Text Classification of Cancer Clinical Trials Documents Using Deep Neural Network and Fine Grained Document Clustering

Jasmir JASMIR[1], Siti NURMAINI[2], Reza Firsandaya MALIK[3*], Dodo Zaenal ABIDIN[1]

[1].Intelligence System Research Group, STIKOM Dinamika Bangsa Jambi Indonesia
[2].Intelligence System Research Group, Universitas Sriwijaya Palembang Indonesia
[3].Communication Networks and Information Security Research Lab, Universitas Sriwijaya, Palembang Indonesia
Corresspondent author : rezafm@unsri.ac.id

Abstract

Clinical trials are any research studies involve human participation with health safety outcomes. In clinical trials, there is the most important term called the eligibility criteria (eligible and not eligible). The eligibility criteria for clinical trials are usually written in free text, it requires interpretation from a computer to process them. The purpose of this paper is to classify cancer clinical texts from the public dataset at https://clinicaltrials.gov. The proposed algorithm is Supervised Learning such as K-Nearest Neighbor and Decision Tree, Machine Learning such as Support Vector Machine and Random Forest, Deep Neural Network such as Multilayer Perceptron, and Fine Grained Document Clustering. This research has contributed a new classification model for clinical trial documents and computational value or speed improvement. The results shown the highest accuracy at random forest method 90.5% and the lowest accuracy at multilayer perceptron method that is 72.1%

***Keywords**: text classification, clinical trials, Deep Neural Network, fine grained document clustering*

Introduction

Text classification is defined as labeling natural language text documents with classes or categories of predetermined sets [1]. Text classification is an important component in many NLP applications, such as sentiment analysis [2], relationship extraction [3] and spam detection [4][5]. Text classification has also attracted the attention of researchers to continue to develop innovations and testing, including those sourced from clinical texts commonly referred to as clinical trials.

Clinical trials is a type of research that studies how safe it is to help test or care given to patients. [6]. Clinical trials play an important role in translating scientific research into the practice of medical outcomes [7]. In clinical trials, there is a term or the most important part called the eligibility criteria that determines the cost, duration and success of the clinical trial process [8].

Research on eligibility criteria in clinical trials is usually written in free text, but it will be difficult if interpreted by computer. A popular method for processing eligibility criteria is knowledge representation, which often requires extensive knowledge and hard work from experts in the sector of medical coding to identify eligibility criteria [9]. In completing the problem of the feasibility analysis of clinical trials, the optimal method to solve it is artificial intelligence methods such as rule-based systems, traditional machine learning algorithms, and representation learning, such as deep learning architecture. [10][11].

Currently Deep learning technology [12] has achieved extraordinary results in many area, such as computer vision [13] speech recognition[14], and text classification [15]. Vincent Menger [16] states that in some cases approaches with deep learning techniques applied to the classification of clinical texts can produce conclusions that match expectations, but will be different if tested on other clinical datasets and with different domains and different sizes.

The problems raised by Vincent Menger above can be seen in research Bustos and Pertusa 2018 [17]. Aurelia Bustos who conducts research on clinical trials. In this study, they trained, validated, and compared various classification models namely k-Nearest Neighbor, Support Vector Machine, Convolutional Neural Network and FastText. This research utilizes a dataset from "clinical trial". The calculated values are Precision, Recall, F1, and Cohen's K. SVM produces the lowest accuracy results, and kNN obtains the highest accuracy performance similar to the CNN model, but has the lowest computational performance.

However, the value of computational can be increased by one of the methods discussed by Taufik Sutanto[18] in his paper with the theme of the Fine-Grained Document Clustering (FGC) approach that utilizes the ability of search engines to handle big data efficiently, in this paper only tested on the problem of unsupervised learning clustering

Based on the problems from Bustos and Pertusa above, we see an opportunity to conduct further research, namely increasing the value of computational using the Fine Grained document Clustering method

The remainder of this paper is organized as follows. In Section II, we briefly discuss related works. In section III about material and methods which contains a dataset, preprocessing and classification. In section IV about result which contains Precision Recall and F1-Score (F-Measure) and Model Evaluation and Validation. The conclusion and suggestions for future research are summarized in Section V.

RELATED WORKS

Several studies related to the above theme such as the cancer classification with multilayer perceptron [19] [20], clinical text classification [21] [22][23] and eligibility criteria [24][11].

The other research that discusses the clinical text classification of eligibility criteria [25], which discusses the development of methods for the automatic classification of the eligibility criteria to facilitate the matching of ClinicalTrials.gov dataset patient trials for specific populations such as people living with HIV or woman pregnant. The proposed method is able to act as a filter in the search engine for testing patients.

C. Chuan proposes an active deep learning approach to automatically classify clinical trial eligibility criteria[26]. The experimental results showed that active CNN performed significantly better than the K-Nearest Neighbor method. Thus, Y. Ni et al discuss the development of an automatic screening eligibility algorithm to identify patients who meet the core eligibility characteristics of oncology clinical trials[27].

In this study we made a classification model of the eligibility criteria of cancer clinical texts using the deep neural network method, and improved the computational value of the clinical text classification using the fine grained document algorithm

MATERIALS AND METHOD

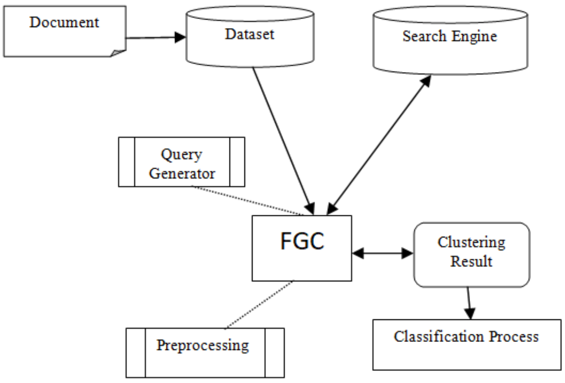

Figure-1. Conceptual Design

Dataset, Data were taken from clinical statements. A total of 6,186,572 extracted from 49,000 Clinical Trial Protocols on cancer originating from the National Library of Medicine, National Institutes of Health: Bethesda, MD, USA, which can be downloaded from https://clinicaltrials.gov. Each clinical trial downloaded is an unstructured XML file [28]. This data comes from the fields of intervention, conditions, and feasibility written in unstructured free text language. Information in the eligibility criteria is a series of phrases and or sentences that are displayed in a free format, such as paragraphs, bulleted lists, enumeration lists, etc.

Preprocessing, Preprocessing has a very important role in the technique and application of text mining. This is the first step in the process of mining text. In this paper, we discuss the three main steps of preprocessing, namely, stopword, stemming and TF / IDF.[29]

All eligibility criteria are converted into a sequence of simple words. information about study interventions and types of cancer added to each feasibility criterion by separating the text into statements, then removing punctuation, white space characters, all non-alphanumeric symbols, separators, and single character words from the extracted text. All words are lowercase letters. We do not delete stop words because, like "or", "and", "for", because they are semantically relevant to clinical statements. Next change numbers, arithmetic marks, comparators to text

Deep Neural Network: Deep neural networks (DNN) are standard feed-forward neural networks that are much greater and much sharper than conventional neural networks[30]. A general deep framework is usually utilized for classification with many hidden layers, and it approves complex hypotheses to be expressed [31][32]. Each layer only receives connections from the previous layer. Networks are trained using DNN supervised learning algorithms

Fine Grained document Clustering (FGC), FGC presume the turned form of cluster hypothesis [33], that is the relevant documents returned in response to a query will inclined to be similar to one-another. FGC uses a combination of loci and relevant clusters concepts to efficiently form clusters. The use of loci makes the computation of clusters' representations efficient, since it only utilizes a small set of documents instead of all documents in a cluster. By using the relevant cluster concept, FGCR does not want pairwise similarity comparisons between a document and all of the clusters. These strategies approve FGC to generate a fine-grained clustering solution efficiently.

Classification. We use some classification methods such as: KNN, MLP, SVM, Random Forest and Decision Tree.

K Nearest Neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).[34] To classify an unknown document, K-NN algorithm identifies the k nearest neighbors in a given document space. K-NN algorithm uses a similarity function such as Euclidean distance or Cosine resemblance to get neighbors. The best option of selecting the grade of k depends upon the dataset or application. The implementation of K-NN algorithm is very easy, but it is computationally intensive, especially when the size of the training documents grows.

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor

$$\text{Eucledian} = \sqrt{\sum_{i=1}^{k} (xi-yi)^2} \tag{1}$$

$$\text{Manhattan} = \sum_{i=1}^{k} |xi-yi| \tag{2}$$

Multilayer Perceptron is consisted of simple neurons named perceptron. As refer to neuron weights in input nodes and generating the output by employing nonlinear activation mathematical function, linear combination will be formed by perceptron through computation of an output neuron from multiple realvalued inputs.[35]

$$\text{Y}= \sigma\left(\sum_{i=1}^{n} wixi+b\right)=\sigma(wTx+b) \tag{3}$$

Support Vector Machine (SVM) is a relatively new classification method [36][37]. Although it is complex algorithm, SVM reaches great classification levels in many areas. SVM is basically a linear two-class classifier. Among the likely hyperplanes between two classes, SVM gets the optimal hyperplane between two classes by maximizing the margin among the closest points of classes. The points prevaricate on the hyperplane boundaries are named support vectors.

For linear kernel the equation for prediction for a new input using the dot product between the input (x) and each When two classes are not linearly partible, SVM projects data points into a higher dimensional scope so that the data points become linearly partible by utilizing kernel techniques. There are several kernels that can be used SVM algorithm. Support vector (xi) is calculated as follows:

$$f(x) = B(0) + \sum (ai * (x,xi)) \tag{4}$$

The polynomial kernel can be written as

$$K(x, xi)=1+\sum (x*xi)^d \tag{5}$$

And exponential as

$$K(x, xi)=\exp(-y*\sum(x-x^2)) \tag{6}$$

Random Forest is an ensemble learning method that grows many random and uncorrelated decision trees. Each tree votes for the test sample class. The most favorite class specifies the last estimation of the RF classifier. This procedure is called bagging [38]. The greater the number of predictors, the more trees that must be planted to improve the good. There are various ways to convert to reading and decorating related individual decision trees, for example, through random feature selection and randomly selected subset of data. While individual decision trees tend to overfitting because they replace them high, RF overcomes this problem by facilitating many decision trees in a random and heterogeneous subset of variables taken.[39]

Decision Tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements [39]. A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label. The paths from root to leaf represent classification rules.

Gain(T,X) = Entropy(T) - Entropy(T,X)

Gain(T,X) = Entropy(T) - Entropy(T,X)

(7)

T = target variable
X = Featuretobespliton
Entropy(T,X) = The entropy calculated after the data is split on feature X

RESULT

Precision Recall and F1-Score (F-Measure)

A number of measures of classification performance can be defined based on the confusion matrix [30] as seen in table 1

Table 1 :number of measures of classification performance

	Predicted Class		
Actual Class		Class = Yes	Class = No
	Class = Yes	TruePositive = TP	False Negatif = FN
	Class = No	FalsePositive = FP	TrueNegative = TN

Precision is a representation of uniformity and repetition of measurements. Precision is the degree of excellence, on the performance of an operation or technique used to get results. Precision measures the level at which the results are close to each other, that is, when measurements are clustered together.

Precision is the ratio of the correctly +ve labelled by our program to all +ve labeled

$$\text{Precision} = \frac{TP}{TP+FP} \tag{8}$$

Recall is the system's success rate in rediscovering information. Furthermore, F-Measure is one of the evaluation calculations in information retrieval that combines recall and precission. The recall value and Precission in a situation can have different weights. The size that displays reciprocity between Recall and Precission is F-Measure which is the mean harmonic weight and reall and precission.

Recall is the ratio of the correctly +velabeleled by our program to all who are diabetic in reality.

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

F-Measure or F1-score is one of the evaluation calculations in information retrieval that combines recall and precission. The recall value and Precission in a situation can have different weights. The size that displays reciprocity between Recall and Precission is F-Measure which is the mean harmonic weight and reall and precission

F1 Score considers both precision and recall. **It is the harmonic mean(average) of the precision and recall.** F1 Score is best if there is some sort of balance between precision (p) & recall (r) in the system. Oppositely F1 Score isn't so good if one measure is increased at the expense of the other.

$$\text{F1 Score} = \frac{2*(Recall*Precision)}{Recall+Precision} \tag{10}$$

Model Evaluation and Validation

Table 2 below is the result of research conducted by Aurelia Bustos and Pertusa, they carried out the Precission, Recall, F-Measure and Kohen's K process [17].

Table 2. Overall results on the validation set for all the classifiers using a dataset of 10^6 samples

Classifier	Evaluation			
	Precision	Recall	F1	Cohen's K
Fast Text	0.88	0.86	0.87	0.75
CNN	0.88	0.88	0.88	0.76
SVM	0.79	0.79	0.79	0.57
kNN	0.92	0.92	0.92	0.83

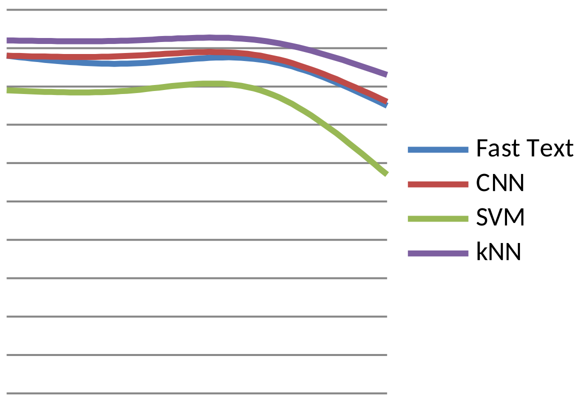

Figure 2. Graph of Overall result the validation set for all the classifiers using a dataset of 10^6 samples

The performance evaluated in this experiment is Recall, Precision, F-Measure and AccuracyTable 3 shows the value of precision, recall and f-measure, after going through the stages of the FGD algorithm, where the highest value of precision, recall and f-measure is found in the KNN, 93, 92 and 93, while the lowest values are in the Multilayer Perceptron, 71, 72 and 72.

Precision shows the accuracy of the classification based on classified documents. In Precision, k-NN outperforms Multilayer perceptron. It can be seen that the Multilayer Perceptron is lowest compared to k-NN with k = 5

The results of the algorithm can be summarized as in the following table :

Table 3. results of Precision Recall and f1-score after using FGC

Evaluation	Classifier				
	MLP	SVM	DT	RF	Knn
Precision	71	78	87	89	93
Recall	72	77	86	88	92
f1-score	72	78	86	88	93

MLP : Multilayer Perceptron
SVM : Support Vector Machine
DT : Decision Tree
RF : Random Forest
KNN : K-Nearest Neighbor

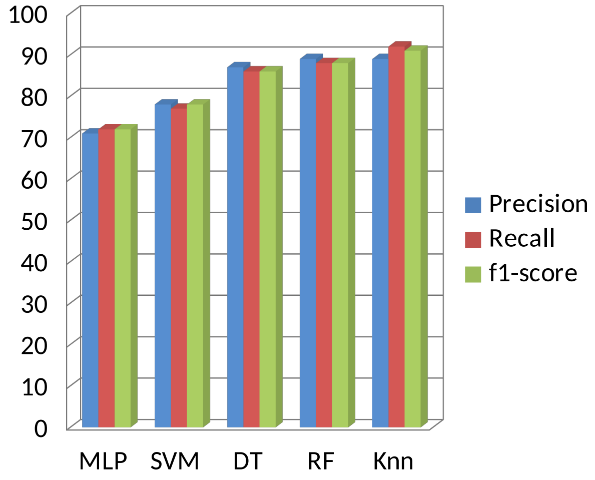

Figure-3. Graph of Result Precision Recall and f1-score

Table 4 shows the results of accuracy of several classifiers as Multilayer Perceptron, Support Vector Machine, Decision Tree, Random Forest, and K-Nearest Neighbor. The author uses 2 to 10-Fold Cross Validation to conduct an evaluation. It can be seen that the highest accuracy is obtained from the Random Forest method which is 90.5, and the lowest value of the Multilayer Perceptron method is 72.1

Table 4. Result of Accuracy process from 2 to 10-fold cross validation

Number of Process	Accuracy %				
	MLP	SVM	DT	RF	KNN
2	71.8	77.3	86.9	90.2	88.5
3	70.1	76.9	85.8	90.1	89.1
4	71.6	77.8	87.3	90.1	90.2
5	73.3	78.1	87.1	89.7	89.7
6	70.2	77.2	85.5	89.9	88.9
7	72.3	77.5	86.3	90.1	89.3
8	69.9	77.1	86.9	90.5	89.9
9	71.4	78.1	85.8	91.4	90.1
10	72.1	77.5	86.2	90.5	89.3

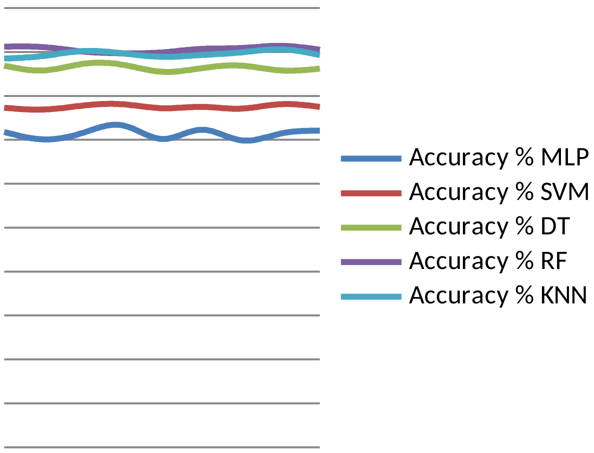

Figure-4. Graph of Accuracy

The evaluation results indicate that cancer-related clinical trial protocols, which are freely available, can be exploited by applying fine grained algorithm, supervised learning, including deep learning techniques, thus opening the potential to explore more ambitious goals by making additional efforts needed to build datasets that corresponding.

CONCLUSIONS

In this study, several classification methods have been trained, validated and compared about the collection of cancer clinical trial protocols (www.clinicaltrials.gov). In this result it can be seen that the value of the recall precision and f-measure have improved slightly, especially in the KNN method after the use of the fine grained document clustering algorithm. As well as accuracy, the highest value is found in the random forest method 90.5, and the lowest value is in the multilayer perceptron method 72.1. So by using the fine grained document clustering method, the computational value of classification can be improved.

Future research is to conduct multilabel classification. The problem will be a multilabel classification task, where classes will be "effective" vs. "ineffective" and "learned" vs "not learned", and both can be true or false. This will allow us to classify four types of cases: effective and studied, potentially effective but not studied, ineffective and learned, and potentially ineffective and not learned. The main effort in this case lies in building a dataset, which includes the efficacy results obtained for each study. New models can be developed to produce potential cancer treatments that can be considered for certain patient cases based on the efficacy of complete clinical trials.

REFERENCES

[1] D. Toruno, E. Çak, M. C. Ganiz, S. Akyoku, and M. Z. Gürbüz, "Analysis of Preprocessing Methods on Classification of Turkish Texts," pp. 112–117, 2011.

[2] A. Y. N. and C. P. Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," no. October, pp. 1631–1642, 2013.

[3] D. Zeng, K. Liu, Y. Chen, and J. Zhao, "Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks," no. September, pp. 1753–1762, 2015.

[4] A. H. Wang, "DON ' T FOLLOW ME Spam Detection in Twitter," 2010.

[5] S. Xie, G. Wang, S. Lin, and P. S. Yu, "Review Spam Detection via Temporal Pattern Discovery," pp. 823–831, 2012.

[6] B. Melinda, "Clinical Trials."

[7] C. Shivade, C. Hebert, M. Lopetegui, M. De Marneffe, E. Fosler-lussier, and A. M. Lai, "Textual inference for eligibility criteria resolution in clinical trials," vol. 58, 2015.

[8] E. Chondrogiannis, V. Andronikou, A. Tagaris, E. Karanastasis, T. Varvarigou, and M. Tsuji, "A novel semantic representation for eligibility criteria in clinical trials," *J. Biomed. Inform.*, vol. 69, pp. 10–23, 2017.

[9] I. S. Samson W. Tu , Mor Peleg , Simona Carini , Michael Bobak , Jessica Ross , Daniel Rubin, "A practical method for transforming free-text eligibility criteria into computable criteria," vol. 44, pp. 239–250, 2011.

[10] B. Mackellar and C. Schweikert, "Analyzing Conflicts between Clinical Trials from a Patient Perspective," pp. 479–482, 2015.

[11] B. Mackellar and C. Schweikert, "Patterns for Conflict Identification in Clinical Trial Eligibility Criteria," 2016.

[12] J. Schmidhuber, "Deep learning in neural networks : An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[13] B. Jou and B. Jou, "From Pixels to Sentiment : Fine-tuning CNNs for Visual Sentiment Prediction NU," 2017.

[14] Ł. Brocki and K. Marasek, "Deep Belief Neural Networks and Bidirectional Long-Short Term Memory Hybrid for Speech Recognition," vol. 40, no. 2, pp. 191–195, 2015.

[15] K. S. Tai, R. Socher, and C. D. Manning, "Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks."

[16] V. Menger, "applied sciences Comparing Deep Learning and Classical Machine Learning Approaches for Predicting Inpatient Violence Incidents from Clinical Text," 2018.

[17] A. Bustos and A. Pertusa, "Learning Eligibility in Cancer Clinical Trials Using Deep Neural Networks," no. 1, pp. 1–19, 2018.

[18] T. Sutanto and R. Nayak, "Fine - grained document clustering via ranking and its application to social media analytics," *Soc. Netw. Anal. Min.*, 2018.

[19] J. Jasmir, S. Nurmaini, R. F. Malik, D. Z. Abidin, A. Zarkasi, and Y. N. Kunang, "Breast Cancer Classification Using Deep Learning," *2018 Int. Conf. Electr. Eng. Comput. Sci.*, vol. 17, pp. 237–242, 2018.

[20] A. M. Abdel-zaher and A. M. Eldeib, "Breast Cancer Classification Using Deep Belief Networks," *Expert Syst. Appl.*, 2015.

[21] V. Garla, C. Taylor, and C. Brandt, “Semi-supervised clinical text classification with Laplacian SVMs : An application to cancer case management,” *J. Biomed. Inform.*, vol. 46, no. 5, pp. 869–875, 2013.

[22] R. L. Figueroa, Q. Zeng-treitler, L. H. Ngo, S. Goryachev, and E. P. Wiechmann, “Active learning for clinical text classification : is it better than random sampling ?,” pp. 809–816, 2012.

[23] B. J. Mara, J. M. Davies, N. S. Bardach, M. L. Dean, and R. A. Dudley, “N-gram support vector machines for scalable procedure and diagnosis classi fi cation , with applications to clinical free text data from the intensive care unit,” pp. 871–875, 2014.

[24] J. J. Cimino, W. J. Lancaster, and M. C. Wyatt, “Classification of Clinical Research Study Eligibility Criteria to Support Multi-Stage Cohort Identification Using Clinical Data Repositories,” vol. 0, 2017.

[25] K. Zhang and D. Demner-fushman, “Automated classification of eligibility criteria in clinical trials to facilitate patient-trial matching for specific patient populations,” vol. 0, no. June 2016, pp. 1–7, 2017.

[26] C. Chuan, “Classifying Eligibility Criteria in Clinical Trials Using Active Deep Learning,” *2018 17th IEEE Int. Conf. Mach. Learn. Appl.*, pp. 305–310, 2018.

[27] I. K. and I. S. Yizhao Ni, Jordan Wright, John Perentesis, Todd Lingren, Louise Deleger, Megan Kaiser, “Increasing the efficiency of trial-patient matching : automated clinical trial eligibility Pre-screening for pediatric oncology patients,” pp. 1–10, 2015.

[28] N. L. of Medicine, “National Library of Medicine, National Institutes of Health. XML Schema for ClinicalTrials.gov Public XML; National Library of Medicine, National Institutes of Health: Bethesda, MD, USA, 2017.,” p. 2017, 2017.

[29] C. Liu, Y. Sheng, Z. Wei, and Y. Yang, “Research of Text Classification Based on Improved TF-IDF Algorithm,” *2018 IEEE Int. Conf. Intell. Robot. Control Eng.*, no. 2, pp. 218–222, 2018.

[30] A. Darmawahyuni and S. Nurmaini, “Coronary Heart Disease Interpretation Based on Deep Neural Network,” *Comput. Eng. Appl.*, vol. 8, no. 1, 2019.

[31] S. Nurmaini, R. U. Partan, M. N. Rachmatullah, and A. Gani, “Cardiac Arrhythmias Classification Using Deep Neural Networks and Principle Component Analysis Algorithm,” *Int. J. Adv. Soft Comput. Its Appl.*, vol. 10, no. 2, 2018.

[32] S. Nurmaini, R. U. Partan, W. Caesarendra, and T. Dewi, “An Automated ECG Beat Classification System Using Deep Neural Networks with an Unsupervised Feature Extraction Technique,” *Appl. Sci.*, vol. 9, 2019.

[33] N. Fuhr, M. Lechtenfeld, B. Stein, and T. Gollub, “The optimum clustering framework : implementing the cluster hypothesis,” 2011.

[34] D. Isa, L. H. Lee, V. P. Kallimani, and R. Rajkumar, “Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine,” vol. 20, no. 9, pp. 1264–1272, 2008.

[35] C. Chatterjee and V. Roychowdhury, “Statistical Risk Analysis for Classification and Feature Extraction by Multilayer,” pp. 1610–1615.

[36] B. Ramesh and J. G. R. Sathiaseelan, “An Advanced Multi Class Instance Selection based Support Vector Machine for Text Classification,” *Procedia - Procedia Comput. Sci.*, vol. 57, pp. 1124–1130, 2015.

[37] N. L. Husni, A. S. Handayani, S. Nurmaini, and I. Yani, “Odor Classification

Using Support Vector Machine," *Int. Conf. Electr. Eng. Comput. Sci.*, pp. 71–76, 2017.

[38] L. E. O. Breiman, "Random Forests," pp. 5–32, 2001.

[39] J. B. J. Albert, E. Aliu, H. Anderhub, P. Antoranz, A. Armada, M. Asensio, C. Baixeras, J.A. Barrio, H. Bartko, D. Bastieri, "Implementation of the Random Forest method for the Imaging Atmospheric Cherenkov Telescope MAGIC," vol. 588, pp. 424–432, 2008.