Prognosis of Diabetes Mellitus with Transfer Learning-Based Naïve Bayes Method

A Andriani1,* and S Hartati1

1 Universitas Gadjah Mada, Sekip Utara Bulaksumur, Yogyakarta 55281, Indonesia
2 Universitas Bina Sarana Informatika, Jl.Kramat Raya No.98, Jakarta 10450, Indonesia

* anikandriani@mail.ugm.ac.id

Abstract. Early detection of diabetes mellitus (DM) prognosis against several diseases cannot be done medically in the short term. Supervised Learning method can be used to classify diabetic patient data to detect the prognosis of DM. There are several problems in the classification of patient medical record data. First, medical record data is not always good, namely structured and complete. Second, DM patient medical record data from various sources do not necessarily have the same parameters. The impact of supervised learning on a dataset of DM patients is not necessarily applicable to a dataset of DM patients from different sources, so that the learning outcomes are not long-lived learning. The proposed method in this study is data classification with the Naïve Bayes method based on transfer learning by applying learning outcomes in the source domain to the target domain as a starting point for learning in the target domain. The method applied uses parameter-based transfer learning so that it can be used to overcome parameter differences in two different datasets.

1. Introduction
Current society’s lifestyle and diet have an impact on the increasing of diabetics’ number. Therefore, the problem of diabetes still becomes a challenge in studies, one of them is for handling the impact of diabetes itself on several comorbidities [1]. Diabetics are at risk for several diseases such as Cardiovascular [2], Pancreatic Cancer [3], Diabetic Nephropathy [4][5][6], Diabetic Rethinopathy [7][8], Stroke [9], Heart Failure [10], dan Glaucoma [11]. Early prognosis in patients of type 2 Diabetes (T2DM) who are diabetics with the highest percentage is required to prolong the average survival of those patients [2]. Prognosis can be made from patient medical records and patient activity data to find the prevalence of diabetes against several diseases [12]. Although Diabetes Mellitus has a high prevalence especially in diseases such as cardiovascular, in medical studies on cardiovascular patients, Diabetes Mellitus has not shown its effect on the disease in the short and medium-term outcomes. Moreover, the prognosis of Diabetes Mellitus against Renal Failure has not shown a strong prevalence in the short term [13]. The impact cannot be detected early on the risk of Diabetes Mellitus patients to the accompanying diseases.

Machine Learning is employed to construct systems that can learn from the existing data [14]. The methods in Machine Learning can be used in diabetes prognosis cases with the learning from patient medical record data. In Machine Learning, data collection becomes the important preliminary stage. The main obstacle in data collection is that the data can come from various domains and the differences in their distribution result in the restrictiveness of fault extraction and classification processes in medical
data which often lead to fault diagnosis [15]. In addition, data noise and incomplete data increase difficulties in learning [16]. Therefore, an appropriate framework is required in activities like diagnosis so that the various kinds of knowledge from dataset can be enhanced so that the interaction between the employed method and knowledge basis in data is achieved [17]. Based on the aforementioned problems, it needs more attention on the appropriate data analysis in data preprocessing stage.

In the case of Diabetes Mellitus prognosis with several diseases that are included in the category of accompanying diseases, this study requires a Machine Learning Method with multiclass classification techniques. Several methods can be used for multiclass classification, one of which is the Bayesian Learning method. In multiclass classification research in the case of modeling and learning patterns of brain activity, this method has an efficient and excellent classification performance [18].

In addition to the problems in data pre-processing stage, another dilemma that often arises in medical record data is that the difference of data from one hospital to the other hospital can be different [19]. This causes the results of learning on a diabetic dataset taken from Hospital A not necessarily applicable to the diabetic dataset at Hospital B. The learning outcomes of a method on a dataset are not long-lived learning since they can be only applied in a case of a dataset. Therefore, it needs learning that yields task which is applicable to new learning on the same case. The concept of transfer learning enables the learning outcome task on dataset A which is the source domain can be reapplied for the learning on dataset B which is the target domain [20]. Table 1 and Table 2 are examples of diabetic patient datasets taken from the UCI Machine Learning dataset and the PIMA Indian Dataset. The sample data shows that the features of diabetic patient dataset are not necessarily the same.

**Table 1.** The first dataset example.

| Age | Sex | Plasma Creatine | Diastolic blood pressure (mmHg) | Tricep skinfold thickness (mm) | 2-hours serum insulin (mmU/ml) | Body Mass Index | The Function of the diabetes pedigree |
|-----|-----|-----------------|---------------------------------|--------------------------------|--------------------------------|----------------|----------------------------------|
| 50  | M   | 148             | 72                              | 35                             | N/A                            | 33.60          | 0.63                             |
| 32  | W   | 183             | 64                              | N/A                            | N/A                            | 23.30          | 0.67                             |
| 33  | W   | 137             | 40                              | 35                             | 168                            | 43.10          | 2.29                             |
| 26  | M   | 78              | 50                              | 32                             | 88                             | 31.00          | 0.25                             |
| 34  | W   | 168             | 74                              | N/A                            | N/A                            | 38.00          | 0.54                             |

**Table 2.** The second dataset example.

| Age | Sex | Cholesterol | Blood Pressure | BMI | Diabetes Duration | Insulin Requirements | Plasma Creatinine | Diabetes Pedigree |
|-----|-----|-------------|----------------|-----|-------------------|----------------------|-------------------|------------------|
| 41  | M   | 200         | 90             | 39.80| 2                 | N/A                  | 196               | 0.45             |
| 41  | M   | 378         | 70             | 31.10| 2                 | 115                  | 125               | 0.21             |
| 43  | W   | N/A         | 76             | 39.40| 1                 | N/A                  | 147               | 0.26             |
| 57  | W   | N/A         | 82             | 22.20| 4                 | 110                  | 145               | 0.25             |

Based on the problems above, this study aims to construct a framework for long-term prognosis study in patients of type 2 Diabetes in which the proposed method is expected to be able to accommodate the problems in either data pre-processing stage or data learning stage. The problems in data pre-processing stage and the problems in data learning stage are how the learning outcomes on a diabetes dataset can be applied in the new dataset.

2. **Dataset**

This study uses medical record data of Diabetes Mellitus patients. The data is obtained from Dr. Sardjito Hospital data, Yogyakarta and Gadjah Mada University Academic Hospital. The patient medical record
data employed are in the forms of laboratory test result data and patient activity data. The medical record data used are patient data from diagnosed diabetes to the emergence of comorbidities which are the impacts of Diabetes Mellitus.

3. Method
In this study, some methods are proposed which are explained in several research stages to create data classification for Diabetes Mellitus prognosis. Proposed research stages comprising Data Collection, Data Preparation, Supervised Learning, Transfer Learning, and Evaluation Stages.

3.1. Data Collection
The dataset of Diabetes Mellitus patients from patient medical record data is transformed based on the Guidelines of Management and Prevention of Diabetes Mellitus by Indonesian Society of Endocrinology (PPDM PERKENI) for determination of comorbidities and the risk diseases of Diabetes Mellitus in which dataset becomes output. Then, the dataset of diabetes mellitus patients is divided into two parts, namely as Source Domain and Target Domain.

3.2. Data Preparation
The aim of this stage is to decrease the complexity of dataset in the real world [21]. This is because the data collected contains a lot of data noise and incomplete data so that it can increase the difficulty in the Machine Learning process [16]. Therefore, in the Data Cleaning and Quality stage, feature selection [22], handling imbalance dataset [23] and imputation missing value [24].

3.3. Supervised Learning
One type of learning in Machine Learning is Supervised Learning. It’s usually used to map input to the desired output in a problem solving such as classification and regression [25]. Machine learning tools research is used for diabetes prognosis, namely predicting the long-term impact of diabetes patients on several diseases.

3.4. Evaluation
The evaluation stage is used to measure the performance of the method used. Performance measurement can be done by measuring the level of accuracy. Confusion Matrix is one of the tools that can be used to measure the accuracy level of learning with Machine Learning method [26].

4. Proposed Method
In this study, several methods are proposed which are described in several stages of research to classify data for the prognosis of Diabetes Mellitus. Figure 1 shows the stages of the proposed research consisting of the stages of Data Collection, Data Preparation, Supervised Learning, Transfer Learning, and Evaluation.
The data collection stage is the stage for constructing the dataset. At this stage, the dataset is divided into two, namely the source domain and the target domain. At the data preparation stage, it is carried out feature selection dan handling missing value. A good feature in dataset becomes a good class predictor in the use of learning method [27]. There are many ways to assess features, one of them is by Information Gain. This method has an ability to select attributes with ranking system [22]. In Information Gain, it uses the calculation of entropy value (en) to measure the score of each attribute towards the value on the objective variable [28].

The problem of missing value is very common in data analysis. The missing value percentage less than 1% is not a problem, the missing value percentage of 1%-5% should be managed, while the missing value percentage of 5%-15% in dataset needs serious handling with sophisticated method. Moreover, missing value percentage more than 15% in dataset highly affects the performance of the applied learning algorithm [29]. In the first and second dataset examples shown in Tables 1 and 2, there are some missing value data indicated by N/A. This study proposes k-Nearest Neighbor method for the problem of incomplete data. In some studies on missing value imputation, the k-Nearest Neighbor method is proven to have a good performance on those case. This method works by finding a number of k data object or the patterns of all existing data that are closest to the input pattern. This study uses the classification method. For Diabetes prognosis using some output, the classification model used is multiclass classification model. The classification learning method used is Bayesian Learning. Bayesian Learning is able to work efficiently on uncertainty data such as medical data.

Not all of the features in dataset are suitable for the model used. The clinical assessment of various hospitals on the handling of a disease is often different. In transfer learning, there are four categories of transfer learning, namely: (1) instance transfer, in which in transfer learning data in source domain should be usable for learning in target domain; (2) feature representation transfer, in transfer learning, it aims to learn general feature representation that is suitable for target domain so that the difference between them is reduced; (3) relational knowledge transfer, is done when the data in source domain and target domain are similar since the knowledge that will be transferred is the relationship between data; (4) parameter transfer, the source domain and the target domain share prior parameter from the model in which the knowledge that will be transferred is in it [19]. The objective of transfer learning in this study is to extract knowledge from the source domain in form of source task and apply it in the learning process in target domain. The learning in DT uses Transfer Learning-based Naive Bayes method called Transfer Naive Bayes Classifier (TNBC). This proposed method uses ST to help the classification in DT using all data in DT and increase knowledge from the previous yielded model.

5. Proposed Method Transfer Learning
The learning outcomes of Supervised Learning using Domain Source are then used for learning in the Target Domain through the transfer learning method. The learning result source task in the source domain is used as the starting point of learning in the target domain.
**Figure 2.** Proposed transfer learning approach.

Transfer learning parameters are an approach that can be applied in this study. This is because the main objective is to find common parameters so that the resulting model of learning in the source domain can be applied to the target domain dataset. The proposed transfer learning method consists of several stages according to the steps in Figure 2. The stages of transfer learning begin with the adjustment of features in the Domain Source (DS) and Domain Target (DT). The same features on the DS will be used in the Adaptation Model in addition to using all the features that the DT has. The results of adjusting the DS and DT features can then be used.

6. Result

This study is the part of dissertation study. The research stage is still in the early stage, that is, data collection and dataset construction. Novelties proposed in this research is the prognosis of Diabetes Mellitus model framework by proposing long-lived learning methods. Naive Bayes approach using parameters-based learning transfer where there is a stage of customization features so that the proposed method can be applied to new datasets diabetic patients with different features.

7. Conclucion

The problems that frequently occur in medical data such as incomplete, noise, and unstructured data become one of the obstacles that impacts on learning outcomes such as classification. Moreover, the number of attributes used for classification also affected on the result of classification. The attribute which doesn’t have high prevalence on the output of classification learning, can impact on the performance of classification result. Accordingly, the feature selection technique and missing value imputation are required in the data preparation stage. Next, in the classification stage with traditional Machine Learning method is not long-lived learning, the source task of learning outcomes can only be applied in the dataset with similar feature and data. The application of Transfer learning in the classification in Target Domain implements knowledge from modeling output in Source Domain so that the source task of modeling outcome in Source Domain can be applied in the Target Domain in the form of database with additional feature. By the application of transfer learning, the learning outcomes can be long-lived learning.

References

[1] Y. Goyal, A. K. Verma, D. Bhatt, A. H. Rahmani, Yasheshwar, and K. Dev, “Diabetes: Perspective and challenges in modern era,” *Gene Reports*, vol. 20, no. June, pp. 1–6, 2020.

[2] M. J. Holzmann *et al.*, “Long-term prognosis in patients with type 1 and 2 diabetes mellitus after coronary artery bypass grafting,” *J. Am. Coll. Cardiol.*, vol. 65, no. 16, pp. 1644–1652, 2015.

[3] S. Mizuno *et al.*, “Diabetes is a useful diagnostic clue to improve the prognosis of pancreatic cancer,” *Pancreatology*, vol. 13, no. 3, pp. 285–289, 2013.

[4] G. Andrésdóttir *et al.*, “Improved prognosis of diabetic nephropathy in type 1 diabetes,” *Kidney Int.*, vol. 87, no. 2, pp. 417–426, 2015.

[5] R. Meza-Palacios, A. A. Aguilar-Lasserre, E. L. Ureña-Bogarín, C. F. Vázquez-Rodríguez, R. Posada-Gómez, and A. Trujillo-Mata, “Development of a fuzzy expert system for the nephropathy control assessment in patients with type 2 diabetes mellitus,” *Expert Syst. Appl.*, vol. 72, no. 1, pp. 335–343, 2017.

[6] M. Masulli *et al.*, “Measurement of the intrarrenal arterial resistance index for the identification and prediction of diabetic nephropathy,” *Nutr. Metab. Cardiovasc. Dis.*, vol. 19, no. 5, pp. 358–364, 2009.

[7] L. Qiao, Y. Zhu, and H. Zhou, “Diabetic Retinopathy Detection Using Prognosis of Microaneurysm and Early Diagnosis System for Non-Proliferative Diabetic Retinopathy Based on Deep Learning Algorithms,” *IEEE Access*, vol. 8, pp. 104292–104302, 2020.

[8] G. Garberg, M. Lövestam-Adrian, S. Nasic, and K. B. Boström, “The prognosis of diabetic retinopathy in patients with type 2 diabetes since 1996-1998: the skaraborg diabetes register.”
[9] S. M. “Prognostic Neural Network Model for Diabetic Risks Prediction Department of Computer Applications,” pp. 392–395, 2012.

[10] E. S. J. Tan et al., “Effect of Diabetes Mellitus on Cardiac Resynchronization Therapy and to Prognosis in Heart Failure (from the Prospective Evaluation of Asian With Cardiac Resynchronization Therapy for Heart Failure Study),” Am. J. Cardiol., vol. 124, no. 6, pp. 899–906, 2019.

[11] R. N. Alloreurung, J. S. M. Saerang, and L. M. Rares, “Prevalensi Glaukoma Akibat Diabetes Melitus Di Poliklinik Mata Rsup Prof. Dr. R. D. Kandou Manado,” e-CliniC, vol. 3, no. 3, pp. 3–6, 2015.

[12] A. Sarwar and V. Sharma, “Comparative analysis of machine learning techniques in prognosis of type II diabetes,” AI Soc., vol. 29, no. 1, pp. 123–129, 2014.

[13] D. Schewel et al., “Impact of interaction of diabetes mellitus and impaired renal function on prognosis and the incidence of acute kidney injury in patients undergoing transcatheter aortic valve replacement (TAVR),” Int. J. Cardiol., vol. 232, pp. 147–154, 2017.

[14] J. Bell, Jason Bell-Machine Learning_ Hands-On for Developers and Technical Professionals-Wiley (2014). Indianapolis: John Willey and Sons, 2015.

[15] P. Ma, H. Zhang, W. Fan, and C. Wang, “A diagnosis framework based on domain adaptation for bearing fault diagnosis across diverse domains,” ISA Trans., no. xxxx, 2019.

[16] X. Lai, X. Wu, L. Zhang, W. Lu, and C. Zhong, “Imputations of missing values using a tracking-removed autoencoder trained with incomplete data,” Neurocomputing, vol. 366, pp. 54–65, 2019.

[17] Q. Zhou and H. Zeng, “A framework for integrated system of fault diagnosis in oil equipments based on neural networks,” 2012 3rd Int. Conf. Syst. Sci. Eng. Des. Manuf. Informatiz. ICSEM 2012, vol. 1, pp. 14–17, 2012.

[18] L. He, D. Hu, M. Wan, Y. Wen, K. M. Von Deneen, and M. C. Zhou, “Common Bayesian Network for Classification of EEG-Based Multiclass Motor Imagery BCI,” IEEE Trans. Syst. Man, Cybern. Syst., vol. 46, no. 6, pp. 843–854, 2016.

[19] G. Wang, G. Zhang, K. S. Choi, K. M. Lam, and J. Lu, “Output based transfer learning with least squares support vector machine and its application in bladder cancer prognosis,” Neurocomputing, no. xxxx, pp. 1–14, 2020.

[20] Z. Chen and B. Liu, “Lifelong Machine Learning,” Synth. Lect. Artif. Intell. Mach. Learn., vol. 10, no. 3, pp. 1–27, 2016.

[21] S. Roshan and S. Asadi, “Improvement of Bagging performance for classification of imbalanced datasets using evolutionary multi-objective optimization,” Eng. Appl. Artif. Intell., vol. 87, no. October 2019, p. 103319, 2020.

[22] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, and F. Herrera, “A survey on data preprocessing for data stream mining: Current status and future directions,” Neurocomputing, vol. 239, pp. 39–57, 2017.

[23] S. Lefkovits and L. Lefkovits, “Gabor Feature Selection Based on Information Gain,” Procedia Eng., vol. 181, pp. 892–898, 2017.

[24] S. E. Roshan and S. Asadi, “Improvement of Bagging performance for classification of imbalanced datasets using evolutionary multi-objective optimization,” Eng. Appl. Artif. Intell., vol. 87, no. October 2019, p. 103319, 2020.

[25] E. T. Capariño, A. M. Sison, and R. P. Medina, “Application of the modified imputation method to missing data to increase classification performance,” 2019 IEEE 4th Int. Conf. Comput. Commun. Syst. ICCCS 2019, pp. 134–139, 2019.

[26] S. Suthaharan, Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, vol. 36. 2016.

[27] O. O. Aremu, R. A. Cody, D. Hyland-Wood, and P. R. McAree, “A relative entropy based feature selection framework for asset data in predictive maintenance,” Comput. Ind. Eng., vol. 145, no. July 2019, p. 106536, 2020.
[28] S. Jadhav, H. He, and K. Jenkins, “Information gain directed genetic algorithm wrapper feature selection for credit rating,” Appl. Soft Comput. J., vol. 69, pp. 541–553, 2018.

[29] E. Acuña and C. Rodriguez, “The Treatment of Missing Values and its Effect on Classifier Accuracy,” Classif. Clust. Data Min. Appl., no. 1995, pp. 639–647, 2004.

Acknowledgment
Special thanks to the Research Directorate of Universitas Gadjah Mada for funding this research.