Prediction Model for Type 2 Diabetes using Stacked Ensemble Classifiers

Norma Latif Fitriyani
Department of Industrial and Systems Engineering
Dongguk University
Seoul, Republic of Korea
norma@dongguk.edu

Muhammad Syafrudin
Department of Industrial and Systems Engineering
Dongguk University
Seoul, Republic of Korea
udin@dongguk.edu

Ganjar Alfian
Industrial AI Research Center, Nano Information Technology Academy
Dongguk University
Seoul, Republic of Korea
ganjar@dongguk.edu

Agung Fatwanto
Informatika
Universitas Islam Negeri Sunan Kalijaga
Yogyakarta, Indonesia
agung.fatwanto@uin-suka.ac.id

Syifa Latif Qolbiyani
Pengembangan Masyarakat Islam
Universitas Islam Negeri Walisongo
Semarang, Indonesia
syifalatifqolbiyani@gmail.com

Jongtae Rhee
Department of Industrial and Systems Engineering
Dongguk University
Seoul, Republic of Korea
jtrhee@dongguk.edu

2020 International Conference on Decision Aid Sciences and Application (DASA) | 978-1-7281-9677-0/20/$31.00 ©2020 IEEE | DOI: 10.1109/DASA51403.2020.9317090

Abstract—**Diabetes is the number one of major causes of death globally. Undetected and untreated diabetes causes serious issues and the individuals with diabetes are at high risk for complication. Thus, an early diabetes prediction is necessary to help the individuals preventing dangerous conditions at the early stage. This study proposed a prediction model to offer early prognostication of type 2 diabetes. The proposed model incorporates isolation forest and synthetic minority oversampling-tomek link technique to detect as well as remove the outlier data, and balance the data distribution, respectively. The stacked ensemble classifiers are the used learn and predict type 2 diabetes at an early stage. We used three publicly available datasets to evaluate the performance of proposed model as compared to other models such as multi-layer perceptron, support vector machines, decision tree, and logistic regression. We applied 10-fold cross-validation and obtain four performance metrics such precision, recall, f-measure, and accuracy. The experimental results show that the proposed model outperformed other models, achieving accuracy up to 93.18%, 98.87%, and 96.09% for dataset I, II, and III, respectively. It is expected that the early diabetes prediction could help the individuals on taking precautions once type 2 diabetes is detected.**

Keywords—diabetes, classification, stacked ensemble, outlier, unbalanced data

I. Introduction

Type 2 diabetes is a metabolic disorder disease caused by body's inability to absorb its produced insulin [1]. Individuals with diabetes are at higher risk for stroke and death [2]. However, by detecting and treating diabetes at the early stage can efficiently prevent the complication [3]. The number of individuals with diabetes are projected to increase up to 220 million in 2030 [4] which will eventually be burdening healthcare systems [5].

As the increasing awareness of diabetes risk, recent studies have reported that machine learning algorithms can be used for early diabetes detection [6, 7] based on present state, and hence supporting them on taking early precautions. Stacked ensemble learning is one of classification method which combines two or more classifier to reduce variance and bias; thus, improving the performance of model [8, 9]. Previous studies have reported the application of stacked ensemble in predicting diabetes [10]. However, the performance of classification model could be affected by outlier and imbalanced data. Isolation forest (iForest)-based outlier detection and data balancing based on synthetic minority oversampling-tomek link technique (SMOTETomek) have been positively reported on improving the model's performance [11, 12].

Thus, the prediction model for type 2 diabetes is proposed in this study by utilizing the combination of iForest and SMOTETomek to enhance the performance of stacked ensemble classifiers. We used three publicly available datasets to evaluate the performance of proposed model as compared to other models such as multi-layer perceptron (MLP), support vector machine (SVM), decision tree (DT), and logistic regression (LR). We applied 10-fold cross-validation and obtain four performance metrics such precision, recall, f-measure, and accuracy. It is expected that the early diabetes prediction could help the individuals on taking precautions once type 2 diabetes is detected.

II. Methodology

This detailed of proposed prediction model for type 2 diabetes is described in this section. Fig. 1 showed the proposed prediction model for type 2 diabetes which comprises of several steps: iForest, SMOTETomek, and stacked ensemble classifiers. The detailed datasets, steps descriptions, and performance metrics are shown in the next subsections. Finally, the performance comparison of the proposed model with other classification models are presented in the results and discussions section.

A. Datasets

We employed three datasets to examine how early type 2 diabetes can be identified by employing classification learning models. We categorized dataset from Chinese, Japanese, and Iranian diabetes as dataset I, II, and III, respectively. Our prediction model used those datasets and projected to produce a general and robust model. Chen et al., (2018) provided dataset I which was obtained from a database established by

This work was supported by Cooperative Research Program of Center for Companion Animal Research (Project No. PJ0139862020), Rural Development Administration, Republic of Korea.

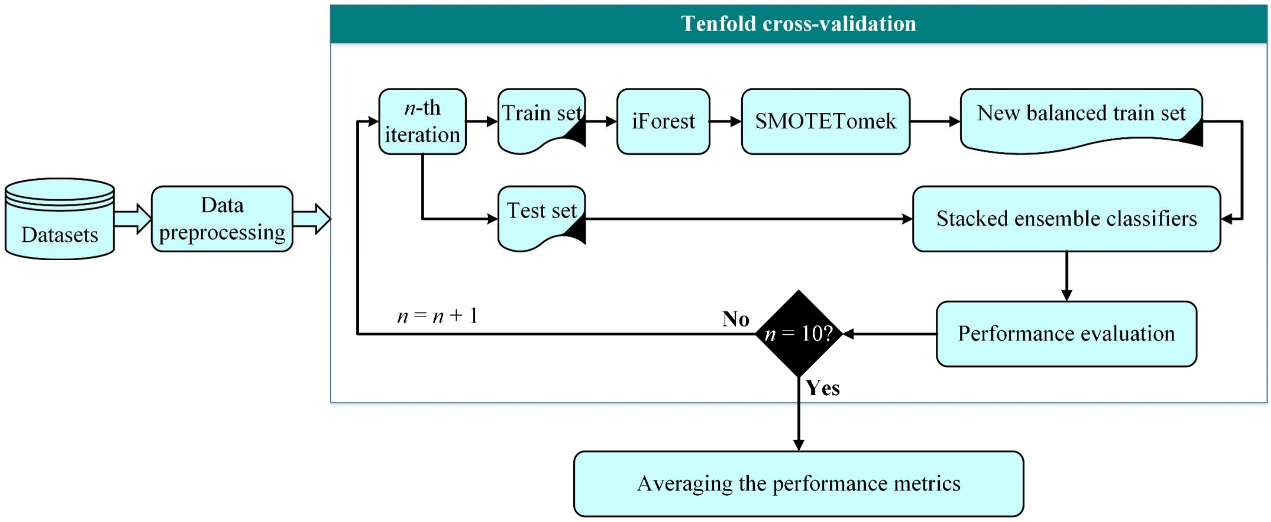

Fig. 1. The flow diagram of the proposed model for type 2 diabetes prediction.

the Rich Healthcare Group in China, which comprises all medical records who obtained a health check from 2010 to 2016 [13]. The original dataset comprised 211833 patients, vary between 20 to 99 years old, with ratio of males and females are 116123 and 95710, respectively. The original number of attributes was 22, however after data preprocessing, we only used 11 selected attributes such as *age, gender, site, height, weight, bmi, sbp, dbp, chol, trig,* and *class*. The value of *class* attribute is binary, where '1' and '0' represents a positive and negative for diabetes, respectively.

Okamura et al., (2018) provided dataset II which is based on longitudinal study of a medical assessment program at Murakami Memorial Hospital in Gifu, Japan [14]. The original dataset consists of 15464 participants and 30 attributes, with individuals ranging from 22 to 70 years old. After data preprocessing, we selected the most significant attributes such as *sex, age, ethanol, fattyliver, bmi, wc, wc9080, bmi25, obesity, followup, ALT, AST, weight, exercise, GGT, HDL-mgdl, chol-mgdl, trig-mgdl, HbA1c, alcohol, smoking, sbp, dbp,* and *class*. The target output *class* is binary, where '1' and '0' means a positive and negative for diabetes, respectively.

Finally, Mozaffary et al. (2016) provided the dataset III which was gathered from Iranians in a population-based study with representative sample of an urban population of Tehran, Iran [15]. The original dataset consists of 3981 participants and 19 attributes such as *sex, age, bmi, wc, changed-fpg, fpg, past-cpg, 2hpast-cpg, trig, chol, ldl, drug, hyper-drug, edu, smoking, trig-hdl, chol-hdl, family-history,* and *class*. The value of *class* attribute is either 0 or 1 for positive and negative of diabetes, respectively.

We anticipate that by employing the three datasets described above (I, II, and III), the proposed prediction model can foresee the diabetes and provide general and accurate model.

B. iForest

We utilized iForest to find and remove the outlier data from type 2 diabetes training sets. Practically, an ensemble of isolated trees for each data were created by iForest, where outliers are defined as records with short length in the isolated trees [16]. Then, the isolated trees is recursively formed by separating the data until each of records is isolated or certain tree height is reached. The tree height can be calculated as follows:

$$tree\ height = ceiling(log_2(maxSample). \quad (1)$$

In iForest, two parameters need to be defined, they are maximum sample size (*maxSample*) and number of trees (*numTree*). The *maxSample* needed to be kept small so that the iForest can perform well. Higher *maxSample* will reduce the ability of iForest on isolating the outlier because normal data can affect the isolation process. Based on different experimental setups, we discovered that *maxSample* and *numTree* is around ten percent (10%) of total data size and 100, respectively. The iForest was realized in python V3.6 and utilized Scikit-learn library V0.20.2 [17]. Finally, the detected outliers from training data are removed and the rest of data are then used for the next analysis.

C. SMOTETomek

The original distributions of class are imbalanced, and we utilized data balancing method called SMOTETomek to solve this issue. For all dataset, the minority class are the participants who identified as type 2 diabetes (positive). First, SMOTE does the over-sampling of the minority class to arbitrarily create records and expanding minority class records, and Tomek does the under-sampling a class to eliminate noise whereas preserving the balanced class distributions. The integration of SMOTE and Tomek (SMOTETomek) offered better outcomes compared either alone [18]. Thus, we employed SMOTETomek technique to solve the imbalance distributions of class. We implemented the SMOTETomek by utilizing Imbalanced-learn V0.4.3 [19]. The goal of classification model is to reduce errors during learning; hence with the balanced training dataset, we expect that the model accuracy will be enhanced.

D. Stacked Ensemble Classifiers

Stacked ensemble learning is one of classification method which combines two or more classifier to reduce variance and bias; thus, improving the performance of model [8, 9]. We adopted stacked ensemble learning with MLP, SVM, and DT for the first layer and logistic regression (LR) for the second layer. The stacked ensemble classifiers structure is shown in Fig. 2 and summarized as following:

- Each classifier in the first layer such as MLP, SVM, and DT are applied to the train set and the prediction outputs are used as the new train set for the second layer.
- The second layer classifier (LR) is the used to learn from the new train set and obtain the final prediction.

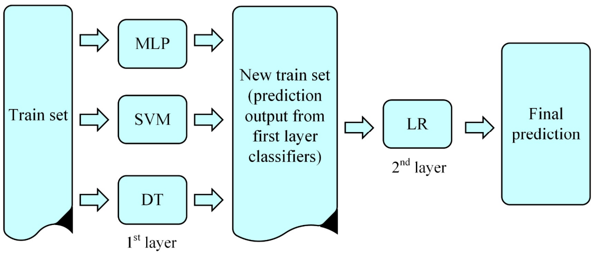
Train set
MLP
SVM
DT
1st layer
New train set (prediction output from first layer classifiers)
LR
2nd layer
Final prediction

Fig. 2. Impact of iForest and SMOTETomek on the stacked ensemble classifiers accuracy.

We implemented the stacked ensemble classifiers in python V3.6 with the Mlxtend V0.15 library [20]. We performed the experiments on a MacBookAir5,2 machine with 8 GB memory and Core i5 processor (1.8 GHz). We utilized default settings provided by the python libraries to offer fewer settings to increase the reproducibility of the present study. We utilized four performance metrics which is presented in such as precision, recall, F-measure, and accuracy to present the performance of our model as compared to other well-known classification models. For each model, we measured TP: true positive, FP: false positive, TN: true negative, and FN: false negative. TP and TN are subject who correctly classified as positive (diabetes) and negative (healthy), respectively. While FP and FN are subject who classified as positive (diabetes) when they are negative (healthy) and classified as negative (healthy) when they are positive (diabetes), respectively. Furthermore, we applied tenfold (10-fold) cross-validation for all classifiers, with the final performance metrics being the average.

TABLE I. PERFORMANCE EVALUATION METRICS.

Metric	Formula
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F-measure	$\frac{2 \times precision \times recall}{precision + recall}$
Accuracy	$\frac{TP+TN}{TP+FN+FP+TN}$

III. Results and Discussions

Table 2 – 4 showed the performance of prediction models to predict type 2 diabetes for each dataset I, II, and III, respectively. In this study, MLP, SVM, DT, and LR are compared with the proposed model which based on the integration of iForest, SMOTETomek and stacked ensemble classifiers. In this scenario, we build two layer of ensemble classifiers and stacked the MLP, SVM, and DT for the first layer and LR for the second layer. The results demonstrated that the proposed model outpaced other models for all dataset and performance metrics used. Table 2 revealed that the proposed model achieved precision, recall, f-measure, accuracy for dataset I by up to 91.76%, 93.57%, 93.76%, and 93.18%, respectively. While Table 3 showed the performance of proposed model on dataset II, by achieving up to 98.44%, 99.31%, 98.87%, and 98.87% for precision, recall, f-measure, and accuracy, respectively. Furthermore, Table 4 showed the result for dataset III and discovered that the proposed model achieved the highest performance in term of precision, recall, f-measure, accuracy by up to 95.07%, 97.28%, 96.1%, and 96.09%, respectively.

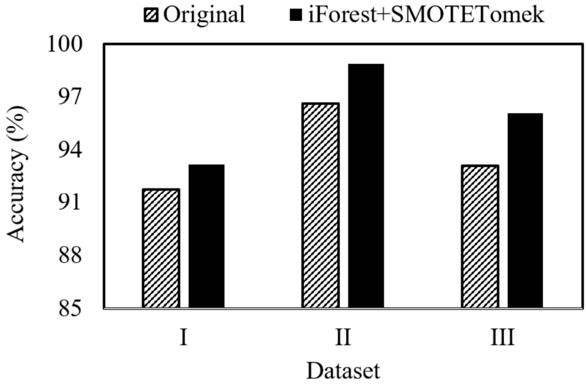

Fig. 3. Impact of iForest and SMOTETomek on the stacked ensemble classifiers accuracy.

TABLE II. PERFORMANCE OF CLASSIFICATION MODEL FOR DATASET I.

Classification model	Performance evaluation metric (%)			
	precision	*recall*	*f-measure*	*accuracy*
MLP	83.73	86.70	84.30	83.80
SVM	66.32	58.80	62.08	73.97
DT	90.46	92.88	92.11	91.75
LR	67.51	76.09	71.43	82.33
Proposed Model	91.76	93.57	93.76	93.18

TABLE III. PERFORMANCE OF CLASSIFICATION MODEL FOR DATASET II.

Classification model	Performance evaluation metric (%)			
	precision	*recall*	*f-measure*	*accuracy*
MLP	86.27	90.23	88.01	87.81
SVM	74.6	93.59	83.02	80.86
DT	97.34	98.26	96.59	96.64
LR	75.6	85.13	80.08	78.82
Proposed Model	98.44	99.31	98.87	98.87

TABLE IV. PERFORMANCE OF CLASSIFICATION MODEL FOR DATASET III.

Classification model	Performance evaluation metric (%)			
	precision	*recall*	*f-measure*	*accuracy*
MLP	93.58	96.01	95.21	95.03
SVM	68.29	89.64	77.49	73.97
DT	90.8	95.92	93.25	93.09
LR	77.52	79.47	78.15	82.33
Proposed Model	95.07	97.28	96.1	96.09

In addition, the impact of iForest and SMOTETomek on the stacked ensemble classifiers accuracy are presented in Fig. 3. We utilized iForest to detect and remove the outlier while the SMOTETomek was used to balance the training data. The result showed that by applying iForest and SMOTETomek methods for the stacked ensemble classifiers, the accuracies were improved for all datasets. The result confirmed that by removing the outlier data using iForest and balancing the training data using SMOTETomek, it could improve the stacked ensemble classifiers accuracy. Finally, employing the iForest and SMOTETomek methods significantly improved the stacked ensemble classifiers for all datasets, with average improvement as much as 2.22%.

IV. Conclusion

This study proposed prediction model to predict type 2 diabetes. The proposed model was developed by utilizing iForest for outlier detection and removal, SMOTETomek for data balancing, and stacked ensemble classifiers as final model. We utilized three diverse publicly available datasets to evaluate the performance of the proposed prediction model. These datasets have never been used by previous studies for predicting type 2 diabetes. Thus, comparison with previous study cannot be presented. The experimental results revealed that the proposed model outperformed other models considered in this study such as MLP, SVM, DT and LR, achieving accuracy up to 93.18%, 98.87%, and 96.09% for dataset I, II, and III, respectively. We also discovered that by employing iForest and SMOTETomek, the performance of stacked ensemble classifiers improved for all datasets with average improvement as much as 2.22%. It is expected that the early diabetes prediction could help the individuals on taking precautions once type 2 diabetes is detected. Furthermore, the developed model in the study could be used as a technological innovation of expert systems, where a more cooperative human-machine coexistence represents a realistic vision by allowing human medical doctors to focus on either finding preemptive ways and/or determining medical treatments of Type 2 Diabetes, while letting computers (predictive models) give earlier signals of the developing possibilities of such a disease in certain patients [21].

Future study should consider the performance comparison of different techniques for detecting the outlier as well as dealing with unbalanced dataset. The managerial implication and clinical trial could also be conducted in the future.

Acknowledgment

This paper is a tribute made of deep respect for a wonderful person, colleague, advisor, and supervisor, Yong-Han Lee (1965–2017).

References

[1] K. G. M. M. Alberti and P. Z. Zimmet, "Definition diagnosis and classification of diabetes mellitus and its complications. Part 1: Diagnosis and classification of diabetes mellitus. Provisional report of a WHO Consultation", *Diabetic Med.,* vol. 15, no. 7, pp. 539-553, Jul. 1998.

[2] N. N. Tun, G. Arunagirinathan, S. K. Munshi and J. M. Pappachan, "Diabetes mellitus and stroke: A clinical update", *World J. Diabetes*, vol. 8, no. 6, pp. 235-248, Jun. 2017.

[3] S. H. Ley, O. Hamdy, V. Mohan and F. B. Hu, "Prevention and management of type 2 diabetes: Dietary components and nutritional strategies", *Lancet*, vol. 383, no. 9933, pp. 1999-2007, Jun. 2014.

[4] S. Wild, G. Roglic, A. Green, R. Sicree and H. King, "Global prevalence of diabetes: Estimates for the year 2000 and projections for 2030", *Diabetes Care*, vol. 27, no. 5, pp. 1047-1053, 2004.

[5] S. Wild, G. Roglic, A. Green, R. Sicree and H. King, "Global prevalence of diabetes: Estimates for the year 2000 and projections for 2030", *Diabetes Care*, vol. 27, no. 5, pp. 1047-1053, 2004.

[6] G. Alfian, M. Syafrudin, M. Ijaz, M. Syaekhoni, N. Fitriyani, and J. Rhee, "A Personalized Healthcare Monitoring System for Diabetic Patients by Utilizing BLE-Based Sensors and Real-Time Data Processing," *Sensors*, vol. 18, no. 7, p. 2183, Jul. 2018.

[7] N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension," *IEEE Access*, vol. 7, pp. 144777-144789, 2019.

[8] D. H. Wolpert, "Stacked generalization", *Neural Netw.*, vol. 5, no. 2, pp. 241-259, Jan. 1992.

[9] C. C. Aggarwal, Ed., *Data Classification*, 0 ed. Chapman and Hall/CRC, 2014.

[10] J. P. Anderson, J. R. Parikh, D. K. Shenfeld, V. Ivanov, C. Marks, B. W. Church, et al., "Reverse Engineering and evaluation of prediction models for progression to type 2 diabetes: An application of machine learning using electronic health records", *J. Diabetes Sci. Technol.*, vol. 10, no. 1, pp. 6-18, Jan. 2016.

[11] R. Domingues, M. Filippone, P. Michiardi and J. Zouaoui, "A comparative evaluation of outlier detection algorithms: Experiments and analyses", *Pattern Recognit.*, vol. 74, pp. 406-421, Feb. 2018.

[12] G. Goel, L. Maguire, Y. Li and S. McLoone, "Evaluation of sampling methods for learning from imbalanced data" in Intelligent Computing Theories, Berlin, Germany:Springer, vol. 7995, pp. 392-401, 2013.

[13] Y. Chen *et al.*, "Association of body mass index and age with incident diabetes in Chinese adults: a population-based cohort study," *BMJ Open*, vol. 8, no. 9, p. e021768, Sep. 2018.

[14] T. Okamura, Y. Hashimoto, M. Hamaguchi, A. Obora, T. Kojima, and M. Fukui, "Ectopic fat obesity presents the greatest risk for incident type 2 diabetes: a population-based longitudinal study," *Int J Obes*, vol. 43, no. 1, pp. 139–148, Jan. 2019.

[15] A. Mozaffary, S. Asgari, M. Tohidi, S. Kazempour-Ardebili, F. Azizi, and F. Hadaegh, "Change in fasting plasma glucose and incident type 2 diabetes mellitus: results from a prospective cohort study," *BMJ Open*, vol. 6, no. 5, p. e010889, May 2016.

[16] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *2008 Eighth IEEE International Conference on Data Mining*, Pisa, Italy, Dec. 2008, pp. 413–422.

[17] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[18] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, Jun. 2004.

[19] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 559–563, Jan. 2017.

[20] S. Raschka, "MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack", *J. Open Source Softw.*, vol. 3, no. 24, pp. 638, Apr. 2018.

[21] O. H. Hamid, N. L. Smith, and A. Barzanji, "Automation, per se, is not job elimination: How artificial intelligence forwards cooperative human-machine coexistence," in 2017 IEEE 15th International Conference on Industrial Informatics (INDIN), Emden, Jul. 2017, pp. 899–904.