Cluster analysis of type II Diabetes Mellitus Patients with the Fuzzy C-means method

Simeftiany Indrilemta Lomo [a,1,*], Endang Darmawan[a,2], Sugiyarto[b,3]

[a] Faculty of Pharmacy, Ahmad Dahlan University, 55164, Yogyakarta, Indonesia
[b] Faculty of Applied Science and Technology, Ahmad Dahlan University, 55191, Bantul, Indonesia
[1] indrilomo@gmail.com *; [2] endang.darmawan@pharm.uad.ac.id, [3] sugiyarto@math.uad.ac.id
* corresponding author

ARTICLE INFO

Keywords
Cluster analysis
Type II Diabetes Mellitus
Fuzzy C-Means
Male

ABSTRACT

Cluster analysis has been widely used in the fields of mathematics and health sciences. This study aims to classify distance-based data which are divided into several clusters. Accurate prediction from the outcome or survival rate of diabetic patients can be the key for the stratification of prognosis and therapy. A retrospective study of 447 medical record data of type II diabetes mellitus patients aged 18 years old or above and were hospitalized in the PKU Muhammadiyah Gamping Hospital from 2015-2019. Clustering is using the PCA-Fuzzy C-Means method based on patients' survival status, demographic characteristics, therapy, and blood glucose (BG) levels. Clustering evaluation by Davies Bouldin Index (DBI). Data analysis is using Jupyter Notebook programme. Cluster formation are first cluster consists of 171 members, second cluster consists of 9 members, third cluster consists of 267 members with DBI 2,2645. 401 patients (89,7%) were recorded as alive and 46 patients (10,3%) were recorded as dead. A total of 447 patients: 54,1% were male; 90,6% were ≥ 45 years old; 66,4% has comorbidities; 51,7% had BG level of more than 200 mg/dl, and 57,7% received combination insulin+oral antidiabetic therapy.

1. Introduction

Hyperglycemia is one of the typical symptoms of diabetes mellitus (DM). It is caused by abnormalities in insulin secretion, insulin action or both, characterized by an increase in glucose levels in the blood beyond normal limits. Diabetes mellitus is one of the threats to global health including Indonesia [1]. Based on data from basic health research in 2018, the prevalence of diabetes mellitus among people aged 15 years in Indonesia increased from 1,5% in 2013 to 2,0% in 2018 [2]. World Health Organization (WHO) states that diabetes causes 1,5 million deaths in 2012. WHO also estimates that people with Diabetes Mellitus will increase to around 21,3 million in 2030 and increase by 2-3 times in 2035 [3].

Diabetes caused 1,5 million deaths in 2012. Blood glucose levels higher than the optimal limit caused additional 2,2 million deaths, by increasing the risk of cardiovascular disease and other diseases. About 43% of these 3,7 million deaths occurred before the age of 70. The percentage of deaths is caused by diabetes that occurs before age of 70 is higher in low- and middle-income countries than in high-income countries [3].

10.33292/amm.v3i1.28

One way that can be used to extract patient data is clustering. Cluster analysis is a statistical method that identifies a particular structure by grouping objects based on similar characteristics into a group. A cluster is identical with the similarities between the members in one cluster, and the differences between one cluster and another[4]. Fuzzy C-Means (FCM) is the most frequently used analysis. This analysis is part of the K-Means method where each data in a cluster is determined by the degree of membership[5], it can be applied to a wide and varied data, and gives more optimal results[6]–[9]. However, this analysis has weaknesses in determining the initial cluster center. The classification and clustering algorithms become significantly challenging for data with high dimensions that might affect the final result. This challenge can be solved by doing dimensional reduction. One of the most widely used data reduction methods is Principal Component Analysis (PCA). PCA is an analysis technique by reducing the dimensions of a large data sets into new variable data that is linear with the initial data [10]. Based on the research conducted by Surono and Putri (2020), was found that the combination of the PCA and FCM algorithms gave better results than the pure FCM algorithm [11].

Cluster analysis using Fuzzy C-Means method has been used in the prognosis of diabetes mellitus[12]–[15]. Fuzzy C-Means analysis in this study aims to facilitate the process of grouping data based on demographic characteristics, pharmacological therapy and patient glucose levels. In addition, the results of the analysis are expected to contribute for the selection of therapy that is suitable for the patients' condition with the intent that it can support the long-term survival of patients with type II diabetes mellitus.

2. Method

This study is an observational study, in which the observations were carried out retrospectively on medical record data of patients with Type II Diabetes Mellitus for the period 2015 to 2019 at PKU Muhammadiyah Gamping Hospital, Yogyakarta. The inclusion criteria were patients aged ≥ 18 years who were hospitalized with a diagnosis of Type II Diabetes Mellitus for the period from 2015 to 2019 with or without comorbidities. Patients with type I DM, gestational diabetes, and patients with unreadable or incomplete medical record data were excluded. The research was held from August to October 2020, and it has received and passed an ethical review statement from the Research Ethics Commission of PKU Muhammadiyah Yogyakarta Hospital with number 0011 / KT.7.4 / VIII / 2020. Data was collected by using the data collection sheet that has been provided, it is including patient characteristics data, records of the use of diabetes mellitus drugs which are consumed by patients, and laboratory results. Patient medical record data that has been obtained are classified and coded according to the type and variety. Dependent variable: Patient status (0= Survivor; 1= Non-Survivor). Independent variabel: (1)Sex (1= Male; 2= Female); (2)Age (1= <45 years old; 2= ≥45 years old); (3)Comorbidity (0= None; 1= With comorbidities); (4)Types of Complications (0= None; 1= Renal Complications; 2= Neurological Complications; 3= Peripheral Circulation Complications; 4= Multiple complications; 5= Coma); (5)Therapy (1= Insulin; 2= Oral Antidiabetics; 3= Insulin+ Oral Antidiabetics); (6)Blood Glucose Levels (1= Normal < 140 mg/dl; 2= Moderate 140-199 mg/dl; 3= High ≥ 200 mg/dl).

Data reduction was performed by using the PCA method. Before implementing PCA, the data set needs to be standardized, in such a way that the attributes with larger domains will not dominate the other attributes with smaller domains. The data set obtained from the application of PCA that was grouped using the Fuzzy C-Means method and the clusters formed were tested for validity by using the Davies Bouldin Index (DBI) method. Data analysis used Jupyter Notebook programme Figure 1.

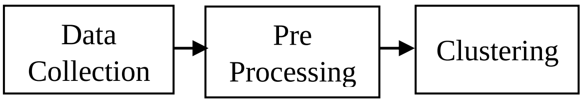

Fig. 1. Research Method

3. Results and Discussion

Based on the results of the study, there were 447 medical record data of Type II Diabetes Mellitus patients who met the inclusion and exclusion criteria (Table 1). The results showed that 401 patients (89,7%) were recorded as survival and 46 patients (10,3%) were recorded as died (non-survival) during the period 2015 to 2019. Overall, 242 patients were female (54,1 %) and 205 patients were male (45,9%). This is similar to a study of Hongdiyanto et al (2014) at Prof. Dr. R.D. Kandou Manado, consisting of 16 male patients and 30 female patients [16]. Gender is a statistically significant factor associated with repeated occurrences of recovery in blood sugar levels. Recovery time is better seen in male patients [17]. This finding is also consistent with a research that was conducted by Terefe et al., (2018) found that female have a longer time than male to recover for reaching normal blood glucose levels [18].

Table 1. Clustering Results

Variable	**Cluster 1 (n=171)**	**Cluster 2 (n=9)**	**Cluster 3 (n=267)**
Patient Status			
Survivor	171	9	221
Non-Survivor	0	0	46
Sex			
Male	120	9	76
Female	51	0	191
Age			
<45 years old	0	0	42
≥45 years old	171	9	225
Type of Complications			
Without Complication	75	0	151
Renal Complication	7	0	13
Neurological Complication	12	0	16
Peripheral Circulation Complication	77	9	66
Multiple Complication	0	0	5
Coma	0	0	16
Comorbidities			
Without Comorbidity	97	0	53
With Comorbidities	74	9	214
Therapy			
Insulin	0	0	37
Oral Antidiabetics	51	3	98
Insulin+ Oral Antidiabetics	120	6	132
Blood Glucose Level			
Normal < 140 mg/dl	0	0	49
Moderate 140-199 mg/dl	71	9	87
High ≥ 200 mg/dl	100	0	131

The age variable used was the age at which patient was first diagnosed with type 2 Diabetes Mellitus at the PKU Muhammadiyah Gamping Hospital. The age range ≥ 45 years was more with a total of 405 patients (90,6%) than patients with age range <45 years with

42 patients (9,4%). The complications category was classified into 6 groups which are without complications 226 patients (50,6%), renal complications 20 patients (4,5%), neurological complications 28 patients (6,3%), peripheral circulation complications 152 patients (34%), multi-complication 5 patients (1,1%), and coma 16 patients (3,6%). In the comorbidity's category 150 patients (33,6%) were treated without comorbidities while 297 other patients (66,4%) were treated with comorbidities. A total of 231 patients had high mean blood glucose level ≥ 200 mg/dL, 167 patients (37,4) had blood glucose level of 140-199 mg/dL, and 49 other patients (11%) had blood glucose level <140 mg/dL. In the therapy category, 258 patients (57,7%) received combination insulin therapy and oral antidiabetics, 152 patients (34%) received oral antidiabetics, and 37 other patients (8,3%) received insulin therapy. According to Shamshirgaran et al (2017), age has a correlation with a high percentage of complications. Although the prevalence of complications was high for older patient group, this group had better blood glucose control than the younger patient group [19]. Other studies assert that duration of diabetes is the strongest determinant for complications. Morbidity and mortality are greatest in patients diagnosed with diabetes at a young age [20].

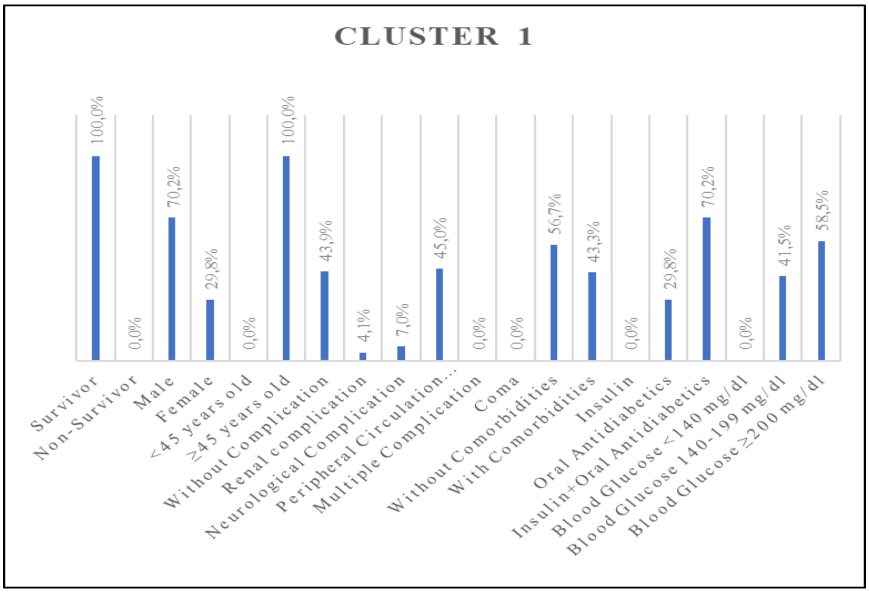

Fig. 2. Cluster 1

Cluster analysis is intended to classify objects or data into groups based on similar characteristics. In this study, the clustering process was carried out by assigning 3 clusters of 100 iterations. Figure 2, cluster 1 consisted of 171 members where all members were patients with surviving status, and the majority were male aged ≥ 45 years (70,2%) with 45% peripheral circulation complications and 43,9% without complications, 56% without comorbidities, with medium-high blood glucose levels, and consuming a combination of insulin therapy and oral antidiabetic as much as 70,2%. Cluster 2 consisted of 9 surviving patients, all patients were male and aged ≥ 45 years with peripheral circulation complications and comorbidities, and also had moderate glucose levels. 33,3% of patients in this cluster received oral antidiabetic therapy and 66,7% combined oral insulin and antidiabetic therapy. Cluster 3 consisted of 267 members with 221 members were surviving patients and 46 members were patients with non-survive status, it was dominated by female patients, 71,5% with age ≥ 45 years, 56,6% of patients without complications, 24,7% had peripheral circulation complications, 80,1% with comorbidities, blood glucose levels were high (49,1%) and also received a combination of insulin therapy and oral antidiabetics.

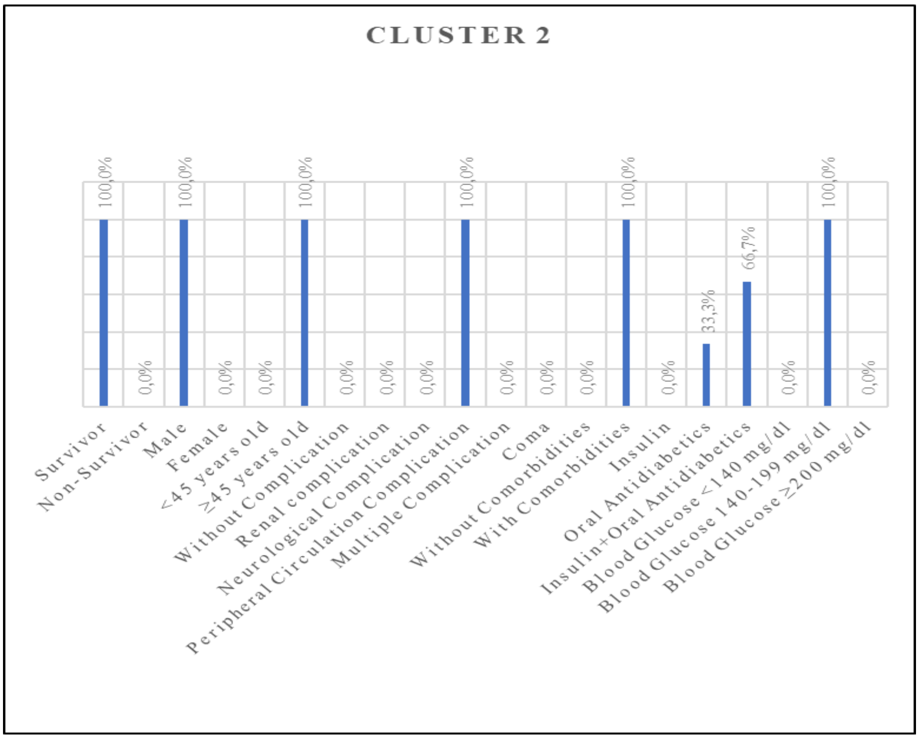

Fig. 3. Cluster 2

The existing pattern forms one similarity or similarities in which type II diabetes mellitus patients aged ≥ 45 years tend to have moderate-high blood glucose level therefore they receive combination therapy, namely insulin and oral antidiabetic Figure 3. Patients with complications and comorbidities, with high blood glucose levels have greater risk of death than patients without complications and comorbidities with normal blood glucose levels. Dewi et al (2018) stated that comorbidities have a value of 0,640, which means that each diabetes mellitus patient with no other disease but only diabetes mellitus has a slower time to attain failure 0,640 times than diabetes mellitus patients who have other diseases [21]. In addition, other diseases related to blood vessel circulation, such as uncontrolled hypertension and dyslipidemia (43,6%) that is followed by complications of diabetes mellitus (33,3%) contribute to lower probability of patient survival [22]. Sanusi et al in their study obtained 0,120 as a time-bound blood glucose coefficient which was positive. It is indicating that patients with high glucose levels have a risk of failure 1,128 times greater than patients who have low and normal blood glucose levels. In other words, patients with normal and low glucose levels have a longer survival time [23]. Juniarti and Nugraha found that the chances of survival for type II diabetes mellitus patients who went through oral antidiabetic therapy were greater than patients who went through insulin therapy treatment for all-time t tested [24]. However, a study conducted by Derebew stated that patients who received only oral antidiabetic therapy and patients who received oral antidiabetic therapy combined with insulin had a longer recovery time than patients who received only insulin therapy [17]. Based on the research of Putri and Astutik, age is also an independent variable that significantly influences the survival time of diabetes mellitus patients [25]. From the best log-normal Accelerated Failure Time (AFT) model obtained by Rachmaniyah et al, it was concluded that patients with one year older tended to have a faster chance of dying than younger patients in Figure 4 [26].

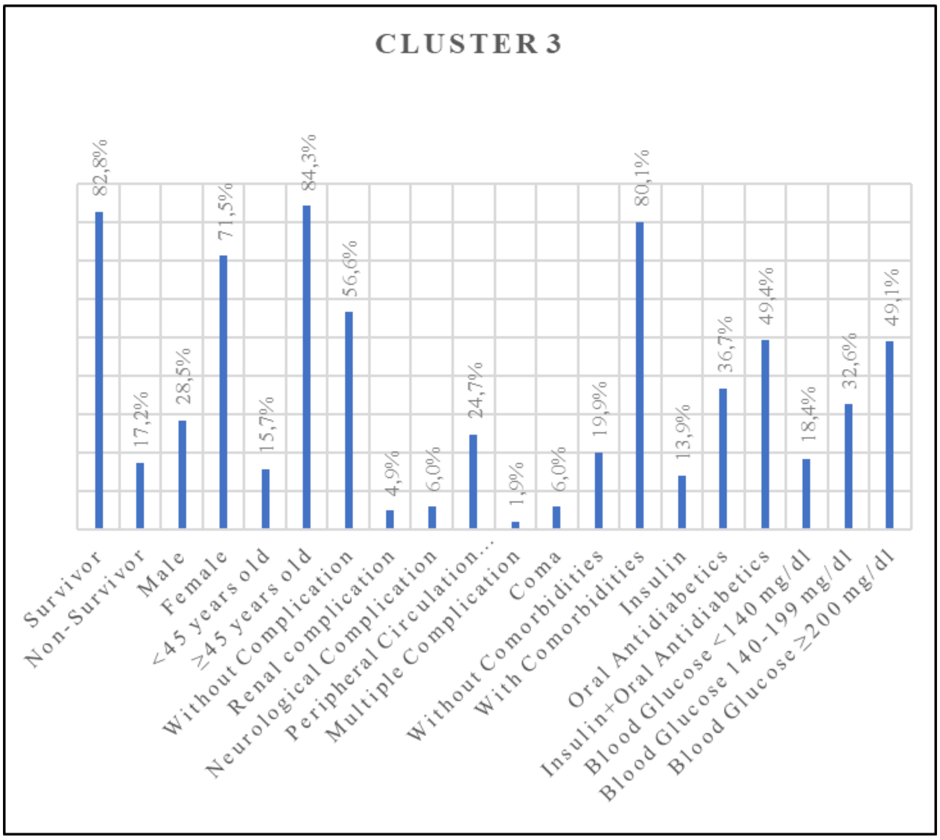

Fig. 4. Cluster 3

To test the validity of the clusters formed, Davies-Bouldin Index/DBI can be used. This method maximizes the distance between clusters and minimizes the distance between points in a cluster. The less DBI value obtained, the better cluster results[27]–[30]. In this study, the DBI value was 2,2645. This value is the smallest DBI value obtained after being compared with the formation of 5 clusters. Essentially, DBI wants the smallest possible value (non-negative ≥ 0) to assess how well the clusters are formed. The index is obtained as decision support data to assess the number of the most suitable clusters. The limitation of this study is that the data used is medical record data with the intent that the available data and variables that can be analyzed are minimal. In addition, the research period was limited because the research was carried out during the 2019 coronavirus disease pandemic (Covid-19).

4. Conclusion

In this work, the investigation of development three-dimensional mathematical model of steady flow in stenotic artery focused on the two type of models. The geometries proposed for TYPE I and TYPE II with each location of stenosis imposed in computational models have considerable impact on streamlines pattern. The flow of the blood has been governed by non-Newtonian of the streaming blood together with the effects of severity of stenosis.The non-Newtonian has been characterized by the generalized power-law model. The numerical computation is done using COMSOL Multiphysic 5.2. The effects of stenosis severity, different type of blood rheology and location of stenosis on streamlines pattern are approximated quantitatively. Based on the analyzation made, it can be observed that as the severity of the stenoses increases, the streamlines shown a abnormal behaviour where recirculation occur as the stenosis severity increases.

References

[1] Kementrian Kesehatan Republik Indonesia, *Pedoman Pelayanan Kefarmasian Pada Diabetes Melitus*. Jakarta: Kementerian Kesehatan Republik Indonesia. Direktorat Jendera Kefarmasian dan Alat Kesehatanl, 2019.

[2] RISKESDAS, "Laporan Riskesdas 2018," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 181–222, 2018.

[3] World Health Organization, "Global Report on Diabetes," *World Heal. Organ.*, vol. 978, pp. 6–86, 2016.

[4] S. Bano and M. N. A. Khan, "A Survey of Data Clustering Methods," *Int. J. Adv. Sci. Technol.*, vol. 113, no. December, pp. 133–142, 2018, doi: 10.14257/ijast.2018.113.14.

[5] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Comput. Geosci.*, vol. 10, no. 2–3, pp. 191–203, 1984, doi: 10.1016/0098-3004(84)90020-7.

[6] A. Praja, C. Lubis, and D. E. Herwindiati, "Deteksi Penyakit Diabetes Dengan Metode Fuzzy C-Means Clustering Dan K-Means Clustering," *Comput. J. Comput. Sci. Inf. Syst.*, vol. 1, no. 1, p. 15, 2017, doi: 10.24912/computatio.v1i1.233.

[7] K. V. Rajkumar, A. Yesubabu, and K. Subrahmanyam, "Fuzzy clustering and Fuzzy C-Means partition cluster analysis and validation studies on a subset of CiteScore dataset," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 4, pp. 2760–2770, 2019, doi: 10.11591/ijece.v9i4.pp2760-2770.

[8] R. Susilowati, A. S. Yazid, and S. Uyun, "Patient Data Clustering using Fuzzy C-Means (FCM) and Agglomerative Hierarchical Clustering (AHC)," *IJID (International J. Informatics Dev.*, vol. 3, no. 1, pp. 17–24, 2019.

[9] W. Wiharto and E. Suryani, "The Comparison of Clustering Algorithms K-Means and Fuzzy C-Means for Segmentation Retinal Blood Vessels," *Acta Inform. Medica*, vol. 28, no. 1, p. 42, 2020, doi: 10.5455/aim.2020.28.42-47.

[10] N. Salem and S. Hussein, "Data dimensional reduction and principal components analysis," *Procedia Comput. Sci.*, vol. 163, pp. 292–299, 2019, doi: 10.1016/j.procs.2019.12.111.

[11] S. Surono and R. D. A. Putri, "Optimization of fuzzy c-means clustering algorithm with combination of minkowski and chebyshev distance using principal component analysis," *Int. J. Fuzzy Syst.*, 2020, doi: 10.1007/s40815-020-00997-5.

[12] S. Jamuna and K. Mohan Kumar, "Prediction of diabetes and clustering based on its levels using fuzzy c means algorithm," *Int. J. Sci. Technol. Res.*, vol. 9, no. 2, pp. 3222–3225, 2020.

[13] K. Polat, "Intelligent Recognition of Diabetes Disease via FCM Based Attribute Weighting," *Int. J. Comput. Inf. Eng.*, vol. 10, no. 4, pp. 783–787, 2016.

[14] R. Sanakal and S. T. Jayakumari, "Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine," *Int. J. Comput. Trends Technol.*, vol. 11, no. 2, pp. 94–98, 2014, doi: 10.14445/22312803/ijctt-v11p120.

[15] K. Saravananathan and T. Velmurugan, "Cluster based performance analysis for Diabetic data," *Int. J. Pure Appl. Math.*, vol. 119, no. 16, pp. 399–410, 2018.

[16] H. S. Hongdiyanto, Arnold; Yamlean, Paulina V. Y.; Supriati, "Evaluasi kerasionalan pengobatan diabetes melitus tipe 2 pada pasien rawat inap di RSUD Prof. Dr. R. D. Kandou manado tahun 2013," *Pharmacon J. Ilm. Farm.*, vol. 3, no. 2, pp. 77–86, 2014, doi: 10.36465/jkbth.v17i1.205.

[17] B. Derebew, "Survival analysisof recurrent events: an application to diabetes mellitus patients in the case of menellik II referral hospital," ADDIS ABABA UNIVERSITY, 2016.

[18] N. Terefe, Y. Getachew, and B. Birlie, "Modeling Time-To-Recovery Recovery of Adult Diabetic Patients : a Comparison of Proportional Hazard and Shared Gamma Frailty Models," 2018.

[19] S. M. Shamshirgaran, A. Mamaghanian, A. Aliasgarzadeh, N. Aiminisani, M. Iranparvar-Alamdari, and J. Ataie, "Age differences in diabetes-related complications and glycemic control," *BMC Endocr. Disord.*, vol. 17, no. 1, pp. 1–7, 2017, doi: 10.1186/s12902-017-0175-5.

[20] A. H. Al-Saeed *et al.*, "An inverse relationship between age of type 2 diabetes onset and complication risk and mortality: The impact of youth-onset type 2 diabetes," *Diabetes Care*, vol. 39, no. 5, pp. 823–829, 2016, doi: 10.2337/dc15-0991.

[21] I. A. P. R. DEWI, N. L. P. SUCIPTAWATI, and N. K. T. TASTRAWATI, "Aplikasi Regresi Cox Proportional Hazard Pada Sintasan Pasien Diabetes Melitus," *E-Jurnal Mat.*, vol. 7, no. 3, p. 278, 2018, doi: 10.24843/mtk.2018.v07.i03.p215.

[22] Z. Zhao *et al.*, "Survival of Chinese people with type 2 diabetes and diabetic kidney disease: A cohort of 12-year follow-up," *BMC Public Health*, vol. 19, no. 1, pp. 1–8, 2019, doi: 10.1186/s12889-019-7859-x.

[23] W. Sanusi, Alimuddin, and Sukmawati, "Model Regresi Cox dan Aplikasinya dalam Menganalisis Ketahanan Hidup Pasien Penderita Diabetes Mellitus di Rumah Sakit Bhayangkara Makassar," *J. Math. Comput. Stat.*, vol. 1, no. 1, pp. 62–77, 2018.

[24] I. Juniarti and J. Nugraha, "ESTIMASI FUNGSI TAHAN HIDUP PENDERITA DIABETES PENGOBATAN TERBAIK DARI DUA METODE PENGOBATAN (Data Berdistribusi Eksponensial Dua Parameter Tersensor Tipe-II)," *Pros. Semin. Nas. Pendidik. Mat. 2016 ~*, vol. 1, no. Dm, pp. 292–300, 2016.

[25] A. A. Putri and S. Astutik, "Survival Analysis On The Rate Of Diabetes Mellitus Patient Recovery With Bayesian Methode," vol. 8, no. 1, pp. 268–272, 2017.

[26] Rachmaniyah, Erna, and Saleh, "Analisis Survival dengan Model Accelerated Failure Time Berdistribusi Log-normal," pp. 1–7, 2011.

[27] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, 1979, doi: 10.1109/TPAMI.1979.4766909.

[28] B. Jumadi Dehotman Sitompul, O. Salim Sitompul, and P. Sihombing, "Enhancement Clustering Evaluation Result of Davies-Bouldin Index with Determining Initial Centroid of K-Means Algorithm," *J. Phys. Conf. Ser.*, vol. 1235, no. 1, 2019, doi: 10.1088/1742-6596/1235/1/012015.

[29] A. F. Khairati, A. A. Adlina, G. F. Hertono, and B. D. Handari, "Kajian Indeks Validitas pada Algoritma K-Means Enhanced dan K-Means MMCA," *Pros. Semin. Nas. Mat.*, vol. 2, pp. 161–170, 2019.

[30] S. Syahidatul Helma *et al.*, "Clustering pada Data Fasilitas Pelayanan Kesehatan Kota Pekanbaru Menggunakan Algoritma K-Means," *Puzzle Res. Data Technol. Fak. Sains dan Teknol.*, vol. 1, no. November, p. 4, 2019.