



***COMPARATIVE STUDY OF NAÏVE BAYES AND ARTIFICIAL NEURAL  
NETWORK TO DETECT PATIENTS' DISEASE TYPES BY USING  
STRUCTURAL AND UNSTRUCTURAL DATA***

*TUGAS AKHIR*

Ibrohim Imam Thohari  
41518120090

**PROGRAM STUDI TEKNIK INFORMATIKA  
FAKULTAS ILMU KOMPUTER  
UNIVERSITAS MERCU BUANA  
JAKARTA  
2021**



***COMPARATIVE STUDY OF NAÏVE BAYES AND ARTIFICIAL NEURAL  
NETWORK TO DETECT PATIENTS' DISEASE TYPES BY USING  
STRUCTURAL AND UNSTRUCTURAL DATA***

*Tugas Akhir*

Diajukan Untuk Melengkapi Salah Satu Syarat  
Memperoleh Gelar Sarjana Komputer

Oleh:  
Ibrohim Imam Thohari  
41518120090

PROGRAM STUDI TEKNIK INFORMATIKA  
FAKULTAS ILMU KOMPUTER  
UNIVERSITAS MERCU BUANA  
JAKARTA  
2021

## LEMBAR PERNYATAAN ORISINALITAS

Yang bertanda tangan dibawah ini:

NIM : 41518120090

Nama : Ibrohim Imam Thohari

Judul Tugas Akhir : *Comparative Study Naïve Bayes dan Artificial Neural Network* untuk Mendeteksi Jenis Penyakit Pasien dengan Menggunakan *Structural Data* dan *Unstructural Data*

Menyatakan bahwa Laporan Tugas Akhir saya adalah hasil karya sendiri dan bukan plagiat. Apabila ternyata ditemukan di dalam laporan Tugas Akhir saya terdapat unsur plagiat, maka saya siap untuk mendapatkan sanksi akademik yang terkait dengan hal tersebut.

Jakarta, 09 Februari 2021



Ibrohim Imam Thohari

UNIVERSITAS  
MERCU BUANA

## SURAT PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR

Sebagai mahasiswa Universitas Mercu Buana, saya yang bertanda tangan di bawah ini :

Nama Mahasiswa : Ibrohim Imam Thohari  
NIM : 41518120090  
Judul Tugas Akhir : *Comparative Study Naïve Bayes dan Artificial Neural Network untuk Mendeteksi Jenis Penyakit Pasien dengan Menggunakan Structural Data dan Unstructural Data*

Dengan ini memberikan izin dan menyetujui untuk memberikan kepada Universitas Mercu Buana **Hak Bebas Royalti Non eksklusif** (*Non-exclusive Royalty Free Right*) atas karya ilmiah saya yang berjudul di atas beserta perangkat yang ada (jika diperlukan).

Dengan Hak Bebas Royalti/Non eksklusif ini Universitas Mercu Buana berhak menyimpan, mengalih media/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat dan mempublikasikan tugas akhir saya.

Selain itu, demi pengembangan ilmu pengetahuan di lingkungan Universitas Mercu Buana, saya memberikan izin kepada Peneliti di Lab Riset Fakultas Ilmu Komputer, Universitas Mercu Buana untuk menggunakan dan mengembangkan hasil riset yang ada dalam tugas akhir untuk kepentingan riset dan publikasi selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 09 Februari 2021

UNIVERSITAS  
MERCU BUANA



1000  
METERAI  
TEMPEL  
42FF3AJX016176742

Ibrohim Imam Thohari



## SURAT PERNYATAAN LUARAN TUGAS AKHIR

Sebagai mahasiswa Universitas Mercu Buana, saya yang bertanda tangan di bawah ini :

Nama Mahasiswa : Ibrohim Imam Thohari  
 NIM : 41518120090  
 Judul Tugas Akhir : Comparative Study of Naïve Bayes and Artificial Neural Network to Detect Patients' Disease Types by Using Structural and Unstructural Data

Menyatakan bahwa :

1. Luaran Tugas Akhir saya adalah sebagai berikut :

No	Luaran	Jenis	Status
1	Publikasi Ilmiah	Jurnal Nasional Tidak Terakreditasi	Diajukan ✓
		Jurnal Nasional Terakreditasi	
		Jurnal International Tidak Bereputasi	Diterima
		Jurnal International Bereputasi ✓	
Disubmit/dipublikasikan di :	Nama Jurnal	: Computer and Information Science	
	ISSN	: 0258-2724	
	Link Jurnal	:	
	Link File Jurnal Jika Sudah di Publish	:	

2. Bersedia untuk menyelesaikan seluruh proses publikasi artikel mulai dari submit, revisi artikel sampai dengan dinyatakan dapat diterbitkan pada jurnal yang dituju.
3. Diminta untuk melampirkan scan KTP dan Surat Pernyataan (Lihat Lampiran Dokumen HKI), untuk kepentingan pendaftaran HKI apabila diperlukan

Demikian pernyataan ini saya buat dengan sebenarnya.

Mengetahui,  
Dosen Pembimbing TA



Dr. Mujiono Sadikin, MT, CISA, CGEIT

Jakarta, 09 Februari 2021




Ibrohim Imam Thohari

## LEMBAR PERSETUJUAN PENGUJI

NIM : 41518120090  
Nama : Ibrohim Imam Thohari  
Judul Tugas Akhir : *Comparative Study of Naïve Bayes and Artificial Neural Network to Detect Patients' Disease Types by Using Structural and Unstructural Data*

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.


Jakarta, 09 Februari 2021



(Desi Ramayanti, S.Kom., MT)  
Ketua Penguji



(Muhammad Rifqi, S.Kom, M.Kom)  
Anggota Penguji 1



(Dwiki Jatikusumo, S.Kom, M.Kom)  
Anggota Penguji 2

## LEMBAR PENGESAHAN

NIM : 41518120090  
Nama : Ibrohim Imam Thohari  
Judul Tugas Akhir : *Comparative Study of Naïve Bayes and Artificial Neural Network to Detect Patients' Disease Types by Using Structural and Unstructural Data*

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 09 Februari 2021

Menyetujui,



UNIVERSITAS

(Dr. Mujiono Sadikin, MT, CISA, CGEIT)

Dosen Pembimbing

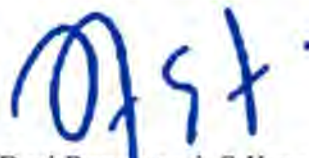
MERCU BUANA

Mengetahui,



(Diky Firdaus, S.Kom, MM)

Koord. Tugas Akhir Teknik Informatika



(Desi Ramayanti, S.Kom, MT)

Ka. Prodi Teknik Informatika

## ABSTRAK

Nama : Ibrohim Imam Thohari  
NIM : 41518120090  
Pembimbing TA : Dr. Mujiono Sadikin, MT. CISA, CGEIT.  
Judul : *Comparative Study of Naïve Bayes and Artificial Neural Network to Detect Patients' Disease Types by Using Structural and Unstructural Data*

Konsep utama dari rumah sakit adalah penyediaan layanan kesehatan kepada masyarakat. Untuk memastikan pelayanan kesehatan bisa maksimal, rumah sakit memanfaatkan teknologi untuk merekam semua data kegiatan yang ada di rumah sakit. Namun, saat ini data tersebut hanya disimpan di *database* dan digunakan sebagai *history* tanpa digunakan lebih lanjut. Banyak pengalaman yang menunjukkan bahwa dengan mengoptimalkan penggunaan data akan sangat membantu dokter dalam mengambil keputusan untuk meminimalisir kesalahan medis. Misalnya data pemeriksaan yang antara lain anamnesis (abstrak medis), tekanan darah, suhu, dll dapat digunakan untuk klasifikasi jenis penyakit. Makalah ini memaparkan hasil studi penggunaan *comparative study* antara Algoritma *Naïve Bayes*, dan *Artificial Neural Network* untuk mengkaji data dalam pengklasifikasian jenis penyakit berdasarkan data pemeriksaan dokter yang terstruktur dan tidak terstruktur. *Natural Language Processing* digunakan untuk merepresentasikan *unstructured text medical abstract* ke dalam bentuk *vector* menggunakan *Word2vec word embedding*. Dengan begitu, *medical abstract* dan data pemeriksaan lainnya dapat diolah dengan menggunakan algoritma *Naïve Bayes* dan *Artificial Neural Network*. Dengan menggunakan kedua algoritma tersebut maka didapatkan hasil klasifikasi jenis penyakit. Eksperimen yang dilakukan menunjukkan bahwa model *ANN* memberikan kinerja yang lebih baik dengan rata-rata akurasi terbaik sebesar 91,46% dibandingkan dengan *Naïve Bayes* yaitu 68,33%. Selain itu, keterlibatan *unstructured text* dari dataset dalam proses pelatihan *word2vec* dapat meningkatkan kinerja *ANN* meskipun tidak signifikan yaitu rata-rata akurasi 90.27% dibandingkan tanpa melibatkan *unstructured text data* yaitu 89.82%.

Kata kunci:

*Natural Language Processing, Word2vec, Medical abstract, Naïve Bayes, Artificial Neural Network.*



## ABSTRACT

Name : Ibrohim Imam Thohari  
Student Number : 41518120090  
Counsellor : Dr. Mujiono Sadikin, MT. CISA, CGEIT.  
Title : Comparative Study of Naïve Bayes and Artificial Neural Network to Detect Patients' Disease Types by Using Structural and Unstructural Data

The main concept of the hospital is the provision of health services to the community. To ensure that healthy services can be maximized, the hospital uses technology to record all activity data in the hospital. However, currently the data is only stored in the database and used as history without any further use. Many experiences show that by optimizing the data usage it will greatly assist doctors in making decisions to minimize medical errors. For example, examination data that among others of medical abstract, blood pressure, temperature, etc. can be used for the classification of the kind of disease. This paper presents the result study of the using a comparative study of naïve bayes and artificial neural network to examine the data in classifying the kind of disease based on structural and unstructural examination data. Natural language processing is used to represent unstructured text medical abstracts in a vector form using Word2Vec word embedding. That way, medical abstract and other examination data can be processed using the Naïve Bayes algorithm and the Artificial Neural Network. By using these two algorithms, the results of the classification of the kind of disease. The performed experiments show that ANN model gives the better performance with the best accuracy average of 91.46% compared to Naive Bayes which is 68.33%. In addition, the involvement of unstructured data from the dataset in the word2vec training process improves the performance ANN even though it is not significant with an accuracy average of 90.27% compared without the involvement of unstructured data which is 89.82%.

Key words:

Natural Language Processing, Word2Vec, Medical abstract, Naïve bayes, Artificial Neural Network.

## KATA PENGANTAR

Puji syukur penulis panjatkan kepada Tuhan Yang Maha Esa yang telah melimpahkan rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan tugas akhir yang berjudul “*Comparative Study of Naïve Bayes and Artificial Neural Network to Detect Patients’ Disease Types by Using Structural and Unstructural Data*” tepat pada waktunya. Tugas akhir ini disusun untuk memenuhi salah satu syarat memperoleh gelar sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer Universitas Mercu Buana.

Penulis menyadari bahwa tanpa bantuan dan bimbingan yang melibatkan banyak pihak, penelitian ini tidak akan terlaksana dengan baik. Oleh karena itu, penulis mengucapkan terima kasih kepada:

1. Bapak Dr. Mujiono Sadikin, MT. CISA, CGEIT. selaku dosen pembimbing Tugas Akhir yang telah meluangkan waktunya untuk memberikan bimbingan serta arahan dalam penyusunan tugas akhir ini hingga selesai.
2. Ibu Umny Salamah, ST selaku dosen pembimbing akademik yang telah membantu persyaratan dalam penyusunan tugas akhir ini.
3. Ibu Desi Ramayanti, S.Kom., MT. selaku Kepala Prodi Teknik Informatika, Fakultas Ilmu Komputer Universitas Mercu Buana.
4. Bapak Diky Firdaus, S.Kom., MM. selaku Koordinator Tugas Akhir Program Studi Teknik Informatika, Fakultas Ilmu Komputer Universitas Mercu Buana.
5. Orang tua yang senantiasa memberikan doa dan dukungan kepada penulis.
6. Teman-teman yang selalu memberikan semangat kepada penulis selama pelaksanaan tugas akhir.
7. Semua pihak yang tidak dapat penulis sebutkan satu persatu yang telah banyak membantu dalam penyusunan tugas akhir.

Akhir kata, penulis berharap tugas akhir ini dapat bermanfaat bagi pembaca guna menambah pengetahuan dan wawasan.

Jakarta, 09 Februari 2021  
Penulis

## DAFTAR ISI

HALAMAN SAMPUL.....	i
HALAMAN JUDUL.....	i
LEMBAR PERNYATAAN ORISINALITAS.....	ii
SURAT PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR..	iii
SURAT PERNYATAAN LUARAN TUGAS AKHIR.....	iv
LEMBAR PERSETUJUAN PENGUJI.....	v
LEMBAR PENGESAHAN.....	vi
ABSTRAK.....	vii
ABSTRACT.....	viii
KATA PENGANTAR.....	ix
DAFTAR ISI.....	x
NASKAH JURNAL.....	1
KERTAS KERJA.....	14
BAGIAN 1. LITERATUR <i>REVIEW</i> .....	16
BAGIAN 2 ANALISIS DAN PERANCANGAN.....	31
BAGIAN 3 <i>SOURCE CODE</i> .....	33
BAGIAN 4 DATASET.....	67
BAGIAN 5 TAHAPAN EKSPERIMEN.....	70
BAGIAN 6 HASIL SEMUA EKSPERIMEN.....	74
DAFTAR PUSTAKA.....	82
LAMPIRAN DOKUMEN HAKI.....	84
LAMPIRAN CV.....	86

ISSN: 0258-2724

DOI : 10.35741/issn.0258-2724.54.6....

Research article

Computer and Information Science

**COMPARATIVE STUDY OF NAÏVE BAYES AND ARTIFICIAL NEURAL NETWORK TO DETECT PATIENTS' DISEASE TYPES BY USING STRUCTURAL AND UNSTRUCTURAL DATA**Ibrohim Imam Thohari <sup>a,\*</sup>, Mujiono Sadikin <sup>a,b</sup><sup>a</sup>Faculty of Computer Science, Mercu Buana University,  
Kembangan, Jakarta, Indonesia, 41518120090@student.mercubuana.ac.id<sup>b</sup>Faculty of Computer Science, Mercu Buana University,  
Kembangan, Jakarta, Indonesia, mujiono.sadikin@mercubuana.ac.id**Abstract**

The main concept of the hospital is the provision of health services to the community. To ensure that healthy services can be maximized, the hospital uses technology to record all activity data in the hospital. However, currently the data is only stored in the database and used as history without any further use. Many experiences show that by optimizing the data usage it will greatly assist doctors in making decisions to minimize medical errors. For example, examination data that among others of medical abstract, blood pressure, temperature, etc. can be used for the classification of the kind of disease. This paper presents the result study of the using a Comparative Study between Naïve Bayes algorithm and Artificial Neural Network to examine the data in classifying the kind of disease based on the structural and unstructural examination data. Natural language processing is used to represent unstructured text medical abstracts in a vector form using Word2Vec word embedding. That way, medical abstract and other examination data can be processed using the Naïve Bayes algorithm and the Artificial Neural Network. By using these two algorithms, the results of the classification of the kind of disease. The performed experiments show that ANN model gives the better performance with the best accuracy average of 91.46% compared to Naive Bayes which is 68.33%. In addition, the involvement of unstructured data from the dataset in the word2vec training process improves the performance ANN even though it is not significant with an accuracy average of 90.27% compared without the involvement of unstructured data which is 89.82%.

**Keywords:** Natural Language Processing, Word2Vec, Medical abstract, Naïve bayes, Artificial Neural Network.**摘要****关键词:**

## I. INTRODUCTION

The hospital is an important agency in providing health services to the community. In providing services to the community, the hospital must always improve quality of services to increase public satisfaction with hospital services. Starting from services for emergency services, outpatient care, inpatient care, doctor consultation, drug administration, to doctor reliability must be maximized.

In this era of technology, data plays an important role in a hospital to provide services to the community. The hospital records all activity data in the hospital, starting from the medical abstract, examination data, doctor's actions, disease diagnosis, type of disease, up to prescribing drugs. However, currently all data has not been fully utilized. Data is only stored in the database and used as history without any further data utilization. In fact, if all data can be processed properly, it will greatly assist doctors in making decisions to minimize medical errors [1].

Data Mining is one of the stages of Knowledge Discovery in Databases (KDD). The steps in KDD include: data cleaning, data integration, data selection, data transformation, data mining, patterns evaluation and knowledge presentation [2]. Data Mining can be used as a Decision Support System for doctors. By using certain methods can provide drug recommendations from medical record data [3]. Automatic recommendation will increase the doctor's awareness in making decisions [4].

Anamnesis data is one of the results of examination data performed by doctors on patients in the form unstructured text medical abstract. To process medical abstract data it is necessary to do text mining using Natural-Language Processing (NLP) [5][6]. With Natural-Language Processing (NLP) medical abstract data can be processed in data mining [7][8]. By using the algorithm Naïve Bayes and Artificial Neural Network, we can get the type of disease classification suffered by a patient based on the doctor's examination.

In this study, we compare Naïve Bayes algorithm and Artificial Neural Network. Generally speaking, comparative analysis allows us to performs several important

functions that are closely interlinked which in this case study are the Naive Bayes algorithms and Artificial Neural Network. In addition, comparisons allow us to test certain theories and to evaluate the significance of certain phenomena, thus contributing to the development of universally accepted theories [9]. We compared both algorithms combined with Natural Language Processing were used in data processing to obtain a type of disease classification suffered by patients based on the results of a doctor's examination. There are two types of disease, namely "acute" and "chronic". Disease is called acute if it is temporary and can be cured after receiving treatment, while chronic illness is a long disease, recurring, requires a long and well-organized treatment process, and needs the ability to limit a person's lifestyle [10]. There are three main stages carried out in this research, namely: 1) Cleaning data which includes removing the noise from dataset and filling in the blank values; 2) Representation of medical abstracts into vector form by Natural Language Processing model training data using Word2Vec word embedding method [11] from the gensim python library [12]; 3) Classification using Naïve Bayes algorithm which is a probabilistic model and an Artificial Neural Network and tests the accuracy of each algorithm.

The final result of this research is a classification of disease types as a recommendation that can increase the awareness of doctors in making decisions and minimize potential medical errors. In this study, we also want to know whether there is the involvement of unstructured data from the dataset in word2vec training affecting the accuracy results. Thus, the experiment in this study consists of two main scenarios.

## II. RESEARCH METHOD

In this research, the classification technique Naïve Bayes and Artificial Neural Network were used by using the programming language python to run the algorithm. Broadly speaking, the research stages include data collection, training the model Word2Vec, preprocessing data, representing data medical abstract into a vector form length N, data testing, and implementing the



Naïve Bayes algorithm and Artificial Neural Network. Data collection stage, obtained in the dataset was the patient examination form of Medical Abstract, systolic blood pressure, blood pressure diastolic, temperature, pulse, age, gender, giddiness, and type of disease. The Stages training data of Word2Vec resulted in a NLP model used to convert Medical Abstract data. The stages of representing the medical abstract, the selected keywords are generated into vectors form. At the algorithm implementation stage, results of testing were obtained for the Naïve Bayes algorithm and Artificial Neural Network.

The process of representing medical abstracts into vector data requires the Word2Vec model to mapping keywords into vector data. However, some of the keywords of medical abstracts are not found in the Word2Vec model. It is happening because there is an inconsistent input of the medical abstract when the doctor takes the examination. To

avoid keywords missing, the study was conducted using two scenarios for training the model Word2Vec. Figure 1 shows a diagram of the research stages in the first scenario. The process training model The Word2Vec’s first scenario only uses the corpus Wikipedia as a dataset. However, if the medical abstract keyword is not found in the Word2Vec model, update carried out to the Word2Vec model by adding the keyword to the model and then retrain the model. That way the model can accommodate inconsistencies of medical abstract input. Figure 2 shows a diagram of the research stages in the second scenario. The process Word2Vec model training second scenario does not only use the corpus Wikipedia as a dataset. But the corpus Wikipedia is combined with a medical abstract as a dataset. In this way, the model Word2Vec accommodates inconsistencies of medical abstract input.

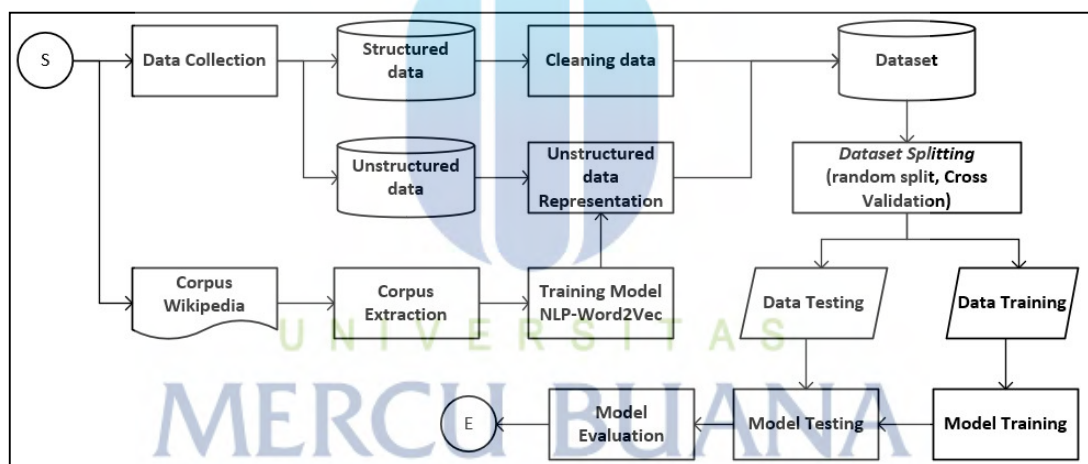


Figure 1. Research Stages The First Scenario

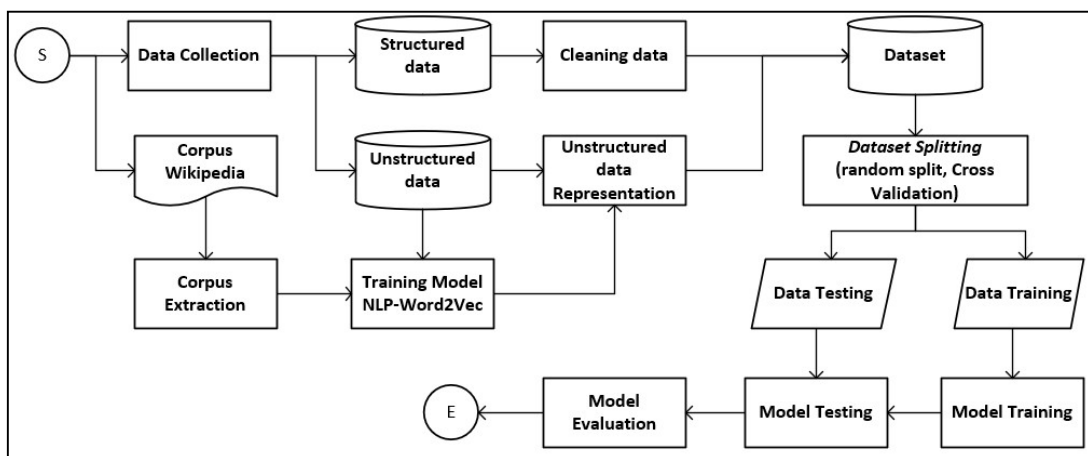


Figure 2 Research Stages The Second Scenario

**Data Collection**

Data collection stages are carried out by conducting observations and interviews. From the results of observations and interviews, it was found that there was no use of doctor's examination data to serve as a Decision Support System in the form of a classification of disease types. Data is only stored and used as patient history so it does not provide added value in the form of more useful information. Table 1 shows the details structure of the 4,404 records examination data collected. Then the examination data is separated between unstructured data and structured data because the treatment of these two types of data is different. Unstructured data includes anamnesis while the data structured include blood systole pressure, blood diastolic pressure, temperature, pulse, age, gender, unsteady condition, and type of disease.

Table 1:  
Structure of the Patient Examination Dataset

Attribute	Data type	Value Range
Anamnesis	Text	Text
Calm	Varchar	Ordinal
Anxious	Varchar	Ordinal
Age	Integer	Continue
Sex	Varchar	Ordinal
Systol	Integer	Continue
Dyastol	Integer	Continue
Temperature	Numeric	Continue
Pulse	Integer	Continue
Weight	Integer	Continue
Height	Integer	Continue
Unsteady condition	Varchar	Ordinal
Risk of falling	varchar	Ordinal
Type of disease	varchar	Ordinal

Description Table 1.

- Anamnesis : Medical abstract in the form of text information from the patient during the examination.
- Continue range value: value that is numerical value
- Ordinal range value: scale that differentiates categories by level / order.
- Range values of Calm, Anxious, Unsteady condition are Y and N.
- Range values of Sex are L and P.
- Range values of Risk of falling are “No risk”, “Low Risk” and “High Risk”.

- Range values of disease types are “Chronic” and “Acute”.

### Cleaning data

Cleaning data is used to remove noise from the dataset. Data that have empty values are re-evaluated to determine whether the data is suitable for use or not. Data that has an empty value within a certain tolerance can still be used. By using a common technique, the treatment of swapping the blank value with the average value recognized from the variable value [13]. This average value is used as a constant to replace the empty values variable dataset regardless of the relationship between properties that affect the Data Mining algorithm used. At this stage the data being cleaned is structured data which includes the variables Calm, Anxious, Age, Sex, Systol, Dyastol, Temperature, Pulse, Weight, Height, Unsteady condition, Risk of falling, Type of disease.

### Conversion of Text Data

Converting Medical Abstract data into data is a vector intended for the data can be processed in Data Mining. The method used in this process is Word2Vec word embedding. Word2Vec is able to understand the meaning and turn it into a vector of words in the document based on the hypothesis that words that have similar meanings have a proximity vector [14]. In the method, Word2Vec there are two models, namely the Continuous Bag-of-Word (CBOW) and Skip-Gram. Each model has several layers, namely the input layer, projection layer, and output layer. Architecturally, the two models have architectures that are opposite to each other. Skip-gram is used to predict the context of a word as input. While CBOW is used to predict words from the surrounding context as input [15] [16]. Figure 3 is the architecture CBOW and Skip-gram proposed by Milokov.

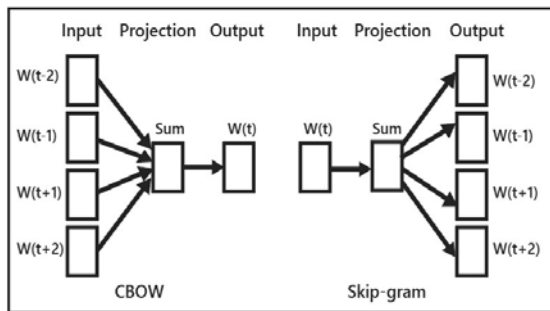


Figure 3. CBOW and Skip-Gram Architecture

The process of representing medical abstract data into the data vector using the method word embedding requires a word2vec model training data. To make the model required a large amount of text in Indonesian. Corpus Wikipedia is a collection of articles from Wikipedia that are used as a dataset for training the model Word2Vec [17]. In this study, the Corpus Wikipedia used several 408,952 articles in Indonesian which were extracted using wiki corpus. The Extraction Process of Wikipedia Corpus is used to convert the Wikipedia corpus into text. Corpus Wikipedia is processed using simply preprocessing then converted into text using wiki corpus.

The next step is to create a model Word2Vec that is used to map the semantic proximity positions between words from an input text [18]. In this research, the making of the model was Word2Vec carried out with two scenarios. Both of these scenarios were carried out because there was an inconsistency of medical abstract input which resulted in the keyword not being found in the model Word2Vec.

**First Scenario:** In the first scenario, we only use the corpus Wikipedia to training the Word2Vec model. However, if the keyword medical abstract is not found in the Word2Vec model, update the model Word2Vec performed by adding keywords to the model and then retraining the model. Models updates are performed while the representation process is medical abstract running. So if the gensim does not find the keyword in the model, the representation processing time will be longer because it requires additional time for model training.

**Second Scenario:** In the second scenario of the Word2Vec model training process, we do not only use the corpus Wikipedia as a

dataset. But we combine the corpus Wikipedia with a medical abstract as a dataset. Medical abstracts are treated with simply preprocessing to eliminate the character of numbers and symbols. Then the medical abstract is added to the Wikipedia corpus for training the model Word2Vec. In this way, the models created already accommodate input medical abstract inconsistencies.

In the representation process for each scenario, we performed cleaning text for the Medical Abstract data. The process of cleaning text data uses a library of Indonesian Natural Language Processing in python. The cleaning process for the Medical Abstract data text includes Lemmatization, Removing number, stopword removal, and Pos tagger.

**Step 1:** We use lemmatization to transform words into the root of words. Lemmatization changes the word by considering the context of the word, which means that it is not just removing some characters in a word. So that the resulting word is more accurate and has meaning.

**Step 2:** The next step is removing numbers, which aim to remove numbers from text data. Numbers are omitted because they don't have much effect on the root word.

**Step 3:** Stopword removal is used to eliminate common words that are considered meaningless. Examples of stopwords in Indonesian are "yang", "dan", "di", "dari", etc. That way the process can be focused on words that are considered important.

**Step 4:** POS Tagging is used to get Part-Of-Speech tags from text which is useful for categorizing word classes. Then the text is separated based on the word class category.

After the text data has been cleaned, it can be represented as vector data using two experimental scenarios of the word2vec model. Table 2 shows the representation of medical abstract data into vector data.

Table 2:  
Vector Data Representation

Teks	V1	V2	....	V25
batuk	-0.0045260135	0.01768395	....	0.00079621444
dahak	0.017375236	0.005819824	....	-0.014331724
nafas	0.011860242	0.014599305	....	-0.015677009
demam	0.0043433174	0.0083342595	....	-0.0074927625

Medical abstracts that have been represented as vectors and structured data that have been cleaned are combined back into a dataset. The dataset is used to implement data mining. In this research, the algorithm used is the Naïve Bayes algorithm and the Artificial Neural Network algorithm. In each algorithm, there are three stages, namely training, testing, and evaluating.

### Dataset Splitting

To carry out the training and testing process, needed to split data is to separate training data and testing data from existing datasets. In this research, using two split data methods namely Random Split and Kfold Cross Validation. Random split divides the dataset randomly into training data and testing data with certain comparisons. Kfold Cross Validation is used to split the data into K parts of the dataset to the same size in order to eliminate bias in the data [19]. The training and testing process is carried out as much as the specified K [20]. Figure 4 shows the iteration process in the K Fold Cross-validation method.

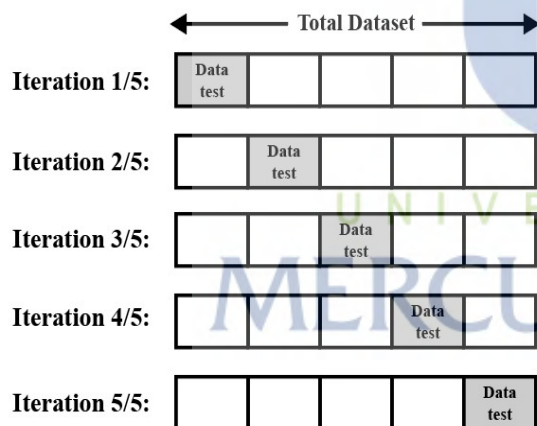


Figure 4 Cross-Validation with 5 kfold value

In this study, two methods were split data carried out with the same comparison for each algorithm and the Word2Vec model training scenario. From a total of 3108 rows of data, three comparisons of random split were made between the training data : testing data, namely 9:1, 8:2, and 7:3. Whereas, in Cross-Validation Kfold there are 3 grades K is 5 K, 10 K, and 15 K.

### Naïve Bayes Algorithm

Naive Bayes algorithm is an algorithm that uses probabilistic and statistical models

invented by the British scientist Thomas Bayes [21]. Classifications are carried out by predicting future opportunities based on past experiences. This algorithm aims to predict a class based on the training data provided [22]. The main element of Naïve Bayes Classifier is about three aspects, they are prior, posterior dan class conditional probability [23].

$$P(y|X) = \frac{P(X|y) \cdot P(y)}{P(X)}$$

Description:

- X : Data with unknown class
- y : Hypothesis class data X is a specific class
- P(y|X) : Hypothesis probability y based on condition X
- P(y) : Hypothesis probability y
- P(X|y) : Probability X based on these conditions
- P(X) : Probability of X

Implementation of the algorithm is Naïve Bayes carried out on patient examination dataset. Implementation is done by creating a model Naïve Bayes using training data. The resulting model is a pattern formed from training data. The model is used to predict the testing data, then the model performance is calculated in the form of prediction accuracy of testing data.

### Artificial Neural Network

This algorithm is a deep learning algorithm inspired by the human brain network and implemented into a computer program [24]. The Artificial Neural Network is running based on a reasoning model of the human brain which is able to complete a number of calculation processes during the learning process . Like the human brain network, an Artificial Neural Network has neurons consisting of a number of simple interconnected processors. Neurons are connected by weight pass signals from neuron to another neuron. Hidden layers and output layers have an additional input called bias [25]. Figure 5 is an example of an architecture of Artificial Neural Network.



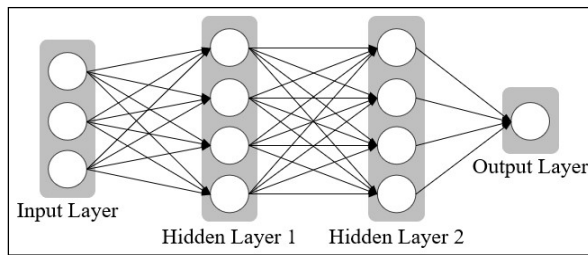


Figure 5. Architecture of ANN

The architecture above is commonly referred as Multi Layer Perceptron (MLP) or Fully-Connected Layer. The architecture above has four layers, that is input layer with 3 neurons, hidden layer 1 with 4 neurons, hidden layer 2 with 4 neurons, and an output layer with 1 output node. So that in the architecture above there are 3x4 weight + 4 bias, 4x4 weight + 4 bias, and 4x1 weight + 1 bias with a total of 41 parameters used.

Implementation of Artificial Neural Network carried out on a dataset for the patient examination. The implementation process is carried out using the python programming language by utilizing the Keras library to create models. In this study, the architectural model made there are 110 nodes on the input layer. Meanwhile, the number of layers created was 4 hidden layers with the activation of ReLU (Rectified Linear Unit activation function) and 1 output layer with the activation of the Sigmoid function.

The next stage is compiled which aims to compile the Keras model so that it is ready for the model training process and to obtain the most efficient computation. At this stage, there are variables that are set during the compile process. variable Optimizer uses "adam" which is quite popular and is often used as an optimization method. The Variable loss used is "binary\_crossentropy". While the matrix variable measured at this stage is the accuracy value as the measurement value.

The next stage is Fit Keras Model. This stage is a model execution of the dataset using the fit() function on the model. At this stage, the training is model carried out as many as 150 epochs (period) and each epoch there are 10 groups (batch\_size). After all the stages have been completed, the last stage is the Hard Evaluate Model. Evaluation is done using the function evaluate() and returns the value accuracy from the model in the dataset.

### III. RESULTS AND DISCUSSION

In the implementation of this research, we compare Naïve Bayes algorithm and Artificial Neural Network combined with Natural Language Processing. Natural Language Processing is used to represent Medical Abstracts as a vector form so that Medical Abstracts which are originally unstructured text data can be processed with data mining algorithms. The Naïve Bayes Algorithm and Artificial Neural Network are used to classify types of diseases as well as test the accuracy of each algorithm.

The process of representing medical abstracts into vector data using the Word2Vec model size 25 vectors. We built the model using the gensim library with a training dataset of 408,952 Indonesian language articles from the Corpus Wikipedia extracted using wiki corpus. The modeling was carried out in two different scenarios. In the first scenario, the medical abstract representation uses the Word2Vec model trained only with the Wikipedia corpus dataset. Whereas in the second scenario the medical abstract representation uses the Word2Vec model trained with the Wikipedia corpus dataset plus medical abstracts from patient examination data. So this experiment uses two different datasets, namely the first scenario dataset and the second scenario. Then we tested the two datasets with the Naïve Bayes algorithm and Artificial Neural Network, each of which used the random split and kfold cross-validation methods.

In this study, we took four medical abstract keywords which were represented as vector data. Keywords are represented using two different model scenarios that have been made. Each model scenario size is 25 data vectors semantic proximity position keyword input. So that from one record Medical Abstract obtained 100 vectors data for each scenario. Table 3 shows the results of a medical abstract representation into vector form.

Table 3:  
Word2Vec Representation Results

w1v1	w1v2	w1v3	w1v4	.....	w4v25
-8.26756	1.569373	1.973389	2.199834	.....	-0.0198
-8.26756	1.569373	1.973389	2.199834	.....	-2.06854
-0.27905	-0.41774	1.174765	0.931454	.....	-2.06854



-8.26756	1.569373	1.973389	2.199834	.....	-2.06854
0.427454	0.283031	-2.07111	-7.61578	.....	-2.06854
.....	.....	.....	.....	.....	.....
0.098014	0.148359	0.115822	0.007574	.....	-2.06854

### Algorithm Implementation

In this research, we conducted experiments using the Naive Bayes algorithm and Artificial Neural Network with two scenarios of the Natural Language Processing model. Naive Bayes is the Bayesian network's simplest form, where all the attributes do not depend on the value of the class variable. The advantage of Naïve Bayes is that it is a machine learning algorithm effective and efficient. Artificial Neural Network is an algorithm that is able to extract from a certain data pattern so that it can create a pattern of knowledge. In the implementation of the Artificial Neural Network, we made an Architectural model with details of 112 input layer nodes, 10 neurons in the first hidden layer, 8 neurons in the second hidden layer, 6 neurons in the third hidden layer, 8 neurons in the fourth hidden layer, and 1 neuron in the output layer.

The test was carried out four times in the experiment for each scenario of the Natural Language Processing model. The experiments we did include the Naïve Bayes Algorithm with the random split method, the Naïve Bayes Algorithm with the cross-validation method, the Artificial Neural Network with the random split method, and the Artificial Neural Network with the cross-validation method.

### First Scenario Algorithm Implementation

In the implementation of the first scenario, there are two data split methods used, namely Random Split and Kfold Cross-Validation to test the Naive Bayes algorithm and Artificial Neural Network. In the first scenario testing of the Naïve Bayes Algorithm with the random split method, we divide the dataset into three comparisons of training data and test data, namely testing data 10% from the dataset, 20% from the dataset, and 30% from the dataset. Then performed testing performance using operator classification reports for each comparison of split data. Figure 6 shows the results of the Naive Bayes first scenario random split method. In

the first scenario, the performance of the model obtained in this test is 65% accuracy for testing data of 10% of the dataset, 68% accuracy for testing data of 20% of the dataset, and 69% accuracy for testing data of 30% of the dataset.

-----					
Classification Report Data Testing 10%					
	precision	recall	f1-score	support	
0.0	0.19	0.75	0.30	28	
1.0	0.96	0.64	0.77	248	
accuracy			0.65	276	
macro avg	0.57	0.69	0.53	276	
weighted avg	0.88	0.65	0.72	276	
Elapsed time: 0:00:00.011504					
-----					
Classification Report Data Testing 20%					
	precision	recall	f1-score	support	
0.0	0.19	0.62	0.29	56	
1.0	0.94	0.69	0.80	495	
accuracy			0.68	551	
macro avg	0.56	0.66	0.54	551	
weighted avg	0.87	0.68	0.75	551	
Elapsed time: 0:00:00.006488					
-----					
Classification Report Data Testing 30%					
	precision	recall	f1-score	support	
0.0	0.18	0.63	0.28	78	
1.0	0.95	0.70	0.80	748	
accuracy			0.69	826	
macro avg	0.56	0.66	0.54	826	
weighted avg	0.87	0.69	0.75	826	
Elapsed time: 0:00:00.005597					

Figure 6. Naive Bayes First Scenario Random Split

In the first scenario testing of the Naïve Bayes Algorithm with the Kfold Cross-Validation method, the dataset is divided into several K values. Then iteration is carried out to calculate the average accuracy value. Figure 7 shows the results of the Naive Bayes first scenario Kfold Cross-Validation method. There are three grades K used is 5 K, 10 K, and 15K. From the three values Kfold, the performance is accuracy model 66.85% for 5K, 66.65% for 10 K, and 66.78% for the15K.

Elapsed time to compile 5 Kfold Data testing: 0:00:00.041494
Avg accuracy 5 Kfold : 66.85886817356872
Elapsed time to compile 10 Kfold Data testing: 0:00:00.063588
Avg accuracy 10 Kfold : 66.65006587615284
Elapsed time to compile 15 Kfold Data testing: 0:00:00.090393
Avg accuracy 15 Kfold : 66.78268789102717

Figure 7. Naive Bayes First Scenario Cross-Validation

In the first scenario experiments of the model Artificial Neural with the random split

method, the dataset is divided into several comparisons of training data and testing data. The comparisons made include 10% testing data of the whole dataset, 20% testing data of the whole dataset, and 30% testing data of the whole dataset. Figure 8 shows the results of the Artificial Neural Network first scenario random split method. From this experiment, the accuracy model is 86.96% for testing data 10%, 87.84% for testing data 20%, and 90.19% for testing data 30%.

```
-----
Elapsed time to compile 10 percent Data testing: 0:00:37.892766
Accuracy 10 percent Data testing: 86.96

-----
Elapsed time to compile 20 percent Data testing: 0:00:33.068614
Accuracy 20 percent Data testing: 87.84

-----
Elapsed time to compile 30 percent Data testing: 0:00:28.354010
Accuracy 30 percent Data testing: 90.19
```

Figure 8. ANN First Scenario Random Split

In the first scenario implementation of the Artificial Neural Network using the method Kfold Cross-Validation, We divide the dataset into several K values that among others the value of 5 K, 10 K, and 15K. Figure 9 shows the results of the first scenario Artificial Neural Network performance using the Kfold Cross-Validation method. From the test of each Kfold value, it shows the accuracy of the performance of the Artificial Neural Network model of 89.06% for 5K, 91.82% for 10K, and 93.03% for 15K.

```
-----
Elapsed time to compile 5 Kfold Data testing: 0:02:29.694699
Avg accuracy 5 Kfold: 89.06503796577454

-----
Elapsed time to compile 10 Kfold Data testing: 0:05:38.205834
Avg accuracy 10 Kfold: 91.82819545269012

-----
Elapsed time to compile 15 Kfold Data testing: 0:08:41.898047
Avg accuracy 15 Kfold: 93.03199529647827
```

Figure 9. ANN First Scenario Cross-Validation

**Second Scenario Implementation**

Like in the first scenario, the implementation of second scenario, there are two data split methods used, namely Random Split and Kfold Cross-Validation to test the Naive Bayes algorithm and Artificial Neural Network. In the second scenario testing of the Naïve Bayes Algorithm with the random split method, the data is divided into training data and testing data as many as three

comparisons, namely testing data 10% from the dataset, 20% from the dataset, and 30% from the dataset. Then performed testing performance using operator classification reports for each comparison of split data. Figure 10 shows the results of the Naive Bayes second scenario random split method. In this scenario, the model performance obtained in this test is 65% accuracy for testing data 10% of the dataset, 70% accuracy for testing data 20% of the dataset, and 70% accuracy for testing data of 30% of the dataset.

Classification Report Data Testing 10%				
	precision	recall	f1-score	support
0.0	0.18	0.68	0.28	28
1.0	0.95	0.65	0.77	248
accuracy			0.65	276
macro avg	0.56	0.66	0.52	276
weighted avg	0.87	0.65	0.72	276
Elapsed time: 0:00:00.013883				
Classification Report Data Testing 20%				
	precision	recall	f1-score	support
0.0	0.19	0.59	0.28	56
1.0	0.94	0.71	0.81	495
accuracy			0.70	551
macro avg	0.56	0.65	0.55	551
weighted avg	0.86	0.70	0.75	551
Elapsed time: 0:00:00.006552				
Classification Report Data Testing 30%				
	precision	recall	f1-score	support
0.0	0.17	0.58	0.27	78
1.0	0.94	0.72	0.81	748
accuracy			0.70	826
macro avg	0.56	0.65	0.54	826
weighted avg	0.87	0.70	0.76	826
Elapsed time: 0:00:00.007076				

Figure 10. Naive Bayes Second Scenario Random Split

In the second scenario testing of the Naïve Bayes Algorithm with the K Fold Cross-Validation method, we divide the dataset into several K values. Then we performed a test iteration to calculate the mean accuracy value. Figure 11 shows the results of the Naive Bayes second scenario Kfold Cross-Validation method. There are three K values that we use which include 5 K, 10 K, and 15K. From the three values Kfold, the performance is accuracy model 65.98% for 5K, 65.92% for 10 K, and 65.80% for the 15K.

```

Elapsed time to compile 5 Kfold Data testing: 0:00:00.044712
Avg accuracy 5 Kfold : 65.98719683220591

Elapsed time to compile 10 Kfold Data testing: 0:00:00.061861
Avg accuracy 10 Kfold : 65.92305665349143

Elapsed time to compile 15 Kfold Data testing: 0:00:00.078163
Avg accuracy 15 Kfold : 65.8024471370872

```

Figure 11. Naive Bayes Second Scenario Cross-Validation

In the second scenario experiments of the model Artificial Neural with the random split method, we divide the dataset into several comparisons of training data and testing data. The comparisons made include 10% testing data of the whole dataset, 20% testing data of the whole dataset, and 30% testing data of the whole dataset. Figure 12 shows performance test result of the model Artificial Neural Network in the second scenario random split method. From this experiment, the accuracy model is 87.32% for testing data 10%, 89.47% for testing data 20%, and 90.44% for testing data 30%.

```

-----
Elapsed time to compile 10 percent Data testing: 0:00:36.047640
Accuracy 10 percent Data testing: 87.32

-----
Elapsed time to compile 20 percent Data testing: 0:00:31.992148
Accuracy 20 percent Data testing: 89.47

-----
Elapsed time to compile 30 percent Data testing: 0:00:28.321562
Accuracy 30 percent Data testing: 90.44

```

Figure 12. ANN Second Scenario Random Split

Implementation of Artificial Neural Network In the second scenario of using the method Kfold Cross-Validation, the dataset is divided into several K values that among others the value of 5 K, 10 K, and 15K. Figure 9 is the performance test result of the second scenario Artificial Neural Network model with the Kfold Cross-Validation. From testing with each Kfold value, showed performance accuracy models Artificial Neural Network of 89.93% for the 5 K, 91.64% to 10 K, and 92.81% for the 15K.

```

-----
Elapsed time to compile 5 Kfold Data testing: 0:02:28.693469
Avg accuracy 5 Kfold: 89.93611574172974

-----
Elapsed time to compile 10 Kfold Data testing: 0:05:38.720701
Avg accuracy 10 Kfold: 91.6470354795456

-----
Elapsed time to compile 15 Kfold Data testing: 0:08:42.155843
Avg accuracy 15 Kfold: 92.81262397766113

```

Figure 13. ANN Second Scenario Cross-Validation

## Evaluation and Scenario Comparison

At this stage, we evaluate and compare scenarios to determine the effect of the Natural Language Processing model in representing medical abstracts on the accuracy of the Naïve Bayes algorithm and the Artificial Neural Network. In the first scenario, the implementation of the Naïve Bayes algorithm using the random split method shows an average accuracy model performance of 67.33% with an average testing time of 0.0078 seconds. Meanwhile, the Artificial Neural Network with the random split method obtained an average accuracy performance of 88.33% with an average test duration of 33.1 seconds. In the implementation of the Naïve Bayes algorithm using the kfold cross-validation method, the average accuracy model performance is 66.76% with an average testing time of 0.064 seconds. Meanwhile, the Artificial Neural Network with the kfold cross-validation method obtained an average accuracy performance of 91.30% with an average test duration of 5 minutes 36.59 seconds. Table 4 shows the results of first scenario accuracy for each algorithm.

Table 4: Results of First Scenario Accuracy

Random Split Method			
Algorithm	Data Testing	Testing Time	Accuracy
Naïve Bayes	10%	0.0115 seconds	65%
Naïve Bayes	20%	0.0064 seconds	68%
Naïve Bayes	30%	0.0055 seconds	69%
<b>Average</b>		<b>0.0078 seconds</b>	<b>67.33%</b>
ANN	10%	37.89 seconds	86.96%
ANN	20%	33.06 seconds	87.84%
ANN	30%	28.35 seconds	90.19%
<b>Average</b>		<b>33.1 seconds</b>	<b>88.33%</b>
Cross Validation Method			
Algorithm	Kfold	Testing Time	Accuracy
Naïve Bayes	5 K	0.041 seconds	66.85%
Naïve Bayes	10 K	0.063 seconds	66.65%
Naïve Bayes	15 K	0.090 seconds	66.78%
<b>Average</b>		<b>0.064 seconds</b>	<b>66.76%</b>
ANN	5 K	2 minutes 29.69 seconds	89.06%
ANN	10 K	5 minutes 38.20 seconds	91.82%
ANN	15 K	8 minutes 41.89 seconds	93.03%
<b>Average</b>		<b>5 minutes 36.59 seconds</b>	<b>91.30%</b>

In the second scenario, the implementation of the Naïve Bayes algorithm with the random



split method shows an average accuracy model performance of 68.33% with an average testing time of 0.0088 seconds. Meanwhile, the Artificial Neural Network with the random split method obtained an average accuracy performance of 89.08% with an average test duration of 32.11 seconds. In the implementation of the Naïve Bayes algorithm with the cross-validation method, the average accuracy model performance is 65.90% with an average testing time of 0.061 seconds. Meanwhile, the Artificial Neural Network with the cross-validation method obtained an average accuracy performance of 91.46% with an average test duration of 5 minutes 36.43 seconds. Table 5 shows the results of second scenario accuracy for each algorithm.

Table 5:  
Results of Second Scenario Accuracy

Random Split Method			
Algorithm	Data Testing	Testing Time	Accuracy
Naïve Bayes	10%	0.013 seconds	65%
Naïve Bayes	20%	0.0065 seconds	70%
Naïve Bayes	30%	0.007 seconds	70%
<b>Average</b>		<b>0.0088 seconds</b>	<b>68.33%</b>
ANN	10%	36.04 seconds	87.32%
ANN	20%	31.99 seconds	89.47%
ANN	30%	28.32 seconds	90.44%
<b>Average</b>		<b>32.11 seconds</b>	<b>89.08%</b>
Kfold Cross Validation Method			
Algorithm	Kfold	Testing Time	Accuracy
Naïve Bayes	5 K	0.044 seconds	65.98 %
Naïve Bayes	10 K	0.061 seconds	65.92 %
Naïve Bayes	15 K	0.078 seconds	65.80 %
<b>Average</b>		<b>0.061 seconds</b>	<b>65.90 %</b>
ANN	5 K	2 minutes 28.69 seconds	89.93%
ANN	10 K	5 minutes 38.72 seconds	91.64%
ANN	15 K	8 minutes 41.89 seconds	92.81%
<b>Average</b>		<b>5 minutes 36.43 seconds</b>	<b>91.46%</b>

Overall, from all the experiments conducted, it was found that the highest average accuracy of the Naïve Bayes algorithm was 68.33% by using the random split method in the second scenario. Meanwhile, the Artificial Neural Network obtained the highest average accuracy is 91.46% by using the cross-validation method in the second scenario. From these two results, we can see that the accuracy value of the two algorithms is quite significant, namely the Artificial

Neural Network is 23.12% higher than the Naïve Bayes algorithm.

In addition, from these experiments, it was also found that the Naïve Bayes Algorithm with the second scenario random split method was 1% better than the first scenario. Meanwhile, the Naïve Bayes algorithm with the cross-validation method in the first scenario is 0.86% better than the second scenario. In the Artificial Neural Network with the random split method the second scenario is 0.74% better than the first scenario and the Artificial Neural Network with the cross-validation method, the second scenario is 0.15% better than the first scenario. So that we know that adding a medical abstract to the training dataset of the Word2Vec model can affect the accuracy of testing even if only slightly. Table 6 shows the results of overall scenario accuracy comparison.

Table 6:  
Results of Overall Scenario Accuracy Comparison

First Scenario			
Algorithm	Metode	Testing Time	Accuracy
Naïve Bayes	Random Split	0.0078 seconds	67.33%
ANN	Random Split	33.1 seconds	88.33%
Naïve Bayes	Cross Validation	0.064 seconds	66.76%
ANN	Cross Validation	5 minutes 36.59 seconds	91.30%
Second Scenario			
Algorithm	Metode	Testing Time	Accuracy
Naïve Bayes	Random Split	0.0088 seconds	68.33%
ANN	Random Split	32.11 seconds	89.08%
Naïve Bayes	Cross Validation	0.061 seconds	65.90%
ANN	Cross Validation	5 minutes 36.43 seconds	91.46%

#### IV. CONCLUSION

This paper presents the results of a study applying a Comparative Study between Naive Bayes and ANN Classifier to detect disease types based on a doctor's diagnosis dataset. The experimental results show that in general ANN provides better performance than Naive Bayes. In the random split dataset, the highest average accuracy performance of

Naive Bayes is 68.33%, while the average accuracy performance of ANN is 89.08%. Both are generated from the second scenario, namely the involvement of the unstructured dataset in the word2vec training process. In the cross validation dataset, the highest Naive Bayes average accuracy performance was 66.76% generated from the first scenario, while the ANN average accuracy performance was 91.46% generated from the second scenario. On average, ANN in the first scenario gave an average accuracy of 89.82%, while in the second scenario it was 90.27%. It can be concluded that the involvement of unstructured data from the dataset in the word2vec training process improves the performance of the ANN model even though it is not significant.

## V. REFERENCES

- [1] P. S. Roshanov et al., "Computerized clinical decision support systems for chronic disease management: A decision-maker-researcher partnership systematic review," *Implement. Sci.*, vol. 6, no. 1, pp. 1–17, 2011, doi: 10.1186/1748-5908-6-92.
- [2] M. Ridwan, H. Suyono, and M. Sarosa, "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," *Eeccis*, vol. 7, no. 1, pp. 59–64, 2013, doi: 10.1038/hdy.2009.180.
- [3] A. Aziz Priatna, R. Megasari, and J. Kusnendar, "Penerapan Association Rules Menggunakan Algoritma Apriori Pada Sistem Rekomendasi Pemilihan Resep Obat Berdasarkan Data Rekam Medis," *J. Teor. dan Apl. Ilmu Komput.*, vol. 1, no. 2, pp. 55–60, 2018, [Online]. Available: <http://jaticom.cs.upi.edu>.
- [4] Guardian Y. Sanjaya, S. Harry, L. Lazuardi, and N. Faizah, "Datamining pereseapan elektronik di pelayanan kesehatan primer: potensi pengembangan sistem pendukung keputusan klinis," *Semin. Nas. Inform. Medis*, no. September, pp. 26–30, 2012.
- [5] N. Indrawati, "NATURAL LANGUAGE PROCESSING ( NLP ) BAHASA INDONESIA adalah :," no. 1, 2010.
- [6] N. I. Widiastuti, "Deep Learning - Now and Next in Text Mining and Natural Language Processing," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 407, no. 1, 2018, doi: 10.1088/1757-899X/407/1/012114.
- [7] T. F. M. Raj and S. Prasanna, "Implementation of ML using naïve bayes algorithm for identifying disease-treatment relation in bio-science text," *Res. J. Appl. Sci. Eng. Technol.*, vol. 5, no. 2, pp. 421–426, 2013, doi: 10.19026/rjaset.5.4968.
- [8] S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, and V. Osmani, "Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review," *arXiv*, vol. 7, pp. 1–18, 2019, doi: 10.2196/12239.
- [9] F. Esser and R. Vliegthart, "Comparative Research Methods," *Int. Encycl. Commun. Res. Methods*, pp. 1–22, 2017, doi: 10.1002/9781118901731.iecrm0035.
- [10] J. Olamaei and S. Ashouri, "Demand response in the day-ahead operation of an isolated microgrid in the presence of uncertainty of wind power," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 23, no. 2, pp. 491–504, 2015, doi: 10.3906/elk-1301-164.
- [11] B. S. Prakoso, D. Rosiyadi, H. S. Utama, and D. Aridarma, "Klasifikasi Berita Menggunakan Algoritma Naive Bayes Classifier Dengan Seleksi Fitur Dan Boosting," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 2, pp. 227–232, 2019, doi: 10.29207/resti.v3i2.1042.
- [12] Y. D. Prabowo, T. L. Marselino, and M. Suryawiguna, "Pembentukan Vector Space Model Bahasa Indonesia Menggunakan Metode Word to Vector," *J. Buana Inform.*, vol. 10, no. 1, p. 29, 2019, doi: 10.24002/jbi.v10i1.2053.
- [13] W. I and S. S. U. Rahman S, "Treatment of Missing Values in Data Mining," *J. Comput. Sci. Syst. Biol.*, vol. 09, no. 02, pp. 51–53, 2015, doi: 10.4172/jcsb.1000221.



- [14] irwan budiman, M. R. Faisal, and D. T. Nugrahadi, "Studi Ekstraksi Fitur Berbasis Vektor Word2Vec pada Pembentukan Fitur Berdimensi Rendah," *J. Komputasi*, vol. 8, no. 1, pp. 62–69, 2020, doi: 10.23960/komputasi.v8i1.2517.
- [15] A. Kao and S. R. Poteet, *Natural language processing and text mining*. 2007.
- [16] B. Jang, I. Kim, and J. W. Kim, "Word2vec convolutional neural networks for classification of news articles and tweets," *PLoS One*, vol. 14, no. 8, pp. 1–20, 2019, doi: 10.1371/journal.pone.0220976.
- [17] F. Rahutomo, P. Y. Saputra, and C. F. P. Putra, "Implementasi Explicit Semantic Analysis Berbahasa Indonesia Menggunakan Corpus Wikipedia Indonesia," *J. Inform. Polinema*, vol. 4, no. 4, pp. 252–257, 2018, doi: 10.33795/jip.v4i4.215.
- [18] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: An introduction," *J. Am. Med. Informatics Assoc.*, vol. 18, no. 5, pp. 544–551, 2011, doi: 10.1136/amiajnl-2011-000464.
- [19] D. Berrar, "Cross-validation," *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, no. January 2018, pp. 542–545, 2018, doi: 10.1016/B978-0-12-809633-8.20349-X.
- [20] F. Tempola, M. Muhammad, and A. Khairan, "Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, p. 577, 2018, doi: 10.25126/jtiik.201855983.
- [21] F. E. Prabowo and A. Kodar, "Analisis Prediksi Masa Studi Mahasiswa Menggunakan Algoritma Naïve Bayes," *J. Ilmu Tek. dan Komput.*, vol. 3, no. 2, p. 147, 2019, doi: 10.22441/jitkom.2020.v3.i2.008.
- [22] N. Salmi and Z. Rustam, "Naïve Bayes Classifier Models for Predicting the Colon Cancer," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 546, no. 5, 2019, doi: 10.1088/1757-899X/546/5/052068.
- [23] A. P. Wibawa et al., "Naïve Bayes Classifier for Journal Quartile Classification," *Int. J. Recent Contrib. from Eng. Sci. IT*, vol. 7, no. 2, p. 91, 2019, doi: 10.3991/ijes.v7i2.10659.
- [24] A. S. Kurniawansyah, "Implementasi Metode Artificial Neural Network dalam Memprediksi Hasil Ujian Kompetensi Kebidanan," vol. V, 2018.
- [25] F. Amato, A. López, E. M. Peña-Méndez, P. Vañhara, A. Hampl, and J. Havel, "Artificial neural networks in medical diagnosis," *J. Appl. Biomed.*, vol. 11, no. 2, pp. 47–58, 2013, doi: 10.2478/v10136-012-0031-x.

## KERTAS KERJA

### Ringkasan

Kertas kerja ini merupakan material kelengkapan artikel jurnal yang berjudul “*Comparative Study of Naïve Bayes and Artificial Neural Network to Detect Patients’ Disease Types by Using Structural and Unstructural Data*”. Kertas kerja berisi semua material hasil penelitian Tugas Akhir yang tidak dimuat/atau disertakan di artikel jurnal. Di dalam kertas kerja ini disajikan beberapa bagian sebagai berikut:

1. *Literature review* merupakan bagian yang berisi hasil studi literatur yang dilakukan terkait dengan eksperimen yang dilakukan. Secara garis besar *literature review* yang dilakukan tentang konsep *Natural Language Processing, Data mining, Algoritma Naïve Bayes, Artificial Neural Network, Cross Validation*, serta literatur tentang jenis penyakit.
2. Analisis dan Perancangan merupakan bagian yang terdiri gambaran secara garis besar serta tahapan-tahapan yang dilakukan pada penelitian ini. Pada tahapan ini dijelaskan bahwa penelitian dilakukan dengan menggunakan 2 skenario. Pada skenario pertama proses *training model Word2vec* skenario pertama hanya menggunakan *corpus wikipedia* sebagai dataset. Skenario kedua proses *training model Word2vec* skenario kedua tidak hanya menggunakan *corpus wikipedia* sebagai dataset. Tetapi *corpus wikipedia* dikombinasikan dengan *medical abstract* sebagai dataset.
3. Pada bagian ini membahas *source code* pada penelitian yang dilakukan. Eksperimen dilakukan dengan menggunakan *database Postgresql* dengan *tools Navicat* dan bahasa pemrograman *python* dengan dua *tools* yang berbeda. *Postgresql* digunakan untuk menampung dataset awal serta untuk melakukan proses *cleaning data*. *Python* dengan *tools* sistem operasi *windows 10* yang dijalankan menggunakan *command line* digunakan untuk melatih model *Word2vec* serta representasi *medical abstract*. *Python* dengan *tools google collabs* digunakan untuk melakukan implementasi algoritma *Naïve Bayes* dan *Artificial Neural Network*.

4. Dataset merupakan bagian yang menjelaskan tentang dataset yang digunakan dalam eksperimen. Pada bagian ini dijelaskan struktur dataset awal, *treatment* yang dilakukan agar dataset siap digunakan pada penelitian, serta dataset hasil dari representasi *medical abstract* menjadi bentuk *vector*.
5. Tahapan Eksperimen merupakan bagian yang berisi tahapan eksperimen seluruhnya yang tidak tercakup di jurnal. Bagian ini secara garis besar menjelaskan tentang alur teknis penelitian yang dilakukan secara keseluruhan. Tahapan-tahapan yang dijelaskan pada bagian ini antara lain adalah tahap pengumpulan data, pembersihan dan *treatment data*, konversi *text data*, *data splitting*, implementasi algoritma *Naïve Bayes*, implementasi *Artificial Neural Network*, serta evaluasi dan perbandingan skenario.
6. Hasil Semua Eksperimen merupakan bagian yang menjelaskan tentang hasil eksperimen yang diperoleh dengan dua skenario yang dilakukan. Masing-masing skenario dilakukan sebanyak empat kali eksperimen. Eksperimen yang dilakukan antara lain adalah Algoritma *Naïve Bayes* dengan metode *random split*, Algoritma *Naïve Bayes* dengan metode *Cross Validation*, *Artificial Neural Network* dengan metode *random split*, dan *Artificial Neural Network* dengan metode *Cross Validation*.