Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 7, No. 1, February, Pp. 69-80

Rule-based disease classification using text mining on symptoms extraction from electronic medical records in Indonesian

Alfonsus Haryo Sangaji*[1], Yuri Pamungkas[2], Diah Risqiwati[3], Supeno Mardi Susiki Nugroho[4], Adhi Dharma Wibawa[5]
Dept. of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Indonesia[1,2]
Informatics Engineering, Universitas Muhammadiyah Malang, Indonesia[3]
Dept. of Electrical Engineering and Computer Engineering, Institut Teknologi Sepuluh Nopember, Indonesia[4,5]

Article Info

Keywords:
Electronic Medical Record, Symptoms Extraction, Rule-Based feature extraction, Random Forests, Support Vector Machine, SMOTE

Article history:
Received: December 12, 2021
Accepted: January 15, 2022
Published: February 28, 2022

Cite:
A. H. Sangaji, Y. Pamungkas, S. M. S. . Nugroho, and A. D. Wibawa, "Rule-based Disease Classification using Text Mining on Symptoms Extraction from Electronic Medical Records in Indonesian", *KINETIK*, vol. 7, no. 1, pp. 69-80, Feb. 2022. https://doi.org/10.22219/kinetik.v7i1.1377

*Corresponding author.
Alfonsus Haryo Sangaji
E-mail address:
haryo.alfon@gmail.com

Abstract

Recently, electronic medical record (EMR) has become the source of many insights for clinicians and hospital management. EMR stores much important information and new knowledge regarding many aspects for hospital and clinician competitive advantage. It is valuable not only for mining data patterns saved in it regarding the patient symptoms, medication, and treatment, but also it is the box deposit of many new strategies and future trends in the medical world. However, EMR remains a challenge for many clinicians because of its unstructured form. Information extraction helps in finding valuable information in unstructured data. In this paper, information on disease symptoms in the form of text data is the focus of this study. Only the highest prevalence rate of diseases in Indonesia, such as tuberculosis, malignant neoplasm, diabetes mellitus, hypertensive, and renal failure, are analyzed. Pre-processing techniques such as data cleansing and correction play a significant role in obtaining the features. Since the amount of data is imbalanced, SMOTE technique is implemented to overcome this condition. The process of extracting symptoms from EMR data uses a rule-based algorithm. Two algorithms were implemented to classify the disease based on the features, namely SVM and Random Forest. The result showed that the rule-based symptoms extraction works well in extracting valuable information from the unstructured EMR. The classification performance on all algorithms with accuracy in SVM 78% and RF 89%.

1. Introduction

In the big data era like nowadays, thousands of data have been created on a daily basis. This also happens in the medical world. Health Information System (HIS) is one of a system that transforms digitally all processes in healthcare services into the digital form [1]. Within the HIS a big data environment is actually made [2]. Recently, HIS has been used by many healthcare organizations to deliver services such as patient treatment, laboratory, medication, insurance, and human resources management. Considering those piles of data, insights from data analytics could be important information for the organization and can be used to increase the organizational competitive advantages. Electronic medical records (EMR) are one of many data resulting from the Healthcare Information System (HIS). EMR consists of patients' symptoms, data, treatment, and planning for medication including anamnesis [3][4]. Anamnesis is a process to obtain information from a patient by communication activity between a physician and a patient about the disease that is suffered and other related information such as the history of their allergic and initial condition [5]. That information then can be used to build the diagnosis of the patient's disease. This means that within the EMR data lie many important insights and new knowledge for the clinicians to deliver better service to the patients. However, big data analytics need to be done on those EMR data. Text mining is one of many methods in big data analytics for processing and analyzing the content of EMR data to get insights. With those insights clinicians then can evaluate and improve their healthcare services to patients [6]. Several insights that can be explored from the EMR data are patients' symptoms regarding a specific disease for specific analysis, variation of symptoms within one specific disease among patients, variation of medication and drug use from the same disease, and many other important aspects that could be beneficial for the healthcare organization regarding their future strategic planning.

However, EMR is a text file, and its content is typed by healthcare administration staff by using the non-standard way of typing [7]. Some staff could use different words and different codes to represent the same information. Simply, The EMR data is an unstructured format that uses natural language in its implementation. Most of the valuable information embedded in EMR is scattered and disorganized [8][9]. Extraction of meaningful information and comprehensive knowledge related to disease, its symptoms, and relationships remains a challenging research problem

[10]. Information extraction is a process for extracting the most valuable and relevant information from an unstructured into a structured format such as entity forms, objects, events, and many other types [11]. The extraction needs careful examination so that it can facilitate in obtaining new knowledge and insights for medical services. Several previous studies in extracting information from the EMR data mostly have been done in English, besides other languages such as Germany [12], Dutch [13], Bulgarian [14], and Polish [15]. We found very few studies that extracted information from the Indonesian language. Automatic learning of complex structures is difficult, especially if there are not many clinical data available and there are no adequate medical corpora for Indonesian. Similar information extraction applications in other languages cannot be reused for Indonesian data because the basic language structure is different. In addition, clinical applications also have to deal with different description standards and specifications. Therefore, developing an information extraction system specifically for EMR data written in Indonesian is important to be made.

Rule-based natural language processing is becoming less popular because it requires much manual work and is not easy to reuse. However, a formal rule-based approach is also used for complex and structured template extraction and gives reliable results. In this study, we proposed a rule-based extraction information technique for the patient's symptoms and diagnosis from the Indonesian EMR data. Rule-based means, this research does not only carry out cleansing of the EMR data but also restores the text by translating texts or medical terms that are misunderstood.

After that, the obtained symptoms will be classified into five diagnoses classes using the machine learning algorithm. Tuberculosis, malignant neoplasm, diabetes mellitus, hypertensive, and renal failure, are diagnosis classes with the highest prevalence rate in Indonesia [16]. Random Forest (RF) and Support Vector Machine (SVM) are the most applied and efficient machine learning classifiers in text data. Research by Sun and Zhang [17] regarding the diagnosis of diabetic retinopathy based on electronic health records that compare several machine learning methods resulted in an accuracy of 92.69% using RF. Another study by Jamaluddin [18] on diagnosis classification based on EMR using SVM resulted in an accuracy of 91.03%. A study done by [18] can classify patient diagnoses well, but valuable information regarding insights in EMR has not been presented. Based on previous literature, this study uses rule-based to extract symptoms from EMR then applies RF and SVM for the disease classification.

2. Research Method

This chapter describes the methods used to achieve the objectives of the research. The method in this study consists of seven stages, namely, data acquisition, data exploration, pre-processing, symptoms extraction, sentence to vector, balanced dataset, and classification, as shown in Figure 1. Each step is explained in more detail in the following sub-chapters:

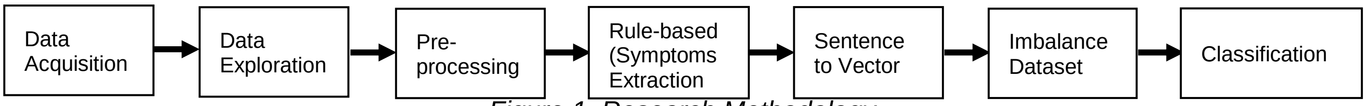

Figure 1. Research Methodology

2.1 Data Acquisition

There were 42431 EMRs in the Indonesian language used in this study. The dataset was obtained from outpatient visits from 2017 to 2018 at a public hospital in Surabaya City, Indonesia. The dataset consists of two columns, the first is patient records as corpus, and the second is physician's diagnosis as a label. Patient records contain patient complaints, disease symptoms, patient history, and other information obtained from the anamnesis. A physician's diagnosis is written with an International Classification of Diseases (ICD) code that the World Health Organization (WHO) standards [19]. The labels were grouped into five diseases based on ICD-11.

1. For physician's diagnosis code in C00-C97: grouped into Malignant Neoplasms.
2. For physician's diagnosis code in I10-I15: grouped into Hypertensive.
3. For physician's diagnosis code in N17-N19: grouped into Renal Failure.
4. For physician's diagnosis code in A15-A19: grouped into Tuberculosis.
5. For physician's diagnosis code in E10-E14: grouped into Diabetes Mellitus.

After labels are grouped into five diseases, the data distribution is changed, as shown in Figure 2.

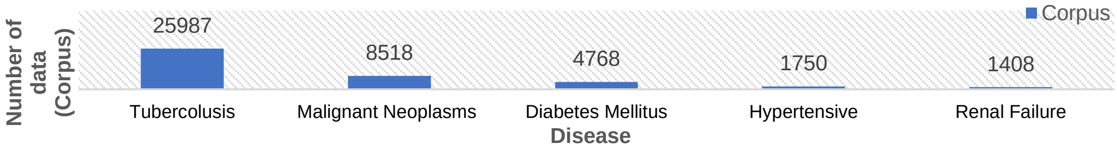

Figure 2. Data Distribution

2.2 Data Exploration

Data exploration is carried out to find parameters in the corpus. Parameters are terms that are often located close to the symptoms in the corpus. Parameters are searched manually. There are seven parameters found from data exploration located close to the symptoms. The parameters found are 'positive sign (+)', 'negative sign (-)', 'no', 'with', 'complaint', 'history', and 'since'. Table 1 shows examples of symptoms located close to Indonesian (Id) and English (Eng) parameters.

Table 1. Location of Symptoms in The Corpus

Parameters	language	Corpus
'positive sign (+)' 'tanda positif (+)'	Eng	The patient came with complaints of loss of consciousness, headache **+,** vomiting **+,** convulsion -
	Id	Pasien datang dengan keluhan penurunan kesadaran, sakit kepala **+,** muntah **+,** kejang -
'negative sign (-)' 'tanda negatif (-)'	Eng	The patient came with complaints of loss of consciousness, headache +, vomiting +, convulsion **-**
	Id	Pasien datang dengan keluhan penurunan kesadaran, sakit kepala +, muntah +, kejang **-**
'no' 'tidak'	Eng	Referral patient from pulmonary polyclinic with adenocarcinoma lung, complained of right chest pain, **no** cough phlegm
	Id	Pasien rujukan dali poli paru dengan paru adenokarsinoma, keluhan nyeri dada kanan, **tidak** batuk dahak
'with' 'dengan'	Eng	Referral patient from pulmonary polyclinic **with** adenocarcinoma lung, complained of right chest pain, no coughing up phlegm
	Id	Pasien rujukan dali poli paru **dengan** paru adenokarsinoma, keluhan nyeri dada kanan, tidak batuk dahak
'complaint' 'keluhan'	Eng	The patient came with **complaints** of loss of consciousness, headache +, vomiting +, convulsion -
	Id	Pasien datang dengan **keluhan** penurunan kesadaran, sakit kepala +, muntah +, kejang -
'history' 'riwayat'	Eng	Referral patients with chronic bronchitis. **history** of cough with blood
	Id	Pasien rujukan dengan brinkitis kronis, **riwayat** batuk berdarah
'since' 'sejak'	Eng	lump in the right jaw **since** a year ago
	Id	Benjolan di rahang kanan **sejak** 1 tahun lalu

2.3 Pre-processing

Text data contains noise in various forms. This step uses to transform raw data into data ready for analysis. There are six steps carried out in this stage. First, remove duplicate corpus and missing value. Second, remove punctuation exclude positive sign (+), negative sign (-), dot (.), and commas (,). The third is to remove numbers except for the numbers that are a part of the term (for example, DMT2 means Diabetes Mellitus Type 2). Forth, replace words that were in parameter terms with base words. Fifth, the stopword removal to remove word/term considered unimportant exclude 'no', 'with', and 'since' ('tidak', 'dengan', dan 'sejak'). The stopword document is based on research conducted by Tala on A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia [20]. Sixth, replace terms in the corpus consisting of 2-4 characters based on medical terms.

2.4 Rule-based (Symptoms Extraction)

The rule-based method is used to extract symptoms from EMR. Rule-based is a simple method for expressing knowledge that forms a logical proposition **IF**... **THEN**. **If** the condition is true, **then** do the action, as shown in Figure 3.

IF 'condition1' **AND** 'condition2' …
THEN 'action'

Figure 3. Rule-based Method

In this step, parameters are as conditions and get symptoms as an action. Before or after the parameters (P), there were symptoms (S) as shown in Table 2. w is a word (not symptoms and parameters), k is the number of symptoms, and n is the number of words in a symptom.

Table 2. Symptoms Location

Parameter Location	Symptom location
Before Parameter	$\ldots w, S_1^1 \ldots\, S_n^1, P, w \ldots, S_1^k \ldots\, S_n^k, P, w \ldots$
After Parameter	$\ldots w, P, S_1^1 \ldots\, S_n^1, w \ldots, P, S_1^k \ldots\, S_n^k, w \ldots$

For example, the patient came with complaints of loss of consciousness, headache +, vomiting +, convulsion -. There are two parameters in that corpus, complaint and positive sign (+). Based on that, the symptoms were located after complaint and before positive signs. Then the extracted symptoms for that corpus are loss of consciousness, headache and vomiting. Rule-based that used consists of five rules, as shown in Table 3.

Table 3. Rules (Symptoms Extraction)

Rules	Logical Proposition

1	**IF** '*positive sign (+)* in corpus'
	THEN 'get 1-5 words before "*positive sign (+)"* as symptoms'

2	**IF** '*complaint* in corpus' AND '*no* not in corpus' AND '*negative sign (-)* not in corpus'
	THEN 'get 1-5 words after "*complaint"* as symptoms'

3	**IF** '*with* in corpus' AND '*no* not in corpus' AND '*negative sign (-)* not in corpus'
	THEN 'get 1-5 words after "*with"* as symptoms'

4	**IF** '*history* in corpus' AND '*no* not in corpus' AND '*negative sign (-)* not in corpus'
	THEN 'get 1-5 words after *"history"* as symptoms'

5	**IF** '*since* in corpus' AND '*no* not in corpus' AND '*negative sign (-)* not in corpus'
	THEN 'get 1-5 words before *"since"* as symptoms'

2.5 Sentence to Vector

Numeric symbols are needed to translate natural language into machine learning problems. Sentence embedding is a method to implement that by distributing the representation of sentences on vector space [21]. CBOW (Continuous Bag-Of-Words) model in word to vector (word2vec) technique that proposed by Mikolov et al [22] used to generate sentence to vector (sent2vec). CBOW predicts the probability of word by word within a specific size window [23]. This study used four windows and 400 dimensions. The average of the dimensions word2vec in a corpus used generated sent2vec, as shown in Figure 4.

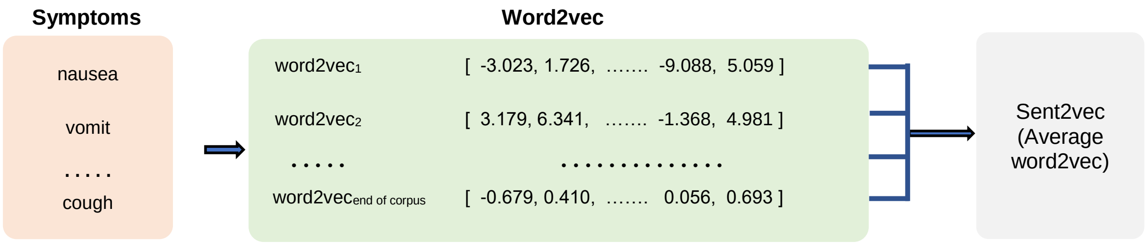
Symptoms
nausea
vomit
.....
cough
Word2vec
word2vec1 [-3.023, 1.726, ... -9.088, 5.059]
word2vec2 [3.179, 6.341, ... -1.368, 4.981]
....
word2vecend of corpus [-0.679, 0.410, ... 0.056, 0.693]
Sent2vec (Average word2vec)

Figure 4. Sentence to Vector

2.6 Imbalance Dataset

Imbalanced data is a classification set with biased or skewed class proportions [24]. As shown in Figure 5, EMR data after-preprocessing is imbalanced. A large proportion of the data set are called majority classes, and the smaller proportion is minority classes. Imbalanced can cause problems in the classification task because the model can over-fit the majority class and under-fit the minority class [25]. To solve that problem, in this step, the re-sampling technique is applied [26]. The re-sampling technique that is applied is oversampling technique. Oversampling is a technique to increase the number of samples in the minority class [27]. SMOTE (Synthetic Minority Over-sampling Technique) is an oversampling technique used [28]. Chawla et al [29] showed that SMOTE could improve classifier performance for minority classes. The basic idea of the SMOTE algorithm for balancing datasets is to synthesize new minority samples by adopting linear interpolation on homogeneous neighbour samples [30]. The number of nearest neighbours to consider when creating a new synthetic element (K) in this study is 1.

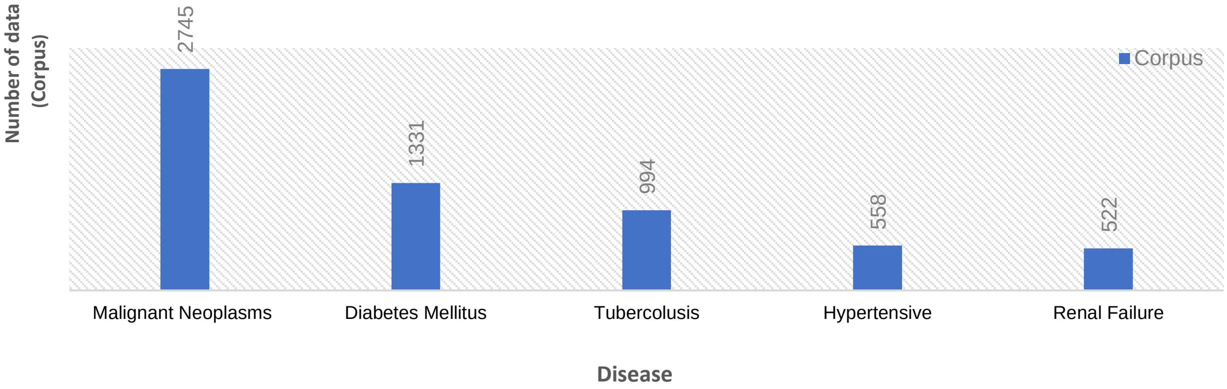

Figure 5. Data Distribution (after symptoms extraction)

2.7 Classification

Comparing classification with more than one kind of model was performed to determine which one is the best model in classifying diseases, especially in text data. The models are Random Forest (RF) and Support Vector Machine (SVM), such as in several medical studies [17][18].

2.7.1 Random Forest (RF)

RF is an ensemble learning method that grows many random decision trees [31]. The procedure for determining the final prediction is called bagging. Bagging works by random resampling with the replacement of the original data. The new dataset will be used to grow multiple random decision trees. Then each decision tree casts a vote for the class. The most popular class determines the final prediction as called the majority vote. The larger the number of predictors, the more trees need to be planted for good performance. The steps in the RF method are: First specifies how many trees to build. Second, from the training data, pick random data points by bootstrap. Third, split the node using the GINI method. Forth, perform training tasks on each decision tree. Fifth, vote to determine the optimal solution. The Gini method is shown in Equation 1, where D is data, n is the number of classes, i is a class attribute, and p_i is the ratio of the number of data labelled i class in the D data.

$$Gini\ Index\ (D) = 1 - \sum_{i=1}^{n} {p_i}^2 \tag{1}$$

2.7.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning classifier that can solve non-linear problems. SVM transforms the original feature space into a high-dimensional feature space by fitting a distance-maximizing hyperplane between classes through the use of kernels that can accommodate any functional form [32]. Selecting the appropriate kernel function is important to build an optimal hyperplane. There four SVM kernels are used for the experiments in this paper Linear in Equation 2, Radial Bias Function (RBF) in Equation 3, Polynomial in Equation 4, and Sigmoid in Equation 5. Where d is a degree, r is coef 0, and γ is gamma.

$$K(x_i, x_j) = (x_i, x_j) \tag{2}$$

$$K(x_i, x_j) = (\gamma(x_i, x_j) + r)d \tag{3}$$

$$K(x_i, x_j) = exp(-\gamma\|x_i, x_j\|2)\gamma \tag{4}$$

$$K(x_i, x_j) = tanh(\gamma(x_i, x_j) + r) \tag{5}$$

2.7.3 Evaluation

Evaluations are needed to select the best-performing model. This study used f1-score and confusion matrix for performance measurement. In the confusion matrix, there are values of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Then the measurements were calculated using $Precision$ in Equation 6, $Recall$ in Equation 7, $F1 - Score$ in Equation 8, and $Accuracy$ in Equation 9.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 - Score = 2 * \frac{recall * precision}{recall + precision} \tag{8}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

3. Results and Discussion

This chapter explained results from several experiments and several highlights for discussion, consisting of two main topics. The first is rule-based to extract the symptoms of each disease. The second is obtaining several findings from the classification experiment.

3.1 Symptoms Extraction

This section shows the symptoms of a disease that the patient complains about. Pre-processing and rule-based symptoms extraction (SE) that applied make the raw dataset significantly reduced from 42431 to 6150 data. Table 4 shows data distribution in each disease. Data reduction has occurred in each data class. Removed data indicates that there are no symptoms in the corpus. That means there were 6150 corpora in the dataset that had symptoms. The highest decrease data (24993 corpora) is tuberculosis class when compared to other classes.

Table 4. Data Distribution (after rule-based symptoms extraction)

Disease	Corpus	
	Before	After
Tuberculosis	25987	994
Malignant Neoplasms	8518	2745
Diabetes Mellitus	4768	1331
Hypertensive	1750	558
Renal Failure	1408	522
Total	42431	6150

Table 5 shows the result of rule-based SE. That algorithm can extract symptoms but not with several descriptions of the symptoms. For example, in a malignant neoplasm patient complains of a lump appearing in the lower right chest, the algorithm extracted the lump appears right chest without lower. That happens because lower is in the stopword list. Another example from tuberculosis, the patient complaint of phlegm (+) yellow color. The algorithm extracted the phlegm without yellow color. That happens because of the first rule, as shown in Table 3, there were positive signs as a parameter before the symptom's description.

Table 5. Symptoms Extraction Result

Disease	language	Corpus	Symptoms Extraction
Renal failure	Eng	Complaint of the swollen foot that since two months, back pain +, shortness of breath - fever -	[Swollen foot], [back pain]
	Id	keluhan kaki bengkak sejak dua bulan. nyeri pinggang +. sesak - demam –	[kaki bengkak], [nyeri pinggang]
Malignant neoplasms	Eng	A lump appears in the lower right chest for three months, pain (+), fever (-)	[lump appears right chest], [pain]
	Id	muncul benjolan di dada kanan bagian bawah sejak tiga bulan, nyeri (+), demam (-)	[muncul benjolan dada kanan], [nyeri]
Diabetes Mellitus	Eng	a new patient with complaints of slowly blurring the left eye for three months. History dm +	[slowly blurring left eye], [diabetes mellitus]
	Id	pasien baru dengan keluhan mata kiri kabur perlahan sejak tiga bulan yg lalu. riwayat dm +	[mata kiri kabur perlahan], [diabetes mellitus]

Tuberculosis	Eng	patients with complaints of shortness of breath for a week, cough (+) > a week, phlegm (+) yellow colour, decreased appetite (+), weight loss (+), night sweats (+)	[shortness of breath], [cough], [phlegm], [decreased appetite], [weight loss], [night sweats]
	Id	pasien dengan keluhan sesak napas satu minggu, batuk (+) > satu minggu, dahak (+) warna kuning, penurunan nafsu makan (+), penurunan berat badan (+), keringat malam (+)	[sesak napas], [batuk], [dahak], [penurunan nafsu makan], [penurunan berat badan], [keringat malam]
Hypertensive	Eng	dizziness (+), shoulder stiffness (+), not taking medication for a week (no time to control)	[dizziness], [shoulder stiffness]
	Id	pusing (+), bahu terasa kaku (+), tidak minum obat satu minggu (tidak sempat kontrol)	[pusing], [bahu kaku]

Table 6, Table 7, Table 8, Table 9, and Table 10 shows the ten highest symptoms in EMR according to their disease. The count is the number of patient complaints according to their symptoms. In 6, nausea is the highest symptom that patients complain about in tuberculosis, with 27 counts obtained. In tuberculosis, the highest symptom is a cough with 157 counts. In diabetes mellitus, diabetes mellitus is the highest with 150 counts. In malignant neoplasm, the highest symptom is a cough with 134 counts. In hypertensive, hypertensive is the highest symptom with 132 counts.

Table 6. Symptoms of Renal Failure

Renal Failure	
Symptoms	Count
Nausea	27
Acquired cystic kidney disease	24
Chronic kidney disease V	22
Anemia	19
Cough	17
Acute Kidney Injury	15
Chronic kidney disease	14
Hypertensive	12
Continuous Ambulatory Peritoneal Dialysis	11
End-stage renal disease Hemodialysis	10

Table 7. Symptoms of Tuberculosis

Tuberculosis	
Symptoms	Count
Cough	157
Night sweats	71
Pulmonary tuberculosis	46
Weight loss	38
Fever	34
Phlegm	33
Lose weight	32
Dyspnea	29
Tuberculous spondylitis	25
Decreased appetite	18

Table 8. Symptoms of Diabetes Mellitus

Diabetes Mellitus	
Symptoms	Count
Diabetes Mellitus	150
Diabetes Mellitus type 2	125
Diabetes Mellitus type	30
Hypertensive	23
Blurred vision	22
Insulin	17
Thrombotic stroke	14

Cough	13
Headache	12
Nausea	12

Table 9. Symptoms of Malignant Neoplasms

Malignant neoplasms	
Symptoms	Count
Cough	134
Chest pain	79
Pain	73
Dyspnea	67
Weight loss	47
Nausea	45
Neck lump	44
Abdominal pain	38
Shortness of breath	34
Fever	33

Table 10. Symptoms of Hypertensive

Hypertensive	
Symptoms	Count
Hypertensive	132
Chest pain	14
Headache	14
Cardiac chest pain	13
Cardiac	13
Cough	13
Dyspnea	13
Swollen Foot	13
Diabetes Mellitus	10
Stroke	9

3.2 Classifier Model Results

This section presents the evaluation results of the classification model experiments. Precision, recall, f1-score, and accuracy were applied for evaluation. The dataset is split into 80% training data and 20% test data. The experiment is divided into two parts, the first experiment without applying the SMOTE method, the second experiment using the SMOTE method.

3.2.1 Classification Without SMOTE

This section presents the evaluation results of the classification model without SMOTE applied. This experiment use random forest for the classifier. Data distribution in this experiment is shown in Figure 6. The datasets split into 80% training data (4920) and 20% test data (1230).

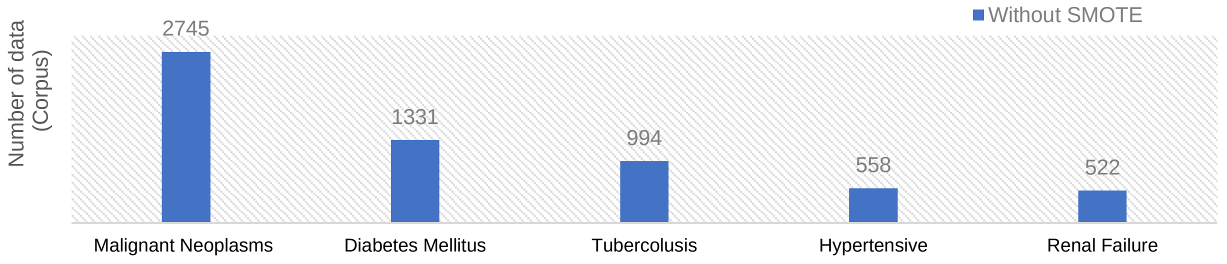

Figure 6. Data Distribution after SMOTE

The evaluation results show that the model cannot be applied even though the accuracy value is good with 0.74. The low recall values were found in hypertensive and renal failure classes (0.42 and 0.38), as shown in Table 11. Based on Equation 7, having a low recall value indicates a high FN value. A high FN value is interpreted as a high incorrect classification result. FN means patients were diagnosed with hypertensive or renal failure in actuality, but the results of the classification model were the opposite. An imbalanced dataset makes poor evaluation values. The evaluation results shown in Table 11 have high value in the majority class (Malignant Neoplasms) and low value in minority classes (Hypertensive and Tuberculosis). A balanced dataset improves the performance of the model as shown in Table 12. Where classes having a similar amount of data are implemented (hypertensive and renal failure). Good recall values in both classes were obtained.

Table 11. Classifier Evaluation without SMOTE

Diseases	*Evaluation*		
	Precision	Recall	F1-score
Diabetes Mellitus	0.71	0.72	0.72
Hypertensive	0.76	**0.42**	**0.54**
Malignant Neoplasms	0.74	0.91	0.82
Renal Failure	0.64	**0.38**	**0.48**
Tuberculosis	0.81	0.64	0.72

Table 12. Classifier Evaluation without SMOTE (similar amount of data)

Penyakit	*Evaluasi*		
	Precision	Recall	F1-score
Hypertensive	0.85	0.76	0.80
Renal Failure	0.75	0.84	0.79

3.2.2 Classification With SOMTE

This experiment uses SMOTE method, the distribution data shown in Figure 7. The dataset splits into 80% training data (10980) and 20% test data (2745).

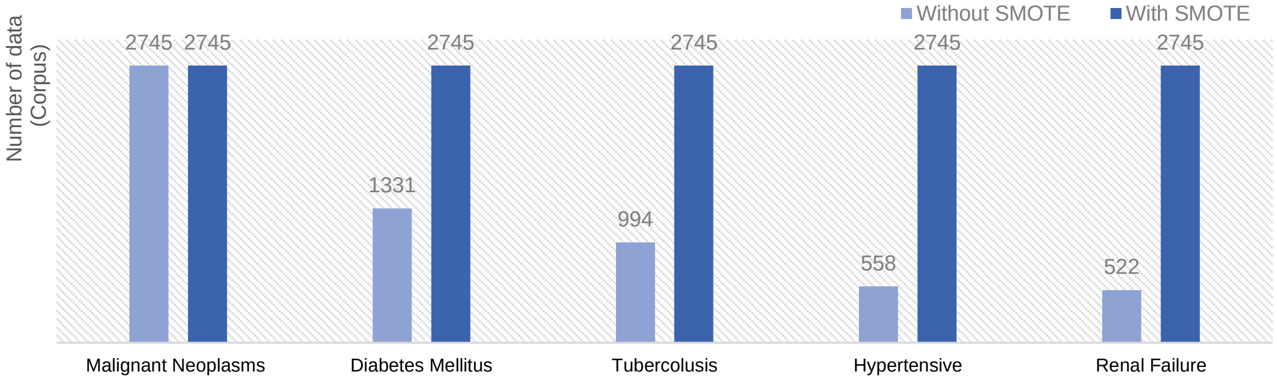

Figure 7. Data Distribution after SMOTE

The best validation score in SVM kernels is RBF with a 0,77 mean F1-score. Validation on SVM (train data) used K-Fold validation (k = 4). The validation result from each kernel shows in Figure 8. The accuracy results on test data shown in Figure 9 are pretty good for both SVM-RBF and RF. RF classifier outperformed SVM-RBF with an accuracy value of 0.89. The result of each disease shows in Table 13. Every disease did not obtain a bad value in precision, recall, or f1-score. The hypertensive and renal failure classes have the highest f1-score with 0.92.

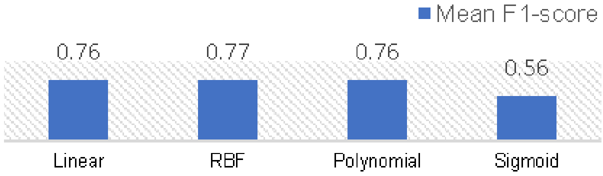

Figure 8. Validation Score Each Kernel

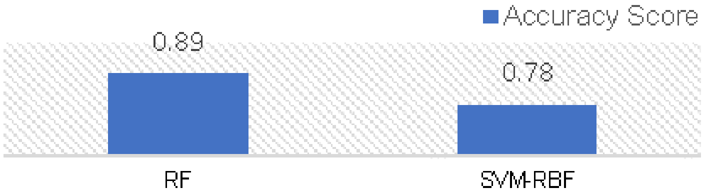

Figure 9. Accuracy Comparison

Table 13. Classifier Evaluation

Class	Classifier	Evaluation		
		Precision	Recall	F1-score
Diabetes Mellitus	RF	0.89	0.85	0.87
	SVM	0.77	0.72	0.75
Hypertensive	RF	0.89	0.94	0.92
	SVM	0.76	0.86	0.80
Malignant Neoplasms	RF	0.86	0.83	0.85
	SVM	0.78	0.77	0.77
Renal Failure	RF	0.92	0.92	0.92
	SVM	0.81	0.75	0.78
Tuberculosis	RF	0.87	0.91	0.89
	SVM	0.80	0.82	0.81

Balanced data makes the model's performance do well. The accuracy of the RF classifier model increased from 0.74 to 0.89. The classifier model with SMOTE improves the evaluation value, as shown in Table 11 and Table 13. Specifically, in minority classes, the F1-score of hypertensive and renal failure classes has increased. Using SMOTE to re-sample the minority data can improve the model's performance.

3.3 Discussion

This study presents the most obtained symptoms from 5 diseases as shown in Table 6 to Table 10. Those symptoms have been validated by our clinicians and confirms well as the most symptoms regarding the disease. Symptoms with a higher frequency mean the most frequent symptoms that are reported by the patient regarding a specific disease. We also found some minor variation in each disease as shown by the symptoms with lower frequency. With this information, clinicians then can be more careful in examining the symptoms reported by the patients.

Moreover, we also found that the result of disease extraction depends fully on the quality of pre-processing steps. As indicated by the extraction result of malignant neoplasms in Table 5 showing that "lower" is a term that represents the location of symptoms. That means the algorithm still has a limitation in extracting the words/terms that affect the description of the symptoms. That happens because the "lower" term is on the stopword list. The research from Tala shows that stopwords can improve classification performance [20]. However, data exploration is needed to observe a word/term that influences the model in some exceptional cases.

Furthermore, the use of SMOTE to oversample the minority data can improve the model's accuracy both in SVM and RF algorithms. The result in this study is also confirmed well with previous research was done by Chawla et al [29] that showed the SMOTE could improve classifier performance for minority classes. In contrast to a previous study [18] which did not apply information extraction, the accuracy results in this study were under Jamaluddin with a difference of 2%. there is a slight difference in the classifier performance results, but with the addition of new information in the form of disease symptoms.

There were Several challenges found remain becoming homework. The first is how to build a stopword removal that does not have a strong influence on the medical text. The second is we need to enhance the dictionary of medical terms so that it will obtain more accurate symptoms data. Then finally is to enhance the parameters on the rule-base that can improve the information extraction.

4. Conclusion

This study attempts to extract information from the EMR data and pesents the most reported symptoms from 5 the most disesases in Indonesian language using rule-based algorithm. Some symptom variations have also been obtained from the EMR data. The obtained symptoms are also being classified using two machine learning algorithms namely Random Forest and SVM. To overcome the imbalance number of data for classification, SMOTE technique is applied. The best result obtained in this study is the SE + SMOTE + RF approach with an accuracy of 89%. The rule-based method can do the extraction well, even though some drawbacks are still present. Screening and sampling the text data are also important to sense the pre-processing algorithm and avoid too many artifacts that can occur. Symptoms Extraction does not improve model performance well but adds much information for the clinicians. For further research, the information extraction for symptoms and symptom characteristics (size, color, time, and others that describe symptoms) and the other entities such as drugs used, patient activities, and other entities need to be explored to enrich information and gain a better understanding regarding insights covered in EMR data.

References

[1] N. Leon *et al.*, "Routine Health Information System (RHIS) improvements for strengthened health system management," *Cochrane Database of Systematic Reviews,* vol 8, 2020. https://doi.org/10.1002/14651858.CD012012.pub2

[2] J. F. Anderson, “Organization, Powers, and Duties of The United States Public Health Service Today,” *American journal of public health (New York, N.Y.: 1912)*, vol. 3, no. 9, pp. 845–852, 1913. https://doi.org/10.2105/ajph.3.9.845-a

[3] C. T. Lye, “Assessment of US Hospital Compliance With Regulations for Patients,” *Requests for Medical Records. JAMA network open*, vol. 1, no. 6, 2018. https://doi.org/10.1001/jamanetworkopen.2018.3014

[4] A. Cesarani, D. Alpini, D. Brambilla, “Anamnesis and Clinical Evaluation,” *In: Cesarani A. et al. (eds) Whiplash Injuries. Springer, Milano*, 1996. https://doi.org/10.1007/978-88-470-2293-5_11

[5] Matthew A. Cottam, Hana A. Itani, Arch A. Beasley IV and Alyssa H. Hasty, “Links between Immunologic Memory and Metabolic Cycling,” *Journal of immunology (Baltimore, Md.: 1950)*, vol. 200, no. 11, pp. 3681–3689, 2018. https://doi.org/10.4049/jimmunol.1701713

[6] L. Faridah, F. R. Rinawan, N. Fauziah, W. Mayasari, A. Dwiartama, and K. Watanabe, “Evaluation of Health Information System (HIS) in The Surveillance of Dengue in Indonesia: Lessons from Case in Bandung, West Java,” *International journal of environmental research and public health,* vol. 17, no. 5, pp. 1795, 2020. https://doi.org/10.3390/ijerph17051795

[7] S. Sharifi, M. Zahiri, H. Dargahi, and F. Faraji-Khiavi, “Medical record documentation quality in the hospital accreditation,” *Journal of education and health promotion,* vol. 10, no. 76, 2021. https://doi.org/10.4103/jehp.jehp_852_20

[8] Z. Fritz, A. Schlindwein, and A. M. Slowther, “Patient engagement or information overload: patient and physician views on sharing the medical record in the acute setting,” *Clinical medicine (London, England)*, vol. 19, no. 5, pp. 386–391, 2019. https://doi.org/10.7861/clinmed.2019-0079

[9] Y. Wang *et al*., “Clinical information extraction applications: A literature review,” *Journal of biomedical informatics*, vol. 77, pp. 34–49, 2018. https://doi.org/10.1016/j.jbi.2017.11.011

[10] S. R. Jonnalagadda *et al.,* “Automatically extracting sentences from Medline citations to support clinicians' information needs,” *Journal of the American Medical Informatics Association: JAMIA*, vol. 20, no. 5, pp. 995–1000, 2013. https://doi.org/10.1136/amiajnl-2012-001347

[11] S. Hassanpour and C. P. Langlotz, “Information extraction from multi-institutional radiology reports,” *Artificial intelligence in medicine*, vol. 66, pp. 29–39, 20156. https://doi.org/10.1016/j.artmed.2015.09.007

[12] Hahn, Udo, Martin Romacker, and Stefan Schulz. "MEDSYNDIKATE—a natural language system for the extraction of medical information from findings reports." *International journal of medical informatics,* vol. 67, pp. 63-74, 2002. https://doi.org/10.1016/S1386-5056(02)00053-9

[13] Spyns, Peter *et al*., "Medical language processing applied to extract clinical information from Dutch medical documents." MEDINFO'98. IOS Press, pp. 685-689, 1998. https://ebooks.iospress.nl/doi/10.3233/978-1-60750-896-0-685

[14] Boytcheva, Svetla, *et al*. "Some aspects of negation processing in electronic health records." Proc. of International Workshop Language and Speech Infrastructure for Information Access in the Balkan Countries. 2005.

[15] A. Mykowiecka, M. Marciniak, and A. Kupść, “Rule-based information extraction from patients’ clinical data,” *Journal of biomedical informatics*, vol. 42, no. 5, pp. 923-936, 2009. https://doi.org/10.1016/j.jbi.2009.07.007

[16] Research and development agency of the Indonesian Ministry of Health. “2018 National Basic Health Research Report”. Lembaga Penerbit Balitbangkes, 2019.

[17] Y. Sun and D. Zhang, "Diagnosis and Analysis of Diabetic Retinopathy Based on Electronic Health Records," in IEEE Access, vol. 7, pp. 86115-86120, 2019, https://doi.org/10.1109/ACCESS.2019.2918625

[18] M. Jamaluddin and A. D. Wibawa, "Patient Diagnosis Classification based on Electronic Medical Record using Text Mining and Support Vector Machine," *2021 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pp. 243-248, 2021. https://doi.org/10.1109/iSemantic52711.2021.9573178

[19] M. S. C. Almeida, L. F. de Sousa Filho, P. M. Rabello, and B. M. Santiago, “International Classification of Diseases – 11th revision: from design to implementation”, Rev. saúde pública, vol. 54, pp. 104, Dec. 2020. https://doi.org/10.11606/s1518-8787.2020054002120

[20] Tala, F. Z, “A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia”. M.Sc. Thesis. Master of Logic Project. Institute for Logic, Language and Computation. Universiteit van Amsterdam, The Netherlands. 2003.

[21] K. Blagec, H. Xu, A. Agibetov, and M. Samwald, “Neural sentence embedding models for semantic similarity estimation in the biomedical domain,” *BMC bioinformatics*, vol. 20, no. 1, pp. 178, 2019. https://doi.org/10.1186/s12859-019-2789-2

[22] T. Mikolov, K. Chen, G. Corrado and J. Dean, Efficient estimation of word representations in vector space, 2013. https://arxiv.org/abs/1301.3781

[23] M. Arguello Casteleiro *et al.,* “Semantic Deep Learning: Prior Knowledge and a Type of Four-Term Embedding Analogy to Acquire Treatments for Well-Known Diseases,” *JMIR medical informatics*, vol. 8, no. 8, 2020. https://doi.org/10.2196/16948

[24] G. Abdulrauf Sharifai and Z. Zainol, “Feature Selection for High-Dimensional and Imbalanced Biomedical Data Based on Robust Correlation Based Redundancy and Binary Grasshopper Optimization Algorithm,” *Genes*, vol. 11, no. 7, pp. 717, 2020. https://doi.org/10.3390/genes11070717

[25] R. O'Brien and H. Ishwaran, “A Random Forests Quantile Classifier for Class Imbalanced Data,” *Pattern recognition*, vol. 90, pp. 232–249, 2019. https://doi.org/10.1016/j.patcog.2019.01.036

[26] M. Deng, Y. Guo, C. Wang, and F. Wu, “An oversampling method for multi-class imbalanced data based on composite weights,” *PloS one*, vol. 16, no. 11, 2021. https://doi.org/10.1371/journal.pone.0259227

[27] P. Gnip, L. Vokorokos, and P. Drotár, “Selective oversampling approach for strongly imbalanced data,” *PeerJ. Computer science*, vol. 7, 2021. https://doi.org/10.7717/peerj-cs.604

[28] J. Shen, J. Wu, M. Xu, D. Gan, B. An, and F. Liu, “A Hybrid Method to Predict Postoperative Survival of Lung Cancer Using Improved SMOTE and Adaptive SVM,” *Computational and mathematical methods in medicine*, 2021. https://doi.org/10.1155/2021/2213194

[29] N. V. Chawla, K. W. Bowyer, L. O Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-Sampling Technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp.321–357, 2002. https://doi.org/10.1613/jair.953

[30] L. Ma, and S. Fan, “CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests,” *BMC bioinformatics*, vol. 18, no. 1, pp. 169, 2017. https://doi.org/10.1186/s12859-017-1578-z

[31] L. Breiman, “Random forests. Machine learning,” vol. 45, no. 1, pp.5-32, 2001. https://doi.org/10.1023/A:1010933404324

[32] B. Scholkopf and A. J. Smola, “Learning with kernels: Support vector machines, regularization, optimization, and beyond,” *Cambridge, MA: MIT Press*. 2001.