Improving Neural Network Performance with Feature Selection Using Pearson Correlation Method for Diabetes Disease Detection

April Firman Daru[1], Mohammad Burhan Hanif[2], Edi Widodo[3]

[1,2,3]Informatics Engineering, Faculty of Information and Communication, Universitas Semarang, Indonesia
`[1]firman@usm.ac.id,[2]hanifburhan@usm.ac.id,[3]ediwidodo@usm.ac.id`

Abstract - Diabetic or silent killer diseases are an alarming scourge for the world and are classed as serious diseases. In Indonesia, the increase in diabetics occurred by 2% in vulnerable times between 2013 to 2018. This affects all sectors, both medical services and the financial sector. The Neural Network method as a data mining algorithm is present to overcome the burden that arises as an early detection analysis of the onset of disease. However, Neural Network has slow training capabilities and can identify important attributes in the data resulting in a decrease in performance. Pearson correlation is good at handling data with mixed-type attributes and is good at measuring information between attributes and attributes with labels. With this, the purpose of this study will be to use the Pearson correlation method as a selection of features to improve neural network performance in diabetes detection and measure the extent of accuracy obtained from the method. The dataset used is diabetes data 130-US hospital UCI with a record number of 101767 and the number of attributes as many as 50 attributes. The results of this study found that Pearson correlation can improve neural network accuracy performance from 94.93% to 96.00%. As for the evaluation results on the AUC value increased from 0.8077 to 0.8246. Thus Pearson's Correlation algorithm can work well for feature selection on neural network methods and can provide solutions to improved diabetes detection accuracy.

Keywords: Neural Network, Pearson Correlation, Diabetes.

I. INTRODUCTION

Diabetes is known as a silent killer disease caused by a high level of insulin in the blood. The pancreas produced insulin to fight high blood sugar. The abnormal blood sugar level can affect internal organs and neurons in the human body. This thing has become a world concern that makes this disease a serious disease category [1].

According to WHO data, the number of diabetic patients increased to 108 in 1989, and 422 million in 2014 [2]. Diabetic patients in Indonesia also increase up to 2% from 2013 to 2018 [3]. The keep increasing patients become an extra burden for the medics [4]. On the economical side, this disease caused $13,700 or more extra burden for this sector [1]. That is why diabetes disease must be detected accurately [5]. By developing an accurate detection for diabetes, the medical workers can receive primary analysis as anticipation toward disease. Besides that, early analysis can reduce care costs [4].

Previous research on diabetes has been carried out, one of which is Verikas and Bacauskiene. The researchers explained that it is important to predict diabetes disease using the Neural Network approach. However, this approach experienced a problem caused by incorrect attribute selections. To solve this problem, the researchers utilized the Neural Network Feature Selector method to increase the prediction accuracy.

Behloul et al also studied diabetes disease, they examined the performance of the Neural Network which tends to be slow due to high dimensional diabetes data. However, this problem is solved by implementing a Fuzzy algorithm to choose the important attribute. Thus, the accuracy increased from 75.78% to 79.42%.

The next diabetes research by Mehrbackh Nilashi et al that used the Classification Neural Network method to find important factors through preprocessing to increase efficiency and accuracy [6]. Research that discussed the dependency problem in the diabetes data, successfully solve the problem by utilizing Feed Forward Neural Network and Recurrent Neural Network. The result from the research found that the RMSE FNN and RNN are 0.652 and 0.624 with AUC 84.94% and 86.67% [7].

Neural Network algorithm can generalize or identify the non-linear relations in the data. That is why this algorithm is implemented in many applications and researches [8]. However, behind the advantages of Neural networks lies the slow training process [9].

Besides that, Neural networks often experienced slow performance caused by interference in the attribute identification process [10]. To understand which attribute is important and relevant, a proper feature selection is the only way to do. This is one of many important aspects that increase Neural network performance [11].

There is an algorithm called Pearson Correlation that can measure information between attributes and their labels. Besides that, this algorithm capable to handle mixed-type attributes efficiently. Because of that reason, this algorithm is suitable to be used as feature selection in the Neural Network algorithm [12].

Data preprocessing is an important technic to increase the quality of the data. By increasing the quality, the accuracy and efficiency also increase. The algorithm performance is also affected by the quality of the data, which means interferences will occur less than before. The most common problems that occurred in Neural networks are caused by the quality of data.

That is why this research proposed a Neural Network approach to detect diabetes disease with Pearson Correlation to increase the quality of the data. This is important to understand how the Pearson Correlation method can increase the performance of the Neural Network algorithm.

II. METHOD

This research used several steps from data gathering, feature selection using Pearson Correlation, classification using Backpropagation Neural Network, and accuracy evaluations. The used steps in this research will be described as follows:

A. Data Gathering

The data used in this research is a collection of patient data with diabetes disease. The data source of this data is categorized as secondary data since the user data is obtained from existing data [13]. The secondary data is taken from UCI Machine Learning Repository Diabetes Hospital 130-US with 101767 total number of rows [14]. Meanwhile, the total number of attribute data is 50 attributes divided into 49 predictor attributes and 1 label. The 50 attribute parameters presented in the following Table I.

UCI Diabetes Hospital 130-US dataset contains 10 years' worth of data taken between 1999-2008 from 130 hospitals in the US. From 101767 records, only 100 randomly picked records were used in this research to ease the processing stage. agar data. Meanwhile, all 50 features in this research will be completely used and

selected with the Pearson Correlation algorithm in the next step.

B. Implementation Methods

After data gathering, the next step is to select the feature from the randomly picked records with the Pearson Correlation method. In statistics, Pearson Correlation is a method to measure the relation between two variables.

It is based on a data covariance matrix to evaluate the strength of the relationship between two vectors [12]. Meanwhile, to calculate the Pearson Correlation used (1) as follows [15]:

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{\{n \sum x^2 - (\sum x)^2\}\{n \sum y^2 - (\sum y)^2\}}} \quad (1)$$

Where: r correlation coefficient is the product of the total number of variables x and y minus the total number of variables x times the total number of variables y divided by the root of the number of squares of the total number of variables x times the number of squares of the total number of variables y.

After the correlation coefficient value of each feature is obtained, the next step is to sort the correlation coefficient interpretation value from the highest value to the lowest value based on the following Guildford Table II.

From this feature selection process, the dominant features will be selected based on the highest weight calculation that has been calculated using the Pearson Correlation algorithm. After the top dominant weight feature is obtained, the next step is to classify using the Neural Network Backpropagation algorithm.

The Backpropagation Neural Network algorithm uses a binary sigmoid function where values 0 to 1 are the desired output. The main problem of the Backpropagation Neural Network is the iteration uncertainty that must be done so that it takes a relatively long time. These uncertainties affect the value of the epoch process, to produce the desired conditions for iterations. Each of the researchers has a different opinion in determining the value of the parameter. In this study, the stages of the Backpropagation Neural Network algorithm are recognized as follows:

The first step is to initialize the weight by setting the initial value with a random value. The next step is applying the maximum value of epoch, Target Error, and Learning Rate. In this research, the value of epoch is set to 0 and MSE to 1. The second step is to loop the process by one increment with criteria: 1) the value of epoch is less than maximum epoch (epoch

epoch), 2) and MSE more than Target Error (MSE > Target Error). Each pair of elements will be carried out to the learning stage of Feed Forward, Backpropagation, and MSE.

TABLE I
FEATURE LIST AND DESCRIPTIONS

Num	Feature Name	Data Type	Description and Value
1	Encounter ID	Numeric	Unique identifier of an encounter
2	Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other
3	Gender	Nominal	Values: male, female, and unknown/invalid
4	Age	Nominal	Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100)
5	Weight	Numeric	Weight in pounds.
6	Admission type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
7	Discharge disposition	Nominal	An integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
8	Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
9	Time in hospital	Numeric	An integer number of days between admission and discharge
10	Payer code	Nominal	An integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay
11	Medical specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon
12	Number of lab procedures	Numeric	Number of lab tests performed during the encounter
13	Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter
14	Number of medications	Numeric	Number of distinct generic names administered during the encounter
15	Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter
16	Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter
17	Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter
18	Diagnosis 1	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
19	Diagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
20	Diagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
…	…	…	…
27	Readmitted	Nominal	Days to inpatient readmission. Values: "<30" if the patient was readmitted in less than 30 days, ">30" if the patient was readmitted in more than 30 days, and "No" for no record of readmission.

TABLE II
CORRELATION COEFFICIENT INTERPRETATION

Correlation Coefficient	Interpretation
0,9 - 1	Very High
0,7 – 0,899	High
0,4 – 0,699	Enough
0,2 – 0,399	Low
< 0,2	Very Low

The third is to calculate the Feed Forward in every input unit $(x_i, i = 1,2,…,n)$ receive x signal and forward it to all units in the next layer (hidden layer). The fourth step is to sum all units in the hidden layer $(z_j, j = 1,2,…,p)$ with weighted signal input using: $z\_{in}j = v_{0j} + \sum_{i=1}^{n} x_i v_{ij}$. After that, use the activation function to calculate the signal output $z_j = f(z\_in j)$ and send them to all units in the next layer (Output

Layer). The fifth step is to sum the weighted input signal $y_k(k = 1, ..., m)$ with every output unit $y\_in\,j = w_{0k} + \sum_{j=1}^{p} z_j w_{jk}$. After that use the activation function to calculate the output signal $y_k = f(y\_in\,k)$ and send them to the next layer.

The sixth step is to backpropagation all output units $y_k(k = 1, ..., m)$ to receive target pattern related with the input pattern training and calculate the error information with $\delta = (t_k - Y_k).f'(Y\_ink)$. Then the corrected weight is used to fix $\Delta W_{jk} = \alpha\varphi 2_{jk}$ value where $\varphi 2_{jk} = \delta_k Z_j$. After that, calculate the bias correction with $\Delta b2_k = \alpha\beta 2_k$ value where $\beta 2_k = \delta_k$.

The seventh step is to sum the hidden units $Z_j(j = 1,2, ..., p)$ with the delta input from the units in the next layer $\delta\_inj = \sum_{k=1}^{m} \delta 2_k W_{jk}$. Multiply this value with the activation function derivation to calculate the error information $\delta 1_j = \delta_{inj} f'(Z\_inj)$ then calculate the corrected weight $\Delta v_{ij} = \alpha\varphi 1_{ij}$ value where $\varphi 1_{ij} = \delta 1_j X_j$ Calculate the correction bias $\Delta b1_j = \alpha\beta 1_j$ value where $\beta 1_j = \delta 1_j$.

The next step is to fix the output units $Y_k(k = 1,2, ..., m)$ of bias and weight $(j = 0,1,2,..,p)$ with $W_{jk}(new) = W_{jk}(old) + \Delta W_{jk}$s, then all hidden units $Z_j(j = 1,2, ..., p)$ fix bias and weight $(i = 0,1,2, ..., n)$ with $V_{ij}(baru) = V_{ij}(old) + \Delta V_{ij}$.. The last step is to calculate the Mean Square Error (MSE), if MSE input

$\le$ MSE, then the process stop. If not stopped, the process will continue reading from the first data (i=1) until the calculated MSE $\le$ MSE input is fulfilled. In the evaluation stage, the hidden and output layers will be re-calculated. The schematic of the implementation stage of the Backpropagation Neural Network is shown in the following Fig.1.

III. RESULTS AND DISCUSSION

A. Research Data

The used data set consists of 101767 rows and 50 attributes. From 101767 rows, this research only used 100 rows to avoid long processing time (Table III). These 100 rows of data are randomly picked, then divided randomly into two data set with 70% dan 30% ratios. The 70% portion of the data is used as training data, and the rest is used as test data.

From this dataset, the data will be transformed or converted into numeric data. The data transformation is done by sorting ordinal data into numerical form, e.g. Female as 0 and Male as 1 (Table IV).

B. Feature Selection

From calculations using the Pearson Correlation method, the weighted values are obtained as shown in the Table V below:
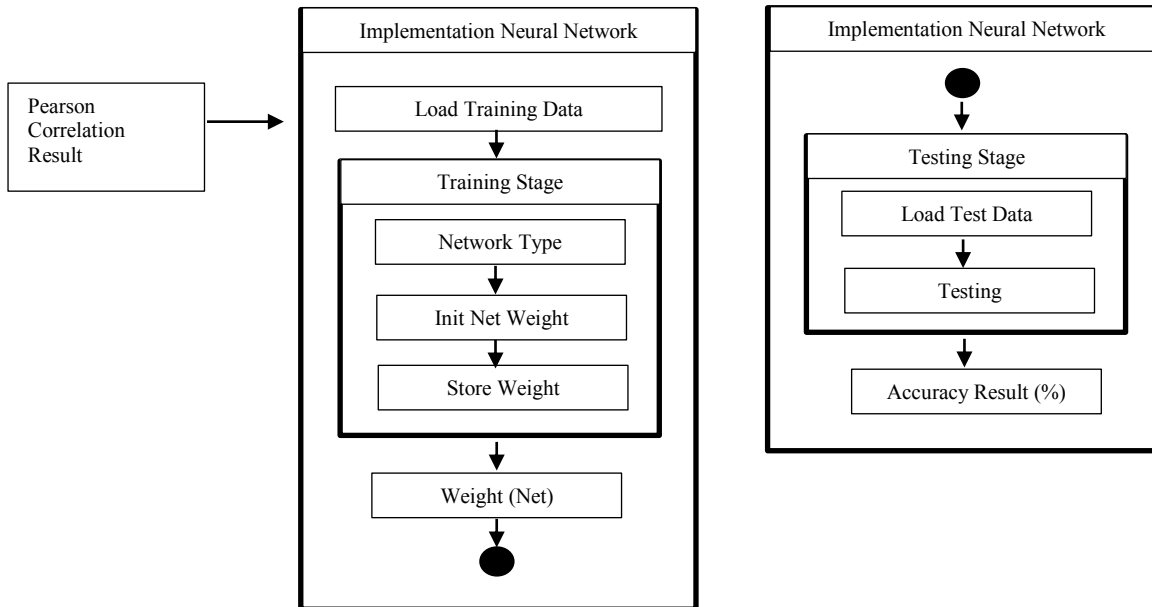

Fig.1 Neural network implementation phase

TABLE III
DIABETIC PATIENTS DATA

No	encounter_id	Patient_nbr	race	gender	age	……	diabetes meds
1	2278392	8222157	Caucasian	Female	[0-10)	.....	No
2	149190	55629189	Caucasian	Female	[10-20)	.....	Yes
3	64410	86047875	African	Female	[20-30)	.....	Yes
4	500364	82442376	Caucasian	Male	[30-40)	.....	Yes
5	16680	42519267	Caucasian	Male	[40-50)	.....	Yes
6	35754	82637451	Caucasian	Male	[50-60)	.....	Yes
…	…	…	…	.	.	.	.
97	42802962.0	19316196.0	Caucasian	Female	[80-90)	.....	No
98	42805614.0	19316448.0	Caucasian	Male	[50-60)	.....	No
99	42808980.0	987381.0	Caucasian	Female	[50-60)	.....	No
100	42809670.0	23724657.0	Caucasian	Male	[80-90)	.....	Yes

TABLE IIIV
DIABETIC PATIENTS DATA TRANSFORMATION

No	encounter_id	Patient_nbr	race	gender	age	……	diabetes meds
1	2278392	8222157	2	1	0	.....	1
2	149190	55629189	2	1	1	.....	0
3	64410	86047875	1	1	2	.....	0
4	500364	82442376	2	0	3	.....	0
5	16680	42519267	2	0	4	.....	0
6	35754	82637451	2	0	5	.....	0
…	…	…	.	.	.	.	.
67	33846834	106467156	2	1	8	.....	0
68	33848646	55265958	2	1	3	.....	1
69	33855372	114415785	2	1	7	.....	0
70	33856296	6919632	2	0	8	.....	0

After obtaining the correlation value of each attribute, the next step is to sort the correlation value from the largest value to the smallest value. Then determine the threshold level of importance (weight) of each of these attributes. Henceforth, attributes that have the same importance (weight) as the threshold or greater will still be used or maintained, but for attributes that have a level of importance or weight value that is smaller or below the threshold value, they will be ignored or will not be used in the process. subsequent calculations.

To determine the threshold, it is done by testing five times, the test results can be seen in the following Table VI. From the experimental data, the highest accuracy lies in the number of attributes as many as 6 attributes with an accuracy value of 96.00%. So that the threshold that will be used is the attribute threshold to 6, which is worth 0.301.

The next step is to reduce the attributes according to the calculation result of Pearson Correlation. The reducing process of attributes is done by threshold limit with value 0.301, any attributes with a lower value will be removed from the data. The final data only consists of 6 attributes as shown in the following Table VII.

C. Evaluation and Validation

To measure the model performance, this research calculates the level of the prediction level using the validation model. By understanding the model performance can help to optimize the parameter and choose the optimal algorithm. Cross-Validation is one of many validation models used in this research. This technic used to validate many training and test subset data repeatedly. Every iteration tests subset data with leftover data as training data.

TABLE V
ATTRIBUTE WEIGHT

No	Attribute	Weight
1	patient_nbr	0.208563381
2	age	0.19734909
3	readmitted	0.176081758
4	glyburide	0.165050817
5	diag_3	0.161492917
6	num_procedures	0.112734179
7	glimepiride	0.11253393
8	number_outpatient	0.109065735
9	A1Cresult	0.090986512
10	admission_type_id	0.090482848
11	number_emergency	0.078335894
12	rosiglitazone	0.078335894
13	troglitazone	0.078335894
14	discharge_disposition_id	0.073947196
15	repaglinide	0.054695081
16	tolbutamide	0.054695081
17	acarbose	0.054695081
18	time_in_hospital	0.045841155
19	diag_2	0.039308495
20	num_medications	0.039246832
21	race	0.036631773
22	gender	0.033794283
23	num_lab_procedures	0

TABLE VI
THRESHOLD LIMIT TEST

No	Attribute	Weight
1	7	93,00 %
2	6	96,00 %
3	5	95,00 %
4	4	93,00 %
5	3	88,00 %

TABLE VII
CHOOSEN ATTRIBUTES

No	Attributes	Weight
1	insulin	1
2	change	0.641250605
3	encounter_id	0.385994862
4	diag_1	0.336798213
5	medical_specialty	0.325478235
6	number_inpatient	0.301011352

The validation result showed a confusion matrix table. This matrix contains prediction results from the system and actual condition. This table also contains the accuracy value, precision, recall, and UAC. The following Table VIII shows the measurement result of the confusion matrix without Pearson Correlation.

1) Accuracy: The following is the accuracy value between the Neural Network only and Neural Network methods + Pearson Correlation

- Accuracy of Neural Network only

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
$$= \frac{92 + 2}{92 + 2 + 1 + 5} * 100$$
$$= 94.93$$

- Accuracy of Neural Network + Pearson Correlation

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
$$= \frac{91 + 5}{91 + 5 + 2 + 2} * 100$$
$$= 96.00$$

2) Precision: The following precision value is the value between Neural Network only and Neural Network + Pearson Correlation

- The precision of Neural Network.

Meanwhile the measurement result of Neural Network with Pearson Correlation was presented at Table IX with its value:

$$Precision = \frac{TP}{TP + FP}$$
$$= \frac{92}{92 + 1}$$
$$= 0.989$$

TABLE VIII
CONFUSION MATRIX NEURAL NETWORK

		Actual Class	
		+ (positive)	- (negative)
Prediction class	+ (positive)	92	1
	- (negative)	5	2

TABLE IX
CONFUSION MATRIX NEURAL NETWORK + PEARSON CORRELATION

		Actual Class	
		+ (positive)	- (negative)
Prediction class	+ (positive)	91	2
	- (negative)	2	5

- The precision of Neural Network + Korelasi Pearson

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

$$= \frac{91}{91 + 2}$$

$$= 0.9784$$

3) Recall: The following is the recall value between Neural Network only and Neural Network + Pearson Correlation methods

- Recall of Neural Network

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

$$= \frac{92}{92 + 5}$$

$$= 0.9484$$

- Recall of Neural Network + Pearson Correlation

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

$$= \frac{91}{91 + 2}$$

$$= 0.9784$$

4) AUC (Area Under Curve): The following is the AUC value of the Neural Network and Neural Network + Pearson Correlation methods

- AUC of Neural Network

$$AUC = \frac{1 + TPrate - FPrate}{2}$$

$$= \frac{1 + 0.9484 - 0.333}{2}$$

$$= 0.8077$$

- AUC of Neural Network + Pearson Correlation

$$AUC = \frac{1 + TPrate - FPrate}{2}$$

$$= \frac{1 + 0.9784 - 0.2857}{2}$$

$$= 0.8246$$

Comparison of the results of the Neural Network method without the Pearson Correlation with the Neural Network method with Pearson Correlation which is presented in the following Table X.

TABLE X
RESULT COMPARISON BETWEEN SCENARIOS

Methods	Accuracy	AUC
Neural Network	94.93%	0.8077
Neural Network + Pearson Correlation	96.00%	0.8246

From this Table X, it can be seen that the results of the Neural Network method test have an accuracy value of 94.93% with an AUC value of 0.8077. While the test results of the Neural Network method using the Pearson Correlation method as feature selection get the results of an accuracy value of 96.00% with an AUC value of 0.8246.

IV. CONCLUSION

Through this research, it can be seen that the accuracy value and the AUC value from the experiment using the Neural Network method based on Pearson Correlation have an increase compared to the results obtained from the Neural Network method experiment alone. This can be seen from the results of increasing the accuracy of the method by 1.07%. The increase in accuracy is calculated from the experimental results of the Neural Network method alone which only has an accuracy value of 94.93% to 96.00% when experimenting with the Neural Network method using the person correlation method as feature selection. Thus it can be concluded that the Pearson Correlation algorithm has succeeded in improving the performance of the Backpropagation algorithm. Neural Network by selecting important features on UCI Diabetes Hospital 130-US data. The feature selection carried out by the Pearson correlation algorithm is based on computational calculations that do not pay attention to the important features that are used in the medical world. Therefore, this research can still be developed, one of which is by choosing features that suit the needs of the medical world by working with existing medical experts. However, overall this research is following and in line with the aims of the researcher, which is to only measure the level of achievement of the algorithm used.

REFERENCES

[1] N. Kumar and M. Abedin, "Comparative Approaches for Classification of Diabetes Mellitus Data: Machine Learning Paradigm," *Comput. Methods Programs Biomed.*, 2017.

[2] M. Chan, "Global Report On Diabetes," *World Heal. Organ.*, 2013.

[3] K. Kesehatan, "HARI DIABETES SEDUNIA TAHUN 2018." Kesehatan, Infodatin pusat data dan informasi kementrian RI, 2019.

[4]   M. Rigla, B. Pons, and M. Elena, "A web-based clinical decision support system for gestational diabetes : Automatic diet prescription and detection of insulin needs," *Int. J. Med. Inform.*, 2017.

[5]   F. Mansourypoor and S. Asadi, "Development of a Reinforcement Learning-based Evolutionary Fuzzy Rule-Based System for Diabetes Diagnosis Fatemeh," *Comput. Biol. Med.*, 2017.

[6]   N. Mehrbakhsh, I. Othman, D. Mohammad, A. Hossein, and S. Leila, "Accuracy Improvement for Diabetes Disease Classification A Case on a Public Medical Dataset," *Fuzzy Inf. Eng.*, 2017.

[7]   S. Kang, "Personalized prediction of drug efficacy for diabetes treatment via patient-level sequential modeling with neural networks," *Artif. Intell. Med.*, pp. 2–7, 2018.

[8]   C. J. Bennett, R. A. Stewart, and J. W. Lu, "Forecasting low voltage distribution network demand profiles using a pattern recognition based expert system," *Energy*, vol. 67, pp. 200–212, 2014.

[9]   G. C. Liao, "Hybrid improved differential evolution and wavelet neural network with load forecasting problem of air conditioning," *Int. J. Electr. Power Energy Syst.*, vol. 61, pp. 673–682, 2014.

[10] C. J. Lu, T. S. Lee, and C. M. Lian, "Sales forecasting for computer wholesalers: A comparison of multivariate adaptive regression splines and artificial neural networks," *Decis. Support Syst.*, vol. 54, no. 1, pp. 584–596, 2012.

[11] F. C. Cruz, E. F. S. Filho, M. C. S. Albuquerque, I. C. Silva, C. T. T. Farias, and L. L. Gouvêa, "Efficient feature selection for neural network based detection of flaws in steel welded joints using ultrasound testing," *Ultrasonics*, vol. 73, pp. 1–8, 2017.

[12] Y. Mu, X. Liu, and L. Wang, "A Pearson's correlation coefficient based decision tree and its parallel implementation," *Inf. Sci. (Ny).*, vol. 435, pp. 40–58, 2018.

[13] Sugiyono, *Metode Penelitian Kuantitatif, kualitatif dan R & D*. Bandung: Alfabeta, 2012.

[14] B. Strack *et al.*, "Diabetes Hospital," 2020. https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008 (accessed Sep. 23, 2020).

[15] H. Serge, "Analisis Regresi Dan Korelasi Antara Seleksi Bobot Badan Fase Starter Terhadap Produksi Ayam Ras Petelur Tipe Medium," no. March, 2016.