Received August 9, 2019, accepted September 27, 2019, date of publication October 2, 2019, date of current version October 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2945129

Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension

NORMA LATIF FITRIYANI[1], MUHAMMAD SYAFRUDIN[1], GANJAR ALFIAN[2], AND JONGTAE RHEE[1]

[1]Department of Industrial and Systems Engineering, Dongguk University, Seoul 04620, South Korea
[2]Nano Information Technology Academy, Dongguk University, Seoul 04626, South Korea

Corresponding authors: Muhammad Syafrudin (udin@dongguk.edu) and Jongtae Rhee (jtrhee@dongguk.edu)

This work was supported by the Ministry of Trade, Industry and Energy (MOTIE) and the Korea Institute for Advancement of Technology (KIAT) through the International Cooperative Research and Development Program under Grant N0002301.

ABSTRACT Early diseases prediction plays an important role for improving healthcare quality and can help individuals avoid dangerous health situations before it is too late. This paper proposes a disease prediction model (DPM) to provide an early prediction for type 2 diabetes and hypertension based on individual's risk factors data. The proposed DPM consists of isolation forest (iForest) based outlier detection method to remove outlier data, synthetic minority oversampling technique tomek link (SMOTETomek) to balance data distribution, and ensemble approach to predict the diseases. Four datasets were utilized to build the model and extract the most significant risks factors. The results showed that the proposed DPM achieved highest accuracy when compared to other models and previous studies. We also developed a mobile application to provide the practical application of the proposed DPM. The developed mobile application gathers risk factor data and send it to a remote server, so that an individual's current condition can be diagnosed with the proposed DPM. The prediction result is then sent back to the mobile application; thus, immediate and appropriate action can be taken to reduce and prevent individual's risks once unexpected health situations occur (i.e., type 2 diabetes and/or hypertension) at early stages.

INDEX TERMS Diabetes, disease prediction, ensemble learning, hypertension, imbalanced data, outlier data.

I. INTRODUCTION

The World Health Organization reported that noncommunicable diseases contribute approximately 41 million premature deaths annually, i.e., nearly 71% of all deaths globally [1], [2]. If unmitigated, the total number of noncommunicable disease related deaths is estimated to reach 52 million yearly by 2030 [3]. The most common noncommunicable diseases are diabetes and hypertension [4], [5], which account for nearly 46.2% and 4% of total deaths [6], respectively. Type 2 diabetes is a continuing metabolic disorder that changes blood glucose levels and is usually a consequence of the body's ineffectiveness to employ its produced insulin [7], [8]. Individuals that have diabetes are likely to have higher risk for stroke and mortality [9]. However, continually monitoring blood glucose levels can effectively prevent and/or mitigate diabetes complication [10], [11]. In developing countries, the number of people suffer from diabetes are estimated to increase from approximately 84 to 228 million by 2030 [12], and significantly burdening health care systems [13]. Furthermore, hypertension is a harmful condition where blood flows through blood vessels (veins and arteries) at persistently elevated pressure. One in three adults globally has elevated blood pressure, and it is the root cause for mortality [14]. Approximately 639 million adults in developing countries suffer from hypertension and are estimated to reach nearly 1 billion adults by 2025 [15].

As the growing awareness of the risk of diabetes and hypertension, several recent studies have utilized machine learning models as decision making tool for early detection of disease, providing high accuracy on detecting diabetes [16], [17] and hypertension [18]–[20] based on individual's current

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Cheng.

condition, and hence helping them to take precautions earlier. An ensemble approach is one of machine learning methods that combines the results from multiple classification models and has shown higher accuracy as compared to a single one [21]–[23]. Several previous studies have successfully used the ensemble approach to assist medical decisions [24], [25] and predicting diabetes [26], [27] as well as hypertension [28]–[30]. In machine learning area, challenging problems may arise such as outlier data and imbalanced datasets which cause on reducing the model accuracy. Previous studies have demonstrated that by using isolation forest (iForest) method for eliminating the outlier data [31], [32] and utilizing synthetic minority oversampling technique tomek link (SMOTETomek) to balance the imbalanced data [33]–[35], the machine learning model's performance was significantly improved.

Nevertheless, none of the previous studies has integrated iForest and SMOTETomek to improve the ensemble model accuracy, especially for type 2 diabetes and hypertension dataset. Therefore, we proposed a disease prediction model (DPM) by utilizing iForest, SMOTETomek and ensemble approach to predict type 2 diabetes and hypertension based on an individual's risk factors data. We used four datasets (related with type 2 diabetes and hypertension) to evaluate the performance of the proposed model. In addition, we developed a mobile application and showed that the proposed DPM can practically be applied to IoT-based mobile healthcare system, offering individual's an effective and convenient way to check their current health status through their smartphones. By providing early diagnosis from the DPM, it is expected that immediate and appropriate action can be taken earlier to reduce and prevent the individual risk once an unexpected health situation occurs (i.e., type 2 diabetes and/or hypertension).

The remainder of this paper is organized as follows. Section II summarizes the literature review. Section III presents the proposed DPM including datasets description, overall design of the proposed DPM and performance evaluation metrics. Section IV discusses performance evaluation results for the proposed model, including the impacts analysis of several outlier elimination and data balancing methods. Section V presents the practical applications of the proposed model. Finally, Section VI summarizes and concludes the paper, and discusses future research directions.

II. LITERATURE REVIEW

A. DIABETES AND HYPERTENSION PREDICTION BASED ON ENSEMBLE APPROACH

Diabetes and hypertension are two main risk factors for stroke which cause mortality [9], [14]. A study has revealed that the worse conditions could be prevented by taking healthy diet and doing daily regular exercise [36]. Therefore, an early disease prediction model that could notifies the individuals on the danger of diabetes and hypertension is required, which could give them the chance to take preemptive actions. Machine learning can be utilized as early disease prediction model to foresee diabetes and hypertension diseases based on the current individual's risk factors data. Several studies have utilized machine learning model and revealed significant results for predicting diabetes [16] or prediabetes [17] as well as disclosing the hypertension risk factors [18]–[20].

An ensemble learning approach is a one of well-known and widely used machine learning model [21]–[23]. The main idea is to combine several machine learning models to help reduce bias and variance, and hence improve the prediction results [37]. Previous studies have used ensemble approach and shown significant outcomes for improving medical decision making and diagnosis, predicting diabetes and identifying their risk factors. Bashir *et al.* [24] developed a disease prediction with multi-layer classifier. The model showed significant result on public disease datasets and was applied to help medical decision making. Ozcift and Gulten [25] built a model based on ensemble of 30 classifier and were applied to Parkinson's, diabetes and heart diseases datasets. The result of the study showed that the developed model achieved higher accuracy as compared to other models and could be used to improve medical diagnosis. Nai-arun and Moungmai [26] utilized one of ensemble learning method (i.e., random forest (RF)) to predict diabetes risk from Sawanpracharak Regional Hospital, Thailand. Several attributes were considered such as body mass index (BMI), age, weight, height, blood pressure (systolic and diastolic), family diabetes and hypertension history, gender, and alcohol and smoking behavior. The study revealed that RF model was more robust than the other models considered in the study. Anderson *et al.* [27] proposed a prediction model based on ensemble of classifiers for predicting the progression to diabetes from electronic health record (EHR) data. The results showed that the model can effectively predict the progression to diabetes and revealed several diabetes risk factors including blood glucose, blood pressure, and hypertension.

Furthermore, several studies related with ensemble learning have been conducted for predicting hypertension and shown promising results. Sakr *et al.* [28] utilized random tree forest as ensemble learning for predicting the individual's risk of hypertension. The model was compared with Bayesian network, locally weighted naïve Bayes, LogitBoost, neural network and support vector machine. The results showed that random tree forest model achieved highest area under the curve (AUC) (up to 0.93) compared with other models when applied to dataset from Henry Ford Health Systems. Sun *et al.* [29] applied RF model (ensemble learning) to predict hypertension transition points. The EHR data were gathered from 1294 patient's with hypertension from Vanderbilt University Medical Center. The result showed that RF model can accurately predict the hypertension transition point and it was expected to be employed for managing personalized hypertension plans. Finally, Singh *et al.* [30] proposed an ensemble model to extract rule sets from diabetic patients for predicting the presence of hypertension. The results showed that the extracted rules set was beneficial for medical experts as well as amateur users to diagnose hypertension and other issues associated with diabetes.

Therefore, in this study, we utilized an ensemble learning approach to predict diabetes and hypertension, with the expectation to reduce and/or prevent individual's risk at early stage. In addition, previous studies have shown that outlier and imbalanced data are common classification problems, which significantly can reduce the model accuracy. Thus, this study incorporates suitable outlier detection and data balancing methods to improve model accuracy.

B. IFOREST-BASED OUTLIER DETECTION AND SMOTETOMEK-BASED DATA BALANCING

Most challenging problems in machine learning areas are dealing with outlier data and imbalanced dataset. Outlier detection is a method to identify the patterns in the dataset that do not significantly follow the projected pattern from the anticipated behavior. Such objects are called outliers or anomalies. Previous studies revealed that by eliminating the outlier data, model accuracy was improved [38]–[42]. One of the method of outlier detection is isolation forest (iForest) [43], [44]. The iForest distinguishes outliers by creating isolation trees (iTrees) and treating outliers as instances/points that have short average length in the iTrees. Several studies have shown significant results of utilizing iForest for outlier and anomaly detection. Domingues *et al.* [31] evaluated different outlier detection methods using the University of California (UCI) repository datasets. The results showed that iForest can effectively be used to identify outliers while providing excellent scalability on large datasets and tolerable memory usage. Calheiros *et al.* [32] used iForest for unsupervised anomaly detection to find issues in huge scale cloud datacenters. The results showed that iForest was feasible and effective to detect the anomaly.

Furthermore, the imbalanced dataset occurs where the majority class membership greatly exceeds the minority class membership and can significantly impact machine learning model performance (i.e., biased and unreliable results). Past studies have revealed that an improved classification models were achieved after solving the imbalanced data problem [45]–[48]. One of the techniques to deal with the imbalanced data is synthetic minority oversampling technique tomek link (SMOTETomek) [33]. Regarding the application of SMOTETomek, several studies have been conducted and shown promising results in balancing the data as well as improving the model performance. Batista *et al.* [33] revealed that SMOTETomek provided better AUC value than synthetic minority oversampling technique edited nearest neighbor (SMOTEENN) when applied to several imbalanced datasets. Goel *et al.* [34] evaluated five sampling methods to solve imbalanced data, using eight datasets from the UCI Repository website. The results revealed that SMOTETomek can improve the model accuracy for most datasets used. In addition, Chen *et al.* [35] developed a framework for lane changing behavior by utilizing resampling methods to solve imbalanced data problem and RF model to predict the lane changing risk level. The result showed that SMOTETomek significantly improved the model accuracy by up to 80.3%.

Several previous studies have shown an improved model accuracy after applying iForest for outlier detection and SMOTETomek to address the imbalanced data. In addition, none of previous studies has used the iForest and SMOTETomek especially for diabetes and hypertension dataset. Therefore, in this study, we propose a disease prediction model (DPM) that utilizes iForest to detect and eliminate outlier data, SMOTETomek to balance the dataset, and ensemble learning approach to predict the diabetes and hypertension. The proposed DPM is expected to improve the performance of the model as well as facilitating the individuals to identify early stage of diabetes and hypertension. Hence, the individuals could take immediate and appropriate preemptive actions to prevent risk of the diseases further.

III. DISEASE PREDICTION MODEL

This section explains the detail of proposed disease prediction model (DPM). Figure 1 shows the proposed DPM for predicting the type 2 diabetes and hypertension. The proposed DPM consists of several modules such as iForest-based outlier detection, SMOTETomek-based data balancing, and ensemble approach to classify type 2 diabetes and hypertension. The detailed dataset, modules descriptions and its implementation, and performance metrics are presented in the following subsections. Finally, the performance evaluation of the proposed DPM with other existing models were measured and are presented in the results and discussion sections.

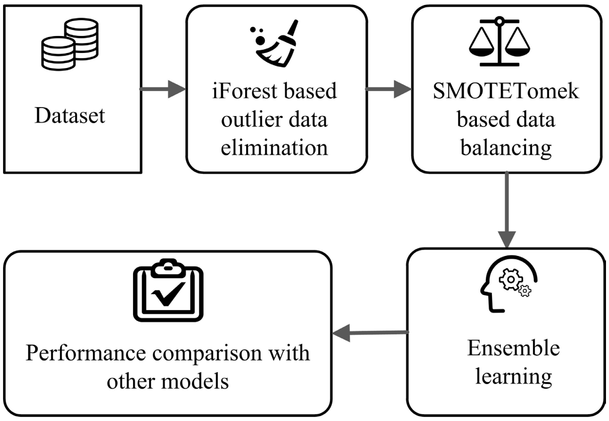

Dataset
iForest based outlier data elimination
SMOTETomek based data balancing
Ensemble learning
Performance comparison with other models

FIGURE 1. Proposed Disease Prediction Model (DPM) for type 2 diabetes and hypertension.

A. DATASETS

We used four different sources of datasets to investigate how type 2 diabetes and hypertension diseases can be identified earlier by utilizing machine learning models. We considered the dataset from type 2 diabetes, hypertension, prehypertension, Chronic Kidney Disease (CKD) as dataset I, II, III, and IV, respectively. The proposed DPM is then applied to those four datasets and expected to generate a general and

TABLE 1. The detailed dataset I attributes definition and distribution.

Attribute				Positive class	Negative class
Symbol	Description	Unit	Type	Mean (± STD)	Mean (± STD)
chol	Total cholesterol	mg/dl	Numeric	227.95 (± 52.49)	203.39 (± 41.07)
stab.glu	Stabilized glucose	mg/dl	Numeric	175.03 (± 80.96)	91.55 (± 26.83)
hdl	High density lipoprotein	mg/dl	Numeric	47.15 (± 16.92)	51.18 (± 17.23)
ratio	chol / hdl	-	Numeric	5.57 (± 2.62)	4.36 (± 1.45)
location	Location	-	Nominal	-	-
age	Age	year	Numeric	56.75 (± 13.29)	44.66 (± 16.08)
gender	Gender	-	Nominal	-	-
height	Height	inch	Numeric	66.33 (± 3.79)	65.95 (± 3.94)
weight	Weight	pound	Numeric	191.29 (± 39.77)	174.6 (± 39.78)
frame	A factor	level	Nominal	-	-
bp.1s	First systolic blood pressure	mmHg	Numeric	144.55 (± 20.2)	135.19 (± 22.89)
bp.1d	First diastolic blood pressure	mmHg	Numeric	84.77 (± 12.98)	83 (± 13.68)
bp.2s	First diastolic blood pressure	mmHg	Numeric	153.06 (± 16.06)	152.17 (± 23.12)
bp.2d	Second diastolic blood pressure	mmHg	Numeric	89.53 (± 10.02)	93.48 (± 11.79)
waist	Waist circumference	inch	Numeric	40.37 (± 5.63)	37.35 (± 5.6)
hip	Hip circumference	inch	Numeric	44.6 (± 5.55)	42.69 (± 5.61)
time.ppn	Postprandial time when labs were drawn	min	Numeric	390.82 (± 334.16)	331.18 (± 302.11)

TABLE 2. The detailed dataset II attributes definition and distribution.

Attribute				Positive class	Negative class
Symbol	Description	Unit	Type	Mean (± STD)	Mean (± STD)
age	Age	year	Numeric	24.7 (± 7.25)	23.48 (± 6.45)
is.obese	Obese	yes/no	Nominal	-	-
bmi	Body mass index	-	Numeric	26.04 (± 5.07)	24.55 (± 3.96)
wc	Waist circumference	inch	Numeric	88.89 (± 13.18)	85.02 (± 10.53)
hc	Hip circumference	inch	Numeric	105.53 (± 9.34)	101.98 (± 8.55)
whr	wc/hc	-	Numeric	83.91 (± 6.81)	83.32 (± 6.7)

TABLE 3. The detailed dataset III attributes definition and distribution.

Attribute				Positive class	Negative class
Symbol	Description	Unit	Type	Mean (± STD)	Mean (± STD)
age	Age	year	Numeric	23.17 (± 5.89)	22.18 (± 4.94)
is.obese	Obese	yes/no	Nominal	-	-
bmi	Body mass index	-	Numeric	24.22 (± 4.56)	22.55 (± 3.5)
wc	Waist circumference	inch	Numeric	78.89 (± 10.45)	75.4 (± 8.29)
hc	Hip circumference	inch	Numeric	102.05 (± 9.21)	99.96 (± 7.9)
whr	wc/hc	-	Numeric	77.23 (± 6.36)	75.4 (± 5.27)

robust classification model. Dr John Schorling [49], [50] provided the dataset I to investigate type 2 diabetes, obesity, and other risk factors in central Virginia. The original dataset comprised 403 subjects, 19 attributes, and no output class. For the present study, the id attribute was removed, and the output class was generated whether the subject was diagnosed with type 2 diabetes (positive) or not (negative) using the available glycosylated hemoglobin (*glyhb*) value. If *glyhb* value < 7.0 then the subject was diagnosed as negative (normal), otherwise as positive (type 2 diabetes). After applying this rule, the updated dataset consists of 73 and 330 for positive (type 2 diabetes) and negative (normal) subjects, respectively. Our proposed DPM uses the dataset I to predict the subject's type 2 diabetes status (positive or negative) from the available risk factor data. Table 1 shows the detailed attributes definition and distribution (mean and standard deviation(STD)) for dataset I.

Golino *et al.* [19] provided the dataset II and III to study the relationship between increased blood pressure and various indicators for male [51] and female [52] subjects, respectively. The original dataset II consists of 175 male subjects, 9 attributes, and one output class (47 and 175 subjects are labelled with positive (hypertension) and negative (normal), respectively). The original dataset III consists of 224 female subjects, 9 attributes, and one output class (95 and 129 subjects are labelled with positive (prehypertension) and negative (normal), respectively). For the present study, *id*, *SBP*, and *DBP* attributes were removed from both of dataset II and III due to their insignificance and similarity with the output class. The hypertension and prehypertension class were defined when systolic blood pressure > 140 mmHg and > 120 mmHg for dataset II and III, respectively. Our proposed DPM uses dataset II and III to foresee the subject's hypertension status (positive or negative) from the available risk factor data. Table 2 and 3 shows the detailed attributes definition and distribution (mean and standard deviation(STD)) for dataset II and III, respectively.

Finally, Dr. P. Soundarapandian, M.D., D.M [53] provided the dataset IV to foresee the chronic kidney disease (CKD) with original dataset consisted of 400 subjects, 24 attributes,

TABLE 4. The detailed dataset IV attributes definition and distribution.

Attribute				Positive class	Negative class
Symbol	Description	Unit	Type	Mean (± STD)	Mean (± STD)
age	Age	year	Numeric	60.12 (± 12.16)	47.05 (± 17.44)
htn	Hypertension	yes/no	Nominal	-	-
bp	Blood pressure	mmHg	Numeric	80.8 (± 15.89)	74.44 (± 11.42)

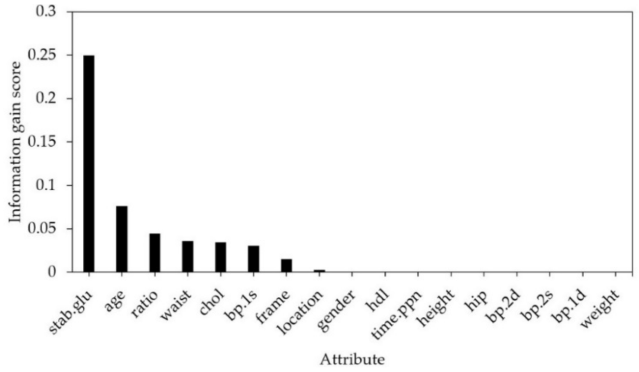


FIGURE 2. Attribute significance scores from the Information Gain (IG) method for dataset I.

and one output class (positive CKD or normal). However, since we focus on the relationship between hypertension and type 2 diabetes; we followed [40] to remove most of the attributes and modify the output class. Finally, the updated dataset consists of three attributes (*age*, *bp*, and *htn*) and one output with 137 and 261 subjects are diagnosed as positive (type 2 diabetes) and negative (normal), respectively. Our proposed DPM uses the dataset IV to diagnose the subject's diabetes status (positive or negative) from the available risk factor data such as age and hypertension; thus, the relationship between hypertension and type 2 diabetes can be revealed. Table 4 shows the detailed attributes definition and distribution (mean and standard deviation(STD)) for dataset IV.

We expect that by using the aforementioned datasets (I, II, III, and IV), our proposed DPM can predict the diabetes and hypertension and reveal their risk factors. The proposed DPM uses dataset I to predict the subject is type 2 diabetes positive or negative; dataset II and III to predict the subject is hypertension (or prehypertension) positive or negative; and dataset IV to investigate the relationship between hypertension and type 2 diabetes.

In addition, we apply data pre-processing to remove missing values and select the most important attribute contributing to the improved model accuracy [54], [55]. Figure 2, 3, 4, and 5 show the attribute significance scores for the dataset I, II, III, and IV, respectively, using the Information Gain (IG) method [56] in Weka V3.8 [57]. Final attributes were selected based on the highest attribute scores for datasets I (*stab.glu*, *age*, *ratio*, *waist*, *chol*, *bp.1s*, *frame*, *location*, *gender*) and IV (*htn*, *age*). However, dataset II and III only had attribute scores for *is.obese*. Therefore, we followed [19] and [40] to utilize all attributes for dataset II and III.

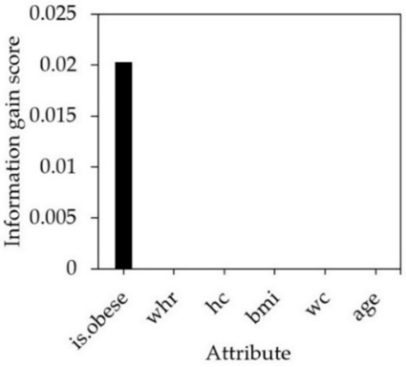


FIGURE 3. Attribute significance scores from the Information Gain (IG) method for dataset II.

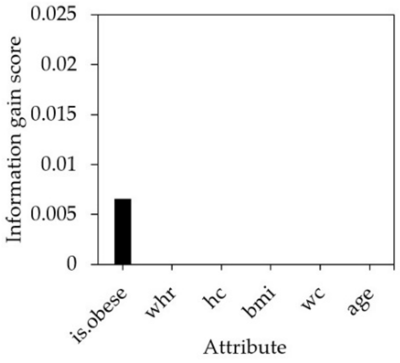


FIGURE 4. Attribute significance scores from the Information Gain (IG) method for dataset III.

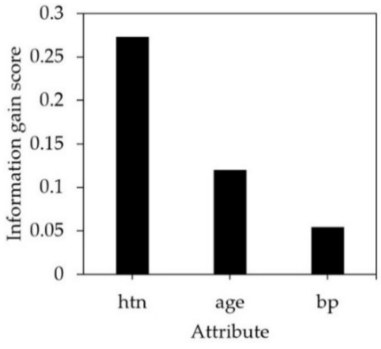


FIGURE 5. Attribute significance scores from the Information Gain (IG) method for dataset IV.

B. OUTLIER DETECTION BASED ON IFOREST

In this study, we used iForest [43], [44] to identify and eliminate the outlier data in diabetes and hypertension dataset. The iForest works by creating an ensemble of isolation trees (iTrees) for each dataset where outliers were defined as instances with short average length in the iTrees. The iTrees is then recursively created by dividing the dataset until all instances is isolated or specific tree height is achieved, where

$$Tree\ height = ceiling(log_2(subsample\ size)) \quad (1)$$

Algorithms 1 and 2 summarize the pseudocode of iForest and iTree, respectively. In iForest, subsample size (*MaxSample*) and number of trees to build (*NumTree*) are two significant parameters to be measured. The iForest works well when MaxSample is kept small, larger MaxSample reduces iForest ability to isolate outlier data since normal data can interfere with isolation. We have tried different setup parameters and found optimal MaxSample is 10% of overall data size and NumTree is 100. The iForest was implemented using Scikit-learn python library V0.20.2 [58]. Table 5 shows the detailed parameters setting as well as the results after applying iForest for each datasets. Finally, we eliminated the outlier data from each datasets and keep the remaining data for further analysis.

Algorithm 1 Isolation Forest Pseudocode

```
Input    Risk data, $D$; number of trees, $NumTree$;
         subsampling size, $MaxSample$
Output   Set of iTrees
Initialize $Forest$
set height limit $l$ = ceiling($\log_2(MaxSample)$)
for $i = 1$ to $NumTree$ do
    $D' \leftarrow$ sample($D, MaxSample$)
    $Forest \leftarrow Forest \cup iTree(D', 0, l)$
end for
return $Forest$
```

Algorithm 2 Isolation Tree Pseudocode

```
Input    Risk data, $D$; current tree height, $C$;
         height limit, $l$
Output   One iTree
if $c \geq l$ or $|D| \leq 1$ then
    return $exNode\ \{Size \leftarrow |D|\}$
else
    let $Q$ be a list of features $D$
    randomly select a feature $q \in Q$
    randomly select a split point $p$ from $min$
    and $max$ values of feature $q$ in $D$
    $D_l \leftarrow$ filter($D, q < p$)
    $D_r \leftarrow$ filter($D, q \geq p$)
    return
    $inNode \left\{ \begin{array}{c} \text{Left} \leftarrow iTree\,(D_l, C+1, l)\,, \\ \text{Right} \leftarrow iTree\,(D_r, C+1, l)\,, \\ SplitAtt \leftarrow q, \\ SplitValue \leftarrow p \end{array} \right\}$
end if
```

C. SMOTETOMEK FOR IMBALANCED DATASET

Our proposed DPM utilizes SMOTETomek [33] to solve the imbalanced dataset. In this study, the majority and minority class distributions are imbalanced for all datasets. In dataset I and IV, the minority class are the subjects who diagnosed as positive (type 2 diabetes), while in dataset II and III, the minority class are the subjects who diagnosed as positive (hypertension and prehypertension, respectively). We applied SMOTETomek method to balance the datasets. SMOTE oversamples the minority class to randomly generate instances and increasing minority class instances, and Tomek undersamples a class to remove noise while maintaining the balanced distributions. Previous studies have shown combining SMOTE and Tomek (SMOTETomek) provides better results compared either alone [33].

Table 6 shows the initial minority class distributions before and after outlier removal for each dataset I, II, III, and IV. We used the Imbalanced-learn python library V0.4.3 [59] to implement SMOTETomek method and applied it to each datasets. As can be seen in Table 6, after applying SMOTETomek, the datasets become balanced. The classification goal is always to minimize errors during learning; hence we expect that an improved model accuracy can be achieved from the balanced datasets.

D. ENSEMBLE LEARNING

Ensemble learning [21]–[23] is type of classification method designed via combination of multiple classification algorithms into a single model to reduce bias and variance and hence improve the prediction accuracy. In this study, we employed ensemble learning with cross validation, which provides two classifier levels: first and second, which can help avoid overfitting [37]. Cross-validation helps to prevent the model using the same training set for first and second level classifiers. Previous studies showed that the multilayer perceptron (MLP), support vector machines (SVM), decision tree (DT) and logistic regression (LR) can be utilized as prediction models and showed significant result on improving the classification accuracy in several health datasets [16]–[19], [40]. Therefore, we used MLP, SVM, and DT as first level classifiers; with LR as the second level classifier. Figure 6 shows the proposed model structure, and Algorithm 3 summarizes the ensemble learning with cross-validation. The proposed ensemble approach works as follows.

1. Training set D is partitioned into K disjoint subsets and the first level classifier is run K times.
2. Each learnt classifier is applied to the remaining subset and the output used as input feature space for the second level classifier.
3. The second level classifier is built from the first level output.
4. The first level classifier is re-trained on the whole dataset so that all dataset instances are used.
5. The second level classifier is then applied to the updated first level output to obtain final predictions.

We implemented the ensemble learning using Scikit-learn V0.20.2 [58] and Mlxtend V0.15.0 [60] python libraries. In this study, the outlier data from diabetes and hypertension

TABLE 5. The detailed iForest parameters setting and result for each dataset I, II, III, and IV.

Dataset	*MaxSample*	*NumTree*	Number of Outliers	Number of subjects before outlier removal	Number of subjects after outlier removal
I	41	100	94	403	309
II	18	100	36	175	139
III	23	100	38	224	186
IV	40	100	156	398	242

TABLE 6. The result of SMOTETomek-based data balancing.

Dataset	Before SMOTETomek		After SMOTETomek	
	Minority (%)	Majority (%)	Minority (%)	Majority (%)
I	15 (4.85%)	294 (95.15%)	293 (50%)	293 (50%)
II	31 (22.30%)	108 (77.70%)	99 (50%)	99 (50%)
III	73 (39.25%)	113 (60.75%)	94 (50%)	94 (50%)
IV	51 (21.07%)	191 (78.93%)	191 (50%)	191 (50%)

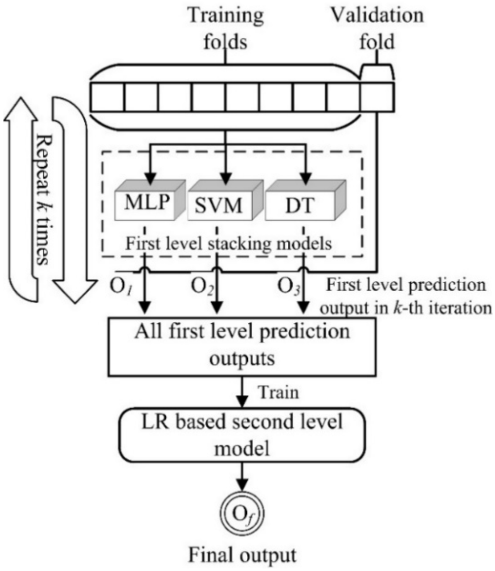

Training folds
Validation fold
Repeat *k* times
MLP
SVM
DT
First level stacking models
O1
O2
O3
First level prediction output in *k*-th iteration
All first level prediction outputs
Train
LR based second level model
Of
Final output

FIGURE 6. Ensemble learning approach structure.

are eliminated by using iForest-based outlier detection method, and SMOTETomek is used to balance the datasets. Finally, the ensemble learning is used to learn from the training dataset, and the prediction results is then compared with the testing dataset to obtain model accuracy. We used performance metrics to compare the performance of proposed DPM with other models and previous studies. In addition, we validated the proposed DPM by applying into mobile application to diagnose the subjects based on theirs's current condition.

E. PERFORMANCE METRICS

Model predictions can have four different potential outcomes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP and TN outcomes are correctly classified, whereas FP outcomes classified as positive when they are actually negative, and FN outcomes classified as negative when they are actually positive. Datasets I and IV outcomes are diabetes, while dataset II and III outcome is hypertension and prehypertension, respectively. We employed 10-fold cross validation generating the models for all classification models, with final performance metric being the average. Several performance metrics are used in this study such as precision (p), which is defined as

$$p = \frac{TP}{TP + FP}, \tag{2}$$

recall (r), which is defined as

$$r = \frac{TP}{TP + FN}, \tag{3}$$

F-measure (f), which is defined as

$$f = \frac{2pr}{p + r}, \tag{4}$$

accuracy (acc), which is defined as

$$acc = \frac{TP + FN}{TP + FN + FP + TN}. \tag{5}$$

In addition, we also used the value of area under the receiver operating characteristic curve (AUC) to compare the performance of proposed DPM with other models. For the given k training data, the AUC can be defined as [61], [62]

$$AUC\left(x^{+}, x^{-}\right) = \frac{1}{k^{+}k^{-}} \sum\nolimits_{i=1}^{k^{+}} \sum\nolimits_{j=1}^{k^{-}} 1_{h\left(x_i^{+}\right)>h\left(x_j^{-}\right)}, \tag{6}$$

where the term $1_{h\left(x_i^{+}\right)>h\left(x_j^{-}\right)}$ corresponds to a '1' when the elements $h\left(x_i^{+}\right) > h\left(x_j^{-}\right)$, $\forall i = 1, 2, \ldots, k^{+}, \forall j = 1, 2, \ldots, k^{-}$, and '0' otherwise. It should be noted that when the model achieves a highest accuracy, the value of AUC is close or equal to 1.

IV. RESULTS AND DISCUSSIONS

A. DPM PERFORMANCE EVALUATION

Table 7 shows the proposed disease prediction model (DPM) compared with other classification models and previous studies results for dataset I. The first previous study [16] used k-means to remove incorrect cluster data and classified the

Algorithm 3 Ensemble Learning Approach With *K*-Fold Cross-Validation Pseudocode

```
Input    Input data, $D = \{x_i, y_i\}_{i=1}^{m}$; cross-validation, $K$; learning models, $T$
Output   Ensemble model $M$
1. Prepare training data for the second level model using cross-validation
Randomly split $D$ into $K$ equal sized subsets, $D = \{D_1, D_2, \ldots, D_K\}$
for $k \leftarrow 1$ to $K$ do
    1.1. Learn the first level model
    for $t \leftarrow 1$ to $T$ do
        Learn model $m_{kt}$ from subsets $\{D_1, D_2, \ldots, D_K\}$
    end for
    1.2. Construct a training set for the second level model
    for $x_i \epsilon D_k$ do
        Get list $\{x_i{}', y_i\}$, where $x_i{}' = \{m_{k1}(x_i), m_{k2}(x_i), \ldots, m_{kT}(x_i)\}$
    end for
end for
2. Learn second level models
Learn new model $m'$ from list $\{x_i{}', y_i\}$
3. Re-learn the first level models
for $t \leftarrow 1$ to $T$ do
    Learn model $m_t$ from subsets $\{D_1, D_2, \ldots, D_K\}$
end for
return $M(x) = m'(m_1(x), m_2(x), \ldots, m_T(x))$
```

TABLE 7. Classification model performance for dataset I.

Classification model	Performance metric				
	p (%)	*r* (%)	*f* (%)	*acc* (%)	AUC
MLP	52.59	48.57	45.08	84.9	0.85
SVM	88.235	41.096	56.075	81.9	0.62
DT	36.37	42.68	32.18	69.69	0.59
LR	52.5	47.14	43.77	84.9	0.91
K-means + LR [16]	91.6	96.4	-	90.7	0.957
DBSCAN + SMOTE + RF [40]	91.497	93.403	92.440	92.555	-
Proposed DPM	94.49	98.62	96.32	96.74	0.99

TABLE 8. Classification model performance for dataset II.

Classification model	Performance metric				
	p (%)	*r* (%)	*f* (%)	*acc* (%)	AUC
MLP	73.23	97.56	83.56	72.02	0.5
SVM	73.04	99.17	84.1	72.58	0.45
DT	71.35	69.42	70.13	56.96	0.46
LR	73.67	95.06	82.77	71.33	0.58
CART [19]	-	58.38	-	-	0.68
DBSCAN + SMOTE + RF [40]	78.788	70.270	74.286	76.419	-
Proposed DPM	93.57	84.89	88.8	85.73	0.87

remaining data using logistic regression (LR). The latter [40] used density-based spatial clustering of applications with noise (DBSCAN) to eliminate outlier data, synthetic minority over-sampling technique (SMOTE) to balance the remaining data, and RF to classify the updated data. In our proposed model, different methods to remove outlier data as well as solve the imbalanced data problem were used. The result showed that our proposed DPM outperformed all other classification models applied to dataset I, including MLP, SVM, DT, and LR, and all previous study results, for all metrics; achieving $acc = 96.74\%$ compared to previous studies $acc = 90.7\%$ [16] and 92.555% [40], respectively. In addition, the proposed DPM also achieved highest AUC (by up to 0.99) compared with others model and all previous studies results.

Table 8 shows the proposed DPM compared with other classification model and previous studies results for dataset II. The first previous study [19] used classification and regression tree (CART) to predict hypertension on the male subjects. The latter [40] used DBSCAN to remove outlier data, SMOTE to balance the remaining dataset, and RF to classify the updated dataset. Our result revealed that the proposed DPM outperformed all other considered models (MLP, SVM, DT and LR) applied to dataset II, and both previous studies, for all metrics, achieving acc = 85.73% and AUC = 0.87.

Table 9 shows the proposed DPM compared with other classification model and previous study result for dataset III. The previous study [19] used classification and regression tree (CART) to predict prehypertension on the female subjects. Our result revealed that the proposed DPM outperformed all other considered models (MLP, SVM, DT and LR) applied to dataset III, and previous study, for all metrics, achieving acc = 75.78% and AUC = 0.76.

Table 10 shows the proposed DPM compared with other classification model and previous study [40] result for dataset IV. The proposed model outperformed all other considered models (MLP, SVM, DT and LR) applied to

TABLE 9. Classification model performance for dataset III.

Classification model	Performance metric				
	p (%)	*r* (%)	*f* (%)	*acc* (%)	AUC
MLP	57.31	69.81	57	54.92	0.53
SVM	56.38	85.26	67.77	53.48	0.44
DT	57.33	54.1	54.69	49.96	0.49
LR	61.53	82.76	69.52	59.41	0.62
CART [19]	-	45.65	-	-	0.566
Proposed DPM	75.6	81.78	77.12	75.78	0.76

TABLE 10. Classification model performance for dataset IV.

Classification model	Performance metric				
	p (%)	*r* (%)	*f* (%)	*acc* (%)	AUC
MLP	67.78	70.49	67.02	80.84	0.89
SVM	64.49	57.69	56.66	75.85	0.77
DT	61.48	56.43	56.13	72.48	0.75
LR	67.98	77.14	69.67	81.57	0.89
DBSCAN + SMOTE + RF [40]	83.665	84.677	84.168	83.644	-
Proposed DPM	100	100	100	100	1

dataset IV, and previous study, for all metrics, achieving acc = 100% and AUC = 1. In addition, the proposed model revealed that age and hypertension are important risk factors for type 2 diabetes.

Finally, our proposed DPM provided significant performance across all metrics for all four datasets (type 2 diabetes and hypertension datasets) when compared with a wide range of models and previous studies results.

B. IMPACTS OF OUTLIER REMOVAL AND DATA BALANCING ON MODEL ACCURACY

Table 5 shows original and post iForest based outlier removal dataset sizes. Outlier removal based on iForest significantly improved ensemble model prediction for all datasets, from acc = 78.17%, 73.2%, 51.7%, and 75.76% to acc = 95.19%, 77.72%, 60.22%, and 100%, for dataset I, II, III, and IV, respectively, with average improvement as much as 13.58%.

We also compared with other outlier detection method, called local outlier factor (LOF) [56], [63] on ensemble model accuracy. LOF detects outliers based on local density deviation for a specific data point with reference to neighboring points. Figure 7 shows iForest and LOF outlier detection impacts on ensemble model accuracy on three datasets. Although LOF also enhances ensemble model accuracy, averaging 2.89% improvement for all datasets, this is still less than the improvement from applying proposed iForest method (13. 58%).

After outlier data are removed, we applied SMOTETomek to balance the datasets (see Table 6). Balancing the dataset provided significant accuracy improvement for the proposed method when compared with omitting the balance step, achieving accuracy (acc) from 95.19%, 77.72%, 60.22%, and 100% to 96.74%, 85.73%, 75.78%, and 100% for dataset I, II, III, and IV, respectively, representing average improvement up to 6.28%.

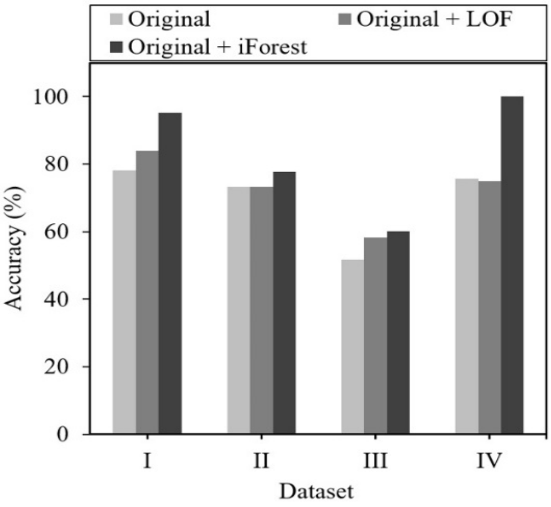


FIGURE 7. Impacts on model accuracy after outlier removal between original, original + LOF, and original + iForest).

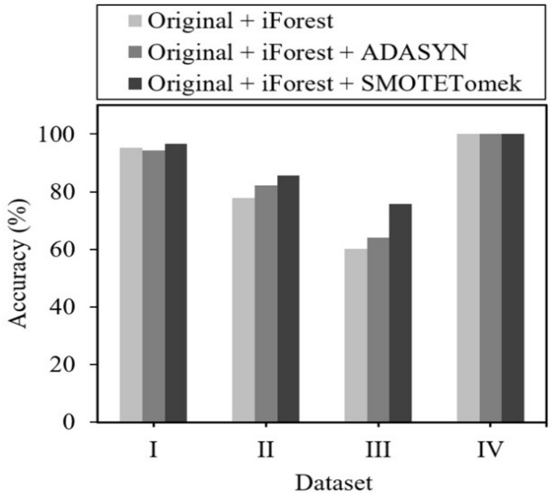


FIGURE 8. Impacts on model accuracy after outlier removal with iForest and data balancing between original + iForest, original + iForest + ADASYN, and original + iForest + SMOTETomek.

In addition, we compared the performance of SMOTE-Tomek with other data balancing method, called adaptive synthetic (ADASYN) [64]. ADASYN adaptively generate synthetic data for minority class according to their weighted distribution, reducing bias caused by an imbalanced dataset. Figure 8 shows prediction accuracy improvement after applying iForest + ADASYN and iForest + SMOTETomek from original datasets. ADASYN also improved accuracy for the ensemble model, achieving acc = 94.37%, 82.3%, 64.11%, and 100% for dataset I, II, III, and IV, respectively (with average improvement as much as 1.91%). However, this result is still less than the improvement from applying iForest + SMOTETomek (6.28%).

Our results showed that the integration of iForest and SMOTETomek generated higher classification performance as compared to other models in our datasets. However, the proposed model might generate different result on different type of datasets.

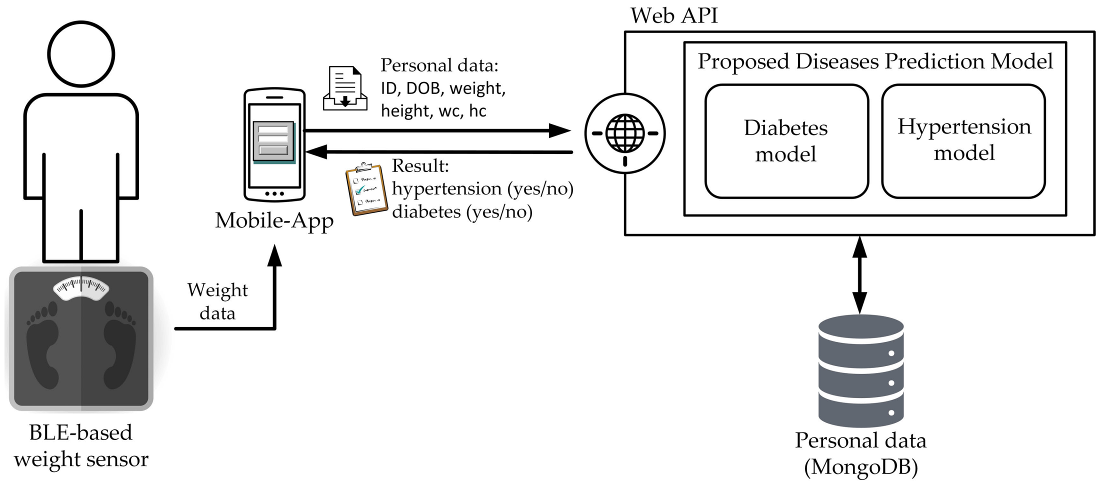
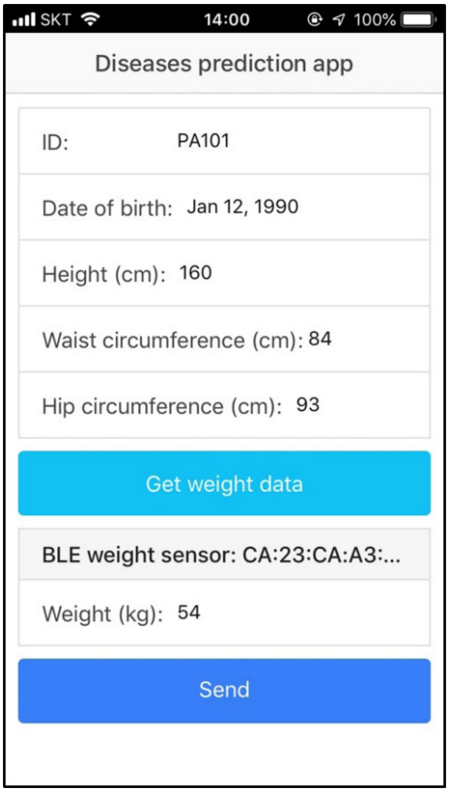


FIGURE 9. The proposed disease prediction model (DPM) system architecture framework.



Diseases prediction app

ID: PA101

Date of birth: Jan 12, 1990

Height (cm): 160

Waist circumference (cm): 84

Hip circumference (cm): 93

Get weight data

BLE weight sensor: CA:23:CA:A3:...

Weight (kg): 54

Send

FIGURE 10. Mobile application interface for gathering user's data.

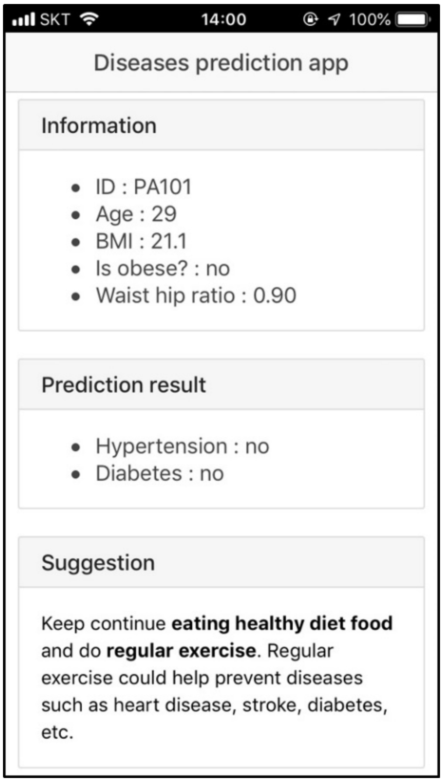
V. PRACTICAL APPLICATIONS

In this section, we implemented the proposed DPM into mobile application (mobile-app) with the integration of IoT-based weight scale sensor (smart scale) to show the feasibility of our proposed model in practical applications. Figure 9 shows the proposed system architecture to diagnose hypertension and type 2 diabetes based on data input by the user through a mobile-app. The mobile-app gathers weight data from a smart scale through Bluetooth low energy (BLE) communication using generic attributes [65]. BLE is a common communication protocol for gathering sensor data wirelessly with low power consumption [66], [67]. Other personal data such as user id (ID), date of birth (DOB), height, waist circumference (wc), and hip circumference (hc) can then be combined with the weight data to be transmitted into a secure server through a web application programming interface (Web API) and stored in a database.

We used MongoDB V4.0.2 database since it has been efficiently used to store data in timely manner for many applications, including healthcare [68], ecommerce [69], manufacturing [70], [71], and supply chain [72]. Hypertension and diabetes model which were created from dataset II and III, and IV, respectively, are then used to predict the subject's disease based on the supplied risk factor data, and the prediction result is then sent to the user's mobile-app.

In this study, the developed Web API consisted of Python V3.6.5, Hug V2.4.1, and Pymongo V3.7.1, running on a Python Web Server Gateway Interface (WSGI) HTTP Server: Gunicorn V19.9.0. The prototype mobile-app was developed using Ionic V1.3.3 and AngularJS V1.5.3. Figure 10 shows the interface of the prototype mobile-app gathering risk factors data input by the user and weight data from smart scale. The weight data is wirelessly retrieved from the smart scale through BLE after pressing the "Get weight data" button. Once all the apps input fields are filled, the user can press "send" button to send all risk factor data to the remote server. The remote server then will execute the trained DPMs to diagnose the user's hypertension and/or type 2 diabetes status.

Figure 11 shows prediction result interface for the user from their mobile-app after sending their risk factor data to the remote server. The result includes suggestions such as to keep eating diet food and/or regular exercise to help prevent and/or reduce the risk of getting the diseases [73]–[75]. The developed mobile-app and disease prediction model could help users effectively reduce their hypertension and type 2 diabetes risk while both diseases are in their early stages.



Diseases prediction app

Information

- ID : PA101
- Age : 29
- BMI : 21.1
- Is obese? : no
- Waist hip ratio : 0.90

Prediction result

- Hypertension : no
- Diabetes : no

Suggestion

Keep continue **eating healthy diet food** and do **regular exercise**. Regular exercise could help prevent diseases such as heart disease, stroke, diabetes, etc.

FIGURE 11. The prediction result interface from the proposed DPM.

VI. CONCLUSION

We developed a disease prediction model (DPM) for type 2 diabetes and hypertension by integrating iForest, SMOTE-Tomek, and ensemble learning. The iForest was used to detect and eliminate the outlier data from dataset while SMOTE-Tomek was utilized to balance the imbalanced dataset and ensemble learning was applied to predict the diseases. Four public datasets with relevant to type 2 diabetes and hypertension were used to evaluate the performance of DPM. Dataset I provided type 2 diabetes status and a wide range of risk factors while dataset II and III provided hypertension risk factors and current hypertension status. Finally, dataset IV provided linked data between age, hypertension, and diabetes. Performance results showed that the proposed DPM significantly enhanced the model accuracy (and all other metrics) when compared with other models and previous studies results, achieving $acc = 96.74\%$, 85.73%, 75.78%, and 100% for datasets I, II, III, and IV, respectively. In addition, the result also revealed a strong connection between age and hypertension for type 2 diabetes.

Furthermore, we also implemented and integrated the proposed DPM into mobile application (mobile-app) with IoT-based sensors to diagnosing the subject/user at their convenience. The mobile-app gathered sensor data through sensors, combined with other user input risk factor data, and transmitted to the remote server. All risk factors data were then stored into MongoDB, which can efficiently handle rapidly increasing incoming data. The proposed DPM was then triggered to diagnose the user's current hypertension and/or type 2 diabetes status, which was later sent to the user's mobile-app. Thus, the mobile-app provides users their current state of health in real-time, allowing them taking an immediate and appropriate action to reduce and/or prevent the risks further.

Future studies will investigate optimal model parameters for the prediction model and applying other clinical datasets with broader features/attributes to create more generalize and robust model. Furthermore, other outlier detection, data sampling options and the effect on the different age's population could be further investigated.

ACKNOWLEDGMENT

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

REFERENCES

[1] World Health Organization. (2018). *Noncommunicable Diseases*. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/noncommunicable-diseases

[2] World Health Organization. (2016). *NCD Mortality and Morbidity*. [Online]. Available: https://www.who.int/gho/ncd/mortality_morbidity/en/

[3] World Health Organization. (2016). *Projections of Mortality and Causes of Death, 2016 to 2060*. [Online]. Available: https://www.who.int/healthinfo/global_burden_disease/projections/en/

[4] M. S. Capehorn, D. W. Haslam, and R. Welbourn, "Obesity treatment in the UK health system," *Current Obesity Rep.*, vol. 5, no. 3, pp. 320–326, Sep. 2016.

[5] S. Hariharan, R. Umadevi, T. Stephen, and S. Pradeep, "Burden of diabetes and hypertension among people attending health camps in an urban area of Kancheepuram district," *Int. J. Community Med. Public Health*, vol. 5, no. 1, p. 140, Dec. 2017.

[6] World Health Organization. (2014). *Global Status Report on Noncommunicable Diseases 2014*. [Online]. Available: https://www.who.int/nmh/publications/ncd-status-report-2014/en

[7] 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.

[8] American Diabetes Association, "Standards of medical care in diabetes–2006," *Diabetes Care*, vol. 29, pp. S4–S42, Jan. 2006.

[9] 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.

[10] C. Hayes and A. Kriska, "Role of physical activity in diabetes management and prevention," *J. Amer. Dietetic Assoc.*, vol. 108, no. 4, pp. S19–S23, Apr. 2008.

[11] 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.

[12] 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.

[13] F. Rubino, "Is type 2 diabetes an operable intestinal disease?: A provocative yet reasonable hypothesis," *Diabetes Care*, vol. 31, pp. S290–S296, Feb. 2008.

[14] World Health Organization. (2012). *World Health Statistics 2012*. [Online]. Available: https://www.who.int/gho/publications/world_health_statistics/2012/en/

[15] P. M. Kearney, M. Whelton, K. Reynolds, P. Muntner, P. K. Whelton, and J. He, "Global burden of hypertension: Analysis of worldwide data," *Lancet*, vol. 365, no. 9455, pp. 217–223, 2005.

[16] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Inform. Med. Unlocked*, vol. 10, pp. 100–107, Aug. 2018.

[17] X.-H. Meng, Y.-X. Huang, D.-P. Rao, Q. Zhang, and Q. Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," *Kaohsiung J. Med. Sci.*, vol. 29, no. 2, pp. 93–99, Feb. 2013.

[18] M. Tayefi, H. Esmaeili, M. S. Karimian, A. A. Zadeh, M. Ebrahimi, M. Safarian, M. Nematy, S. M. R. Parizadeh, G. A. Ferns, and M. Ghayour-Mobarhan, "The application of a decision tree to establish the parameters associated with hypertension," *Comput. Methods Programs Biomed.*, vol. 139, pp. 83–91, Feb. 2017.

[19] H. F. Golino, L. S. de Brito Amaral, S. F. P. Duarte, C. M. A. Gomes, T. de Jesus Soares, L. A. dos Reis, and J. Santos, "Predicting increased blood pressure using machine learning," *J. Obesity*, vol. 2014, Jan. 2014, Art. no. 637635.

[20] B. M. Heo and K. H. Ryu, "Prediction of prehypertenison and hypertension based on anthropometry, blood parameters, and spirometry," *Int. J. Environ. Res. Public Health*, vol. 15, no. 11, p. 2571, Nov. 2018.

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

[22] L. Breiman, "Stacked regressions," *Mach. Learn.*, vol. 24, no. 1, pp. 49–64, Jul. 1996.

[23] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. D. Spyropoulos, and P. Stamatopoulos, "Stacking classifiers for anti-spam filtering of e-mail," Jun. 2001, *arXiv:cs/0106040*. [Online]. Available: https://arxiv.org/abs/cs/0106040

[24] S. Bashir, U. Qamar, and F. H. Khan, "IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework," *J. Biomed. Inform.*, vol. 59, pp. 185–200, Feb. 2016.

[25] A. Ozcift and A. Gulten, "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms," *Comput. Methods Programs Biomed.*, vol. 104, no. 3, pp. 443–451, Dec. 2011.

[26] N. Nai-arun and R. Moungmai, "Comparison of classifiers for the risk of diabetes prediction," *Procedia Comput. Sci.*, vol. 69, pp. 132–142, Nov. 2015.

[27] J. P. Anderson, J. R. Parikh, D. K. Shenfeld, V. Ivanov, C. Marks, B. W. Church, J. M. Laramie, J. Mardekian, B. A. Piper, R. J. Willke, and D. A. Rublee, "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.

[28] S. Sakr, R. Elshawi, A. Ahmed, W. T. Qureshi, C. Brawner, S. Keteyian, M. J. Blaha, and M. H. Al-Mallah, "Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford ExercIse Testing (FIT) project," *PLoS ONE*, vol. 13, no. 4, Apr. 2018, Art. no. e0195344.

[29] J. Sun, C. D. McNaughton, P. Zhang, A. Perer, A. Gkoulalas-Divanis, J. C. Denny, J. Kirby, T. Lasko, A. Saip, and B. A. Malin, "Predicting changes in hypertension control using electronic health records from a chronic disease management program," *J. Amer. Med. Inform. Assoc.*, vol. 21, no. 2, pp. 337–344, Mar. 2014.

[30] N. Singh, P. Singh, and D. Bhagat, "A rule extraction approach from support vector machines for diagnosing hypertension among diabetics," *Expert Syst. Appl.*, vol. 130, pp. 188–205, Sep. 2019.

[31] 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.

[32] R. N. Calheiros, K. Ramamohanarao, R. Buyya, C. Leckie, and S. Versteeg, "On the effectiveness of isolation-based anomaly detection in cloud data centers," *Concurrency Comput., Pract. Exper.*, vol. 29, no. 18, p. e4169, Sep. 2017.

[33] 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," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, p. 20–29, Jun. 2004.

[34] G. Goel, L. Maguire, Y. Li, and S. McLoone, "Evaluation of sampling methods for learning from imbalanced data," in *Intelligent Computing Theories*, vol. 7995, D.-S. Huang, V. Bevilacqua, J. C. Figueroa, and P. Premaratne, Eds. Berlin, Germany: Springer, 2013, pp. 392–401.

[35] T. Chen, X. Shi, and Y. D. Wong, "Key feature selection and risk prediction for lane-changing behaviors based on vehicles' trajectory data," *Accident Anal. Prevention*, vol. 129, pp. 156–169, Aug. 2019.

[36] A. Alloubani, A. Saleh, and I. Abdelhafiz, "Hypertension and diabetes mellitus as a predictive risk factors for stroke," *Diabetes Metabolic Syndrome, Clin. Res. Rev.*, vol. 12, no. 4, pp. 577–584, Jul. 2018.

[37] C. C. Aggarwal, Ed., *Data Classification: Algorithms and Applications*. Boca Raton, FL, USA: CRC Press, 2014.

[38] K. Shin, A. Abraham, and S. Y. Han, "Improving kNN text categorization by removing outliers from training set," in *Computational Linguistics and Intelligent Text Processing*, vol. 3878, A. Gelbukh, Ed. Berlin, Germany: Springer, 2006, pp. 563–566.

[39] A. J. Tallón-Ballesteros and J. C. Riquelme, "Deleting or keeping outliers for classifier training?" in *Proc. 6th World Congr. Nature Biol. Inspired Comput. (NaBIC)*, Porto, Portugal, 2014, pp. 281–286.

[40] M. F. Ijaz, G. Alfian, M. Syafrudin, and J. Rhee, "Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest," *Appl. Sci.*, vol. 8, no. 8, p. 1325, Aug. 2018.

[41] G. Alfian, M. Syafrudin, B. Yoon, and J. Rhee, "False positive RFID detection using classification models," *Appl. Sci.*, vol. 9, no. 6, p. 1154, Mar. 2019.

[42] M. Syafrudin, N. Fitriyani, G. Alfian, and J. Rhee, "An affordable fast early warning system for edge computing in assembly line," *Appl. Sci.*, vol. 9, no. 1, p. 84, Dec. 2018.

[43] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, Pisa, Italy, Dec. 2008, pp. 413–422.

[44] F. T. Liu, K. M. Ting, and Z. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discovery Data*, vol. 6, no. 1, pp. 1–39, Mar. 2012.

[45] M. A. H. Farquad and I. Bose, "Preprocessing unbalanced data using support vector machine," *Decis. Support Syst.*, vol. 53, no. 1, pp. 226–233, Apr. 2012.

[46] A. Yousefian-Jazi, J.-H. Ryu, S. Yoon, and J. J. Liu, "Decision support in machine vision system for monitoring of TFT-LCD glass substrates manufacturing," *J. Process Control*, vol. 24, no. 6, pp. 1015–1023, Jun. 2014.

[47] J. Kim, Y. Han, and J. Lee, "Data imbalance problem solving for smote based oversampling: Study on fault detection prediction model in semiconductor manufacturing process," in *Proc. Inf. Technol. Comput. Sci.*, 2016, pp. 79–84.

[48] R. Harliman and K. Uchida, "Data- and algorithm-hybrid approach for imbalanced data problems in deep neural network," *Int. J. Mach. Learn. Comput.*, vol. 8, no. 3, pp. 208–213, Jun. 2018.

[49] J. P. Willems, J. T. Saunders, D. E. Hunt, and J. B. Schorling, "Prevalence of coronary heart disease risk factors among rural blacks: A community-based study," *Southern Med. J.*, vol. 90, no. 8, pp. 814–820, Aug. 1997.

[50] (1997). *Diabetes Data*. [Online]. Available: http://staff.pubhealth.ku.dk/~tag/Teaching/share/data/Diabetes.html

[51] H. Golino. (2013). *Men's Dataset From the 'Predicting Increased Blood Pressure Using Machine Learning, Figshare*. [Online]. Available: https://doi.org/10.6084/m9.figshare.845665.v1

[52] H. Golino. (2013). *Women's Dataset From the 'Predicting Increased Blood Pressure Using Machine Learning, Figshare*. [Online]. Available: https://doi.org/10.6084/m9.figshare.845664.v1

[53] UCI Machine Learning Repository. (2015). *Chronic_Kidney_Disease Data Set*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease

[54] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artif. Intell.*, vol. 97, nos. 1–2, pp. 245–271, Dec. 1997.

[55] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.

[56] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Diego, CA, USA: Elsevier, 2011.

[57] (2019). *Weka 3: Data Mining Software in Java*. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/

[58] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[59] G. Lemaitre, 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, 2017.

[60] 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, p. 638, Apr. 2018.

[61] C. Marrocco, R. P. W. Duin, and F. Tortorella, "Maximizing the area under the ROC curve by pairwise feature combination," *Pattern Recognit.*, vol. 41, no. 6, pp. 1961–1974, Jun. 2008.

[62] K.-A. Toh, J. Kim, and S. Lee, "Maximizing area under ROC curve for biometric scores fusion," *Pattern Recognit.*, vol. 41, no. 11, pp. 3373–3392, Nov. 2008.

[63] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, 2000.

[64] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Hong Kong, Jun. 2008, pp. 1322–1328.

[65] Bluetooth. (2019). *Specifications GATT Specifications*. [Online]. Available: https://www.bluetooth.com/specifications/gatt/

[66] M. Patel and J. Wang, "Applications, challenges, and prospective in emerging body area networking technologies," *IEEE Wireless Commun.*, vol. 17, no. 1, pp. 80–88, Feb. 2010.

[67] C. Gomez, J. Oller, and J. Paradells, "Overview and evaluation of Bluetooth low energy: An emerging low-power wireless technology," *Sensors*, vol. 12, no. 9, pp. 11734–11753, Aug. 2012.

[68] 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.

[69] G. Alfian, M. F. Ijaz, M. Syafrudin, M. A. Syaekhoni, N. L. Fitriyani, and J. Rhee, "Customer behavior analysis using real-time data processing: A case study of digital signage-based online stores," *Asia–Pacific J. Marketing Logistics*, vol. 31, no. 1, pp. 265–290, Aug. 2019.

[70] M. Syafrudin, N. L. Fitriyani, D. Li, G. Alfian, J. Rhee, and Y.-S. Kang, "An open source-based real-time data processing architecture framework for manufacturing sustainability," *Sustainability*, vol. 9, no. 11, p. 2139, Nov. 2017.

[71] M. Syafrudin, G. Alfian, N. Fitriyani, and J. Rhee, "Performance analysis of IoT-based sensor, big data processing, and machine learning model for real-time monitoring system in automotive manufacturing," *Sensors*, vol. 18, no. 9, p. 2946, Sep. 2018.

[72] G. Alfian, M. Syafrudin, and J. Rhee, "Real-time monitoring system using smartphone-based sensors and nosql database for perishable supply chain," *Sustainability*, vol. 9, no. 11, p. 2073, Nov. 2017.

[73] D. Kranciukaite-Butylkiniene, D. Rastenyte, K. Jureniene, and L. Jancaityte, "[Physical and mental health of stroke survivors and their daily activities]," *Med. (Kaunas)*, vol. 45, no. 11, pp. 896–903, 2009.

[74] M. Kuwabara, R. Kuwabara, K. Niwa, I. Hisatome, G. Smits, C. A. Roncal-Jimenez, P. S. MacLean, J. M. Yracheta, M. Ohno, M. A. Lanaspa, R. J. Johnson, and D. I. Jalal, "Different risk for hypertension, diabetes, dyslipidemia, and hyperuricemia according to level of body mass index in Japanese and American subjects," *Nutrients*, vol. 10, no. 8, p. 1011, Aug. 2018.

[75] M. Włodarczyk and G. Nowicka, "Obesity, DNA damage, and development of obesity-related diseases," *Int. J. Mol. Sci.*, vol. 20, no. 5, p. 1146, Mar. 2019.

NORMA LATIF FITRIYANI received the B.Eng. degree in informatics from Universitas Islam Negeri Sunan Kalijaga, Yogyakarta, Indonesia, in 2014, and the M.S. degree in information management from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2016. She is currently pursuing the Ph.D. degree with the Department of Industrial and Systems Engineering, Dongguk University, Seoul, South Korea. Her research interests include u-health, data mining, information systems, image processing, healthcare, and the Internet of Things.

MUHAMMAD SYAFRUDIN received the B.Eng. degree in informatics from Universitas Islam Negeri Sunan Kalijaga, Yogyakarta, Indonesia, in 2013, and the Dr. Eng. degree in industrial and systems engineering from Dongguk University, Seoul, South Korea, in 2019, where he is currently an Assistant Professor with the Department of Industrial and Systems Engineering. His research interests include machine learning, information systems, edge-computing, the Internet of Things, big data, healthcare, and smart factory.

GANJAR ALFIAN received the B.Eng. degree in informatics from Universitas Islam Negeri Sunan Kalijaga, Yogyakarta, Indonesia, in 2009, and the M.Eng. and Dr. Eng. degrees from the Department of Industrial and Systems Engineering, Dongguk University, Seoul, South Korea, in 2012 and 2016, respectively, where he is currently an Assistant Professor with the Nano Information Technology Academy. His research interests include machine learning, RFID, the Internet of Things, big data, healthcare, simulation, and carsharing service.

JONGTAE RHEE received the B.S. degree from Seoul National University, the M.S. degree from the Korea Advanced Institute of Science and Technology, and the Ph.D. degree from the University of California at Berkeley, all in industrial engineering. He is currently a Professor with the Department of Industrial and Systems Engineering, Dongguk University, Seoul, South Korea, where he is also the Director of the u-SCM Research Center, Nano Information Technology Academy. He has been performing researches and leading various projects related to practical artificial neural network models for production and operation planning, personalized healthcare, and smart factory. His research interests include applied artificial intelligence, data mining, optimization, the Internet of Things, big data, sensors, and healthcare.