan dapat dibuat dan dievaluasi menggunakan data dengan jumlah besar.</p>	<p>Penelitian ini menerapkan beberapa algoritma <i>data mining</i> untuk mengidentifikasi indikator dari penyakit jantung koroner. Algoritma <i>data mining</i> dengan <i>feature selection</i> dan <i>creation</i> digunakan untuk meningkatkan tingkat keakuratan klasifikasi. Akurasi tertinggi dicapai oleh SMO dengan <i>feature selection</i> dan metode <i>feature creation</i> dengan akurasi 94.08%.</p>	<p>Penggunaan teknik <i>data mining</i> untuk membuat <i>knowledge base</i> dari <i>knowledge-based system</i> dapat menjadi sebuah fitur yang penting dalam sistem yang dibuat. Penelitian ini membandingkan tiga algoritma dan berdasarkan hasil penelitian, algoritma J48 mendapatkan hasil evaluasi paling baik. Melalui penelitian ini juga dapat dibuktikan bahwa <i>knowledge-based system</i> dapat membantu proses pembuatan keputusan yang membutuhkan proses pembelajaran dan nalar.</p>

Sumber : Tomita et al. [22], Arslan et al. [34], Khemphila dan Boonjing [24], Alizadehsani et al. [10], Eyasu et al. [35]

Tabel 2.4 juga merupakan tabel perbandingan penelitian terdahulu yang telah dirangkum dan dinilai memiliki kesamaan dengan topik yang diambil dalam penelitian ini. Penelitian – penelitian yang dicantumkan pada tabel 2.4 adalah *Non-invasive detection of fasting blood glucose level via electrochemical measurement of saliva* oleh S. Malik, R. Khadgawat, S. Anand, dan S. Gupta pada tahun 2016, *Real-Data Comparison of Data Mining Methods in Prediction of Diabetes in Iran* oleh L. Tapak, H. Mahjub, O. Hamidi, dan J. Poorolajal pada tahun 2013, *Building predictive models for MERS-CoV infections using data mining techniques* oleh I. Al-Turaiki, M. Alshahrani, dan T. Almutairi pada tahun 2016, *Automatic Prediction of Rheumatoid Arthritis Disease Activity from the Electronic Medical Records* oleh C. Lin, E. W. Karlson, H. Canhao, T. A. Miller, D. Dligach, P. J. Chen, R. N. G. Perez, Y. Shen, M. E. Weinblatt, N. A. Shadick, R. M. Plenge, dan G. K. Savova pada tahun 2013, dan *Word2Vec inversion and traditional text classifiers for phenotyping lupus* oleh C. A. Turner, A. D. Jacobs, C. K. Marques, J. C. Oates, D. L. Kamen, P. E. Anderson, dan J. S. Obeid pada tahun 2017. Penelitian – penelitian ini diperoleh melalui sumber jurnal ilmiah yaitu *Science Direct* dan *National Center for Biotechnology Information* (NCBI) dengan menggunakan kata kunci seperti *data mining*, *medical data mining*, *data mining techniques to predict disease* dan *machine learning techniques to predict disease*.

Tabel 2.4 Penelitian Terdahulu Bagian II

Atribut	Jurnal 6	Jurnal 7	Jurnal 8	Jurnal 9	Jurnal 10
Penulis	S. Malik, R. Khadgawat, S. Anand, dan S. Gupta	L. Tapak, H. Mahjub, O. Hamidi, dan J. Poorolajal	I. Al-Turaiki, M. Alshahrani, dan T. Almutairi	C. Lin, E. W. Karlson, H. Canhao, T. A. Miller, D. Dligach, P. J. Chen, R. N. G. Perez, Y. Shen, M. E. Weinblatt, N. A. Shadick, R. M. Plenge, dan G. K. Savova	C. A. Turner, A. D. Jacobs, C. K. Marques, J. C. Oates, D. L. Kamen, P. E. Anderson, dan J. S. Obeid
Judul	<i>Non-invasive detection of fasting blood glucose level via electrochemical measurement of saliva [7]</i>	<i>Real-Data Comparison of Data Mining Methods in Prediction of Diabetes in Iran [25]</i>	<i>Building predictive models for MERS-CoV infections using data mining techniques [19]</i>	<i>Automatic Prediction of Rheumatoid Arthritis Disease Activity from the Electronic Medical Records [36]</i>	<i>Word2Vec inversion and traditional text classifiers for phenotyping lupus [37]</i>
Teknik DM	<i>Classification</i>	<i>Classification</i>	<i>Classification</i>	<i>Classification</i>	<i>Classification</i>
Obyek Penelitian	Memprediksi penyakit diabetes melitus menggunakan <i>Fasting blood glucose level</i> (FBGL)	Memprediksi penyakit diabetes melitus	Memprediksi penyakit MERS-CoV	Memprediksi <i>Rheumatoid Arthritis</i>	Mengidentifikasi <i>Systemic Lupus Erythematosus</i>
Tipe Penyakit	Diabetes Melitus	Diabetes Melitus	MERS-CoV	Rheumatoid Arthritis	Lupus

Tabel 2.4 Penelitian Terdahulu Bagian II (lanjutan)

Atribut	Jurnal 6	Jurnal 7	Jurnal 8	Jurnal 9	Jurnal 10
Perbandingan	Penelitian ini menggunakan teknik <i>Classification</i> dengan algoritma <i>Logistic Regression</i> , <i>Artificial Neural Network</i> (ANN), dan <i>Support Vector Machine</i> (SVM) untuk memprediksi nilai FBGL dari variasi elektrolit dalam air liur.	Penelitian ini menggunakan teknik <i>Classification</i> dengan algoritma <i>Logistic Regression</i> , <i>Fisher Linear Discriminant Analysis</i> (LDA), <i>Neural Network</i> , <i>Support Vector Machine</i> (SVM), <i>Fuzzy C-Means</i> , dan <i>Random Forests</i> untuk mengklasifikasi pasien yang memiliki penyakit diabetes dan tidak memiliki penyakit diabetes.	Penelitian ini menggunakan teknik <i>Classification</i> dengan algoritma <i>Naïve Bayes</i> dan <i>J48 Decision Tree</i> untuk memprediksi tingkat kestabilan dan kesembuhan pasien dari infeksi MERS-CoV	Penelitian ini menggunakan teknik <i>Classification</i> dengan algoritma <i>Logistic Regression</i> , <i>Support Vector Machine</i> (SVM), <i>Multilayer Perceptron</i> , dan <i>Naïve Bayes</i> untuk mendeteksi rheumatoid arthritis	Penelitian ini menggunakan teknik <i>Classification</i> dengan algoritma <i>Neural Network</i> , <i>Random Forests</i> , <i>Naïve Bayes</i> , <i>Support Vector Machine</i> , dan <i>Word2Vec</i> untuk mendeteksi <i>systemic lupus erythematosus</i>
Keterbatasan	Penelitian ini tidak menggunakan parameter kesehatan yang lebih relevan seperti indeks massa tubuh. Jumlah subyek dalam penelitian juga perlu ditambahkan.	Penelitian ini memiliki <i>sample size</i> pengamatan pasien diabetes yang rendah sehingga mempengaruhi beberapa algoritma <i>Classification</i> yang digunakan.	Penelitian ini tidak memiliki jumlah data pasien yang banyak sehingga performa model yang dihasilkan tidak begitu baik.	Penelitian ini meneliti data medis dari satu institusi saja. Selain itu institusi memakai <i>Unified Medical Language Systems</i> (UMLS) <i>Concept Unique Identifier</i> (CUI) untuk mengklasifikasi rheumatoid arthritis, sementara institusi medis lain belum tentu memakai UMLS CUI	Penelitian ini belum memakai data historis keluarga dalam model. Selain itu masih terdapat kata – kata yang tidak perlu ditambahkan saat memakai algoritma <i>Word2Vec Inversion</i> .

Tabel 2.4 Penelitian Terdahulu Bagian II (lanjutan)

Atribut	Jurnal 6	Jurnal 7	Jurnal 8	Jurnal 9	Jurnal 10
Hasil	<p>Penelitian ini mengaplikasikan teknik <i>machine learning</i> untuk memperlihatkan potensi penggunaan air liur atau <i>saliva</i> sebagai alternatif lain untuk memprediksi FGBL dari pasien. Metode SVM dengan kernel <i>radial basis function</i> (RBF) dapat mendeteksi nilai FGBL diatas atau dibawah 120 mg/dl dengan tingkat akurasi 85%.</p>	<p>Penelitian ini mendapatkan hasil bahwa algoritma SVM memiliki tingkat akurasi, kekhususan, dan sensitivitas yang paling tinggi diantara algoritma lain. Algoritma lain juga memiliki tingkat akurasi yang baik, tetapi memiliki nilai sensitivitas yang rendah.</p>	<p>Penelitian ini menggunakan algoritma <i>Naive Bayes</i> dan <i>J48 Decision Tree</i> untuk menghasilkan model yang dapat memprediksi tingkat kestabilan dan kesembuhan dari pasien MERS-CoV. Performa dari model dievaluasi dan didapatkan rata – rata tingkat akurasi dari model berkisar antara 53.6% - 71.58%.</p>	<p>Penelitian ini menghasilkan sistem yang dapat melihat tanda – tanda rheumatoid arthritis yang dilihat berdasarkan <i>electronic medical record</i>. Melalui penelitian ini didapatkan hasil bahwa algoritma SVM dengan hasil data laboratorium dan pengkodean UMLS memberikan hasil yang paling baik.</p>	<p>Penelitian ini mendapatkan hasil bahwa kombinasi dari algoritma <i>Neural Network</i> dan <i>Random Forest</i> dengan <i>Bag-of-Words</i> (BOW) dan <i>Unified Medical Language Systems</i> (UMLS) <i>Concept Unique Identifier</i> (CUI) memiliki kemampuan untuk mengklasifikasi catatan medis yang tidak terstruktur.</p>

Sumber : Malik et al. [7], Tapak et al. [25], Al-Turaiki et al. [19], Lin et al. [36], Turner et al. [37]

2.9 Ringkasan Penelitian Terdahulu

Penelitian – penelitian yang dijelaskan pada Tabel 2.3 dan Tabel 2.4 menggunakan teknik klasifikasi (*classification*) dalam mendiagnosis atau memprediksi suatu penyakit. Penelitian yang dilakukan oleh Tomita et al. menunjukkan bahwa *Deep Neural Network* (DNN) memiliki tingkat akurasi yang paling baik, diikuti dengan *Support Vector Machine* (SVM) dalam memprediksi penyakit asma [22]. Penelitian yang dilakukan oleh Arslan et al. memprediksi penyakit *ischemic stroke* [34]. Penelitian ini menunjukkan algoritma SVM memiliki performa yang paling baik, yang dievaluasi menggunakan beberapa matriks pengukuran performa. Khemphila dan Boonjing melakukan penelitian untuk membandingkan beberapa algoritma *data mining* untuk memprediksi penyakit jantung [24]. *Artificial Neural Network* (ANN) memiliki akurasi yang lebih tinggi dibandingkan algoritma lain. Penelitian ini memiliki keterbatasan akan jumlah data untuk beberapa faktor resiko penyakit jantung. Penelitian yang dilakukan oleh Alizadehsani et al. menggunakan beberapa algoritma *data mining* untuk mendiagnosis penyakit jantung koroner [10]. Penelitian ini mendapatkan algoritma *Sequential Minimal Optimization* (SMO) memiliki akurasi paling tinggi dengan tambahan *feature selection* dan *creation*. Eyasu et al. mengembangkan *knowledge-based system* menggunakan metode *data mining* untuk mendiagnosis penyakit diabetes [35]. Penelitian tersebut mendapatkan hasil bahwa algoritma J48 mendapatkan hasil evaluasi paling baik. Selain itu dilihat bahwa *knowledge base* dari *knowledge-based system* dapat dibuat menggunakan proses *data mining*. Malik et al. meneliti kandungan kimia yang terdapat pada air liur menggunakan beberapa algoritma *data mining* untuk memprediksi penyakit diabetes [7]. Algoritma SVM dengan kernel *radial basis function* memiliki performa yang paling baik. Tapak et al. menggunakan data *real-time* untuk memprediksi penyakit diabetes [25]. Penelitian ini membandingkan algoritma klasifikasi konvensional dengan algoritma *data mining*. Hasil penelitian ini menunjukkan bahwa SVM memiliki keunggulan dibandingkan algoritma lain. Al-Turaiki et al. membuat model prediktif untuk memprediksi penyakit MERS-CoV menggunakan dua teknik *data mining*, J48

Decision Tree dan *Naïve Bayes* [19]. Hasil yang didapatkan tidak begitu memuaskan dikarenakan jumlah *training data* yang digunakan sedikit sehingga mempengaruhi tingkat keakuratan dari teknik *data mining*. Lin et al. melakukan penelitian untuk memprediksi secara otomatis penyakit *rheumatic arthritis* berdasarkan *electronic health record* [36]. Hasil yang didapatkan menunjukkan SVM dengan menggunakan kernel linear menghasilkan akurasi yang dapat dibandingkan dengan seorang ahli. Turner et al. menggunakan beberapa algoritma *machine learning* untuk mengidentifikasi pasien dengan penyakit lupus [37]. Beberapa algoritma seperti *neural network* dan *random forest* dengan tambahan fitur *bag-of-words* dan *concept unique identifier* terbukti mampu mengklasifikasi rekam medis yang tidak terstruktur jika dibandingkan dengan metode dasar seperti pengkodean ICD-9 dengan metode *Naïve Bayes*.

Penelitian – penelitian terdahulu telah menggunakan berbagai algoritma klasifikasi dalam memprediksi atau mendiagnosis suatu penyakit. Akan tetapi tidak banyak penelitian yang membandingkan lebih dari satu penyakit menggunakan beberapa macam algoritma *data mining*. Selain itu, semua penelitian dilakukan menggunakan Bahasa Inggris dan sebagian besar penelitian yang ditemukan tidak meneliti aspek visualisasi data dari hasil diagnosis atau prediksi. Oleh karena itu, penelitian ini akan melakukan perbandingan algoritma *data mining*, mengevaluasi performa algoritma yang dipakai, dan melakukan visualisasi data penelitian melalui pembuatan model sistem pakar. Penelitian ini menghasilkan diagnosis dalam bentuk Bahasa Indonesia. Hasil evaluasi algoritma dalam menghasilkan diagnosis juga akan divisualisasikan menggunakan program *Python*.