Predicting Prediabetes/Diabetes Diagnosis from Patient Health Indicators

Neil John Catapang
National Graduate School of Engineering
University of the Philippines Diliman
Quezon City, Philippines
nfcatapang@up.edu.ph

ABSTRACT

Diabetes is a prevalent disease worldwide which can lead to serious complications if not detected early. However, current diagnostic methods for diabetes can be expensive, invasive, and insensitive to early cases. This study proposes a diabetes diagnostic tool based on machine learning algorithms to predict diabetes diagnoses from patient health indicators. Three models (artificial neural networks, Random Forest, and XGBoost) were developed for prediction of diabetes on the CDC Diabetes Health Indicators dataset balanced by random under-sampling and Synthetic Minority Oversampling Technique (SMOTE). All three models performed comparably on the under-sampled dataset with 75% accuracy, 80% sensitivity, 70% specificity, 76% F1 score, and 0.82-0.83 ROC-AUC, while the Random Forest classifier performed best on the over-sampled dataset with 92% accuracy, 88% sensitivity, 96% specificity, 92% F1 score, and 0.97 ROC-AUC. The top features found to be relevant for prediction of diabetes cases include general health status, body mass index, age, high cholesterol, and high blood pressure. These models can serve as supplementary diabetes diagnosis aid for healthcare workers in screening of potential diabetes cases.

1. INTRODUCTION

1.1 Background

Diabetes is a globally prevalent disease which leads to serious complications in affected people. The complications which generally arise from diabetes include macrovascular and microvascular diseases which are serious difficulties that require major focus. These reasons require healthcare units to implement effective diabetes diagnosis, prevention, and control measures. However, the current diagnostic criteria utilized by these units have limitations which leads to diagnostic delays and missed detection opportunities [28].

Thus, to prevent cases from leading to complications, healthcare units must gain access to innovative strategies which can diagnose diabetes as early as possible. One possible solution is leveraging machine learning (ML) algorithms to improve the timeliness and accuracy of diagnosis. The advantage of using ML algorithms lies in their capability to process large amounts of data (i.e., patient data such as age, body mass index, high blood pressure occurrence, etc.) and learn patterns to predict disease diagnoses in a short time. Thus, using traditional diagnostic criteria augmented with ML algorithms can greatly accelerate detection of diabetes, preventing the progression of cases to severe complications.

With these reasons, this study proposes to develop machine learning algorithms, such as artificial neural networks and classifier ensembles, to predict pre-diabetic or diabetic cases using the available worldwide patient data. Comparison of these ML algorithms based on performance metrics were done to select the best classifier with highest real-world applicability. Hopefully, this study will serve as a diagnosis tool and reference for healthcare institutions in the development and implementation of early diabetes diagnosis strategies.

1.2 Objectives

The primary objective of this paper is to develop classifiers using machine learning algorithms to predict cases of prediabetes/diabetes from patient health indicators. Specifically, the objectives of this paper are:

1. To develop artificial neural networks, random forest classifiers, and XGBoost classifiers for prediction of prediabetes/diabetes cases from patient health indicators,
2. To balance the CDC Diabetes Health Indicators Dataset by random under-sampling and synthetic minority over-sampling (SMOTE),
3. To determine the most and least important features for prediction of prediabetes/diabetes cases, and
4. To evaluate and compare the performance of the classifiers.

1.3 Scope and Limitations

This study focused on construction of three classical machine learning models which are artificial neural networks, random forest classifiers, and XGBoost classifiers for prediction of prediabetes/diabetes from the publicly-available CDC Surveillance System data. The best performing models were selected based on performance metrics such as accuracy, sensitivity, specificity, F1 score, and ROC-AUC. The ultimate goal of this study was to develop a potential diabetes diagnosis tool which can aid health care units for screening of possible diabetes cases.

The results obtained in this study may have been limited by the inherent quality of the dataset and the complexity of the constructed machine learning models. Furthermore, the class imbalance of the dataset was addressed by random under-sampling and synthetic over-sampling, which possibly affects the applicability of the models in real-world environments. Lastly, while the models in this study were trained to generalize on different instances in the dataset, the performance results of this model may not be applicable to populations in differing countries and regions, as the data in the CDC Diabetes Health Indicators dataset were predominantly collected from American citizens.

2. REVIEW OF RELATED LITERATURE

2.1 Overview and Current Diagnostic Methods of Diabetes

Diabetes is a widespread disease characterized by uncontrollable glucose levels in the blood [16]. According to the World Health Organization, 422 million people are affected by diabetes worldwide, in which most cases were found in low and middle-income countries [24]. Studies have shown that elevated blood glucose levels brought by diabetes increase the risk of numerous cardiovascular diseases such as coronary artery disease, stroke, neuropathy, nephropathy, and retinopathy [19, 9]. Due to its prevalence and severity, it is necessary that healthcare institutions worldwide focus on formulation of effective diagnosis and management of diabetes.

Current diagnostic tests for diabetes include plasma glucose level quantification tests (fasting plasma glucose or 2-h plasma glucose) and the hemoglobin A1C test [3]. These methods, although effective in diagnosing diabetes, are also faced with limitations such as high testing costs, inconvenience for most patients, and low sensitivity in early diabetes cases. In fact, the hemoglobin A1C test has been shown to miss detection of large proportions of asymptomatic early-stage diabetes [5], which could lead to progression of detrimental symptoms when left undetected. Also, these limitations may potentially discourage people afflicted with diabetes to consult their doctors for diagnosis.

2.2 Machine Learning Applications in Medical Diagnosis

Current advancements in technology, including artificial intelligence and machine learning, are becoming more widely utilized in healthcare especially in classification and disease diagnostics. A study by Gray et al. [11] utilized Random Forest-based algorithm to achieve 75-89% accuracies in classification of patients with Alzheimer's disease and healthy controls from MRI, FDG-PET, CSF, and genetic data. Another study [18] have developed a model that detects epileptic seizures from discrete wavelet transform (DWT) of EEG signals using Naive-Bayes and k-Nearest Neighbor (k-NN) algorithm. Furthermore, Mohan et al. [14] designed a heart disease prediction model from 13 clinical features using Hybrid Random Forest with Linear Model (HRFLM) and with 88.7% accuracy. Another study [26] employed convolution neural networks (CNN) to process X-ray image data and predict pneumonia cases with 79-95% accuracy. These studies show promising innovations in medical diagnostics, which address the limitations of traditional diagnostic methodologies.

2.3 Artificial Neural Network

Artificial neural networks (ANN) are systems of interconnected nodes (called *neurons*) which can take and process input data to produce a set of outputs. ANNs were developed as a representation of the mathematical workings of the biological nervous systems [1]. ANN models have components such as *neurons, layers, activation functions, weights*, and *biases*. The output of the artificial neural network depends on the connectivity of neurons, value of weights and biases, and the type of activation functions used [25]. These models have wide range of industrial applications which include information, medicine, economy, control, transportation and psychology [25].

ANN models can be trained using the *backpropagation algorithm*, which updates the weights and biases of the network in order to produce the desired outputs and reduce the classification error rate. Through numerous repeated presentations (epochs) of the randomized training data to the network, the weights and biases of the network converge to a set of values which minimizes the errors in the classification.

2.4 Random Forest Classifier

The Random Forest (RF) Classifier is a tree ensemble classifier algorithm which addresses the over-fitting tendency of individual decision trees. This algorithm applies bootstrap aggregation (bagging), which uses a subset of the training data (without replacement) to train each decision tree in the ensemble [6]. Random forests provide results which can be compared to boosting and adaptive bagging algorithms without adaptive sampling of the training set [7]. Furthermore, in comparison to other classifiers, random forest are advantageous in applications including data anomaly detection and feature importance identification [23].

2.5 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) algorithm is a scalable tree boosting algorithm for machine learning developed by Tianqi Chen [8]. XGBoost has advantages over other models due to its capacity to handle large complex datasets, its parallelizable code design, and its ability to handle missing data [20]. XGBoost was empirically found to be faster than other gradient boosting algorithms [2, 8], and this can be attributed to its sparsity-aware algorithm and parallel computing ability. With these characteristics, this model is usually preferred in applications involving sparse and very large datasets in cases where processing speed is prioritized.

3. METHODOLOGY

3.1 Dataset

The dataset (CDC Diabetes Health Indicators [21]) used for this study was obtained from UC Irvine Machine Learning Repository. This dataset is a cleaned version of the 2015 Behavioral Risk Factor Surveillance System (BRFSS) dataset of Centers of Disease Control and Prevention (CDC). The binary version of the dataset was selected, which is composed of 21 features and 253,680 instances. For the 2-class dataset, the negative class (0) indicates "non-diabetic" patients, while positive class (1) indicates "prediabetic" or "diabetic" patients. The features in this dataset include demographics (e.g., age, sex), laboratory results (e.g., high blood pressure, high cholesterol, etc.), and lifestyle factors (e.g., if patient eats fruits, vegetables, heavy drinker, etc.). This dataset is highly imbalanced as only 14% of instances are labeled as prediabetic/diabetic.

3.2 Data Pre-processing

With imbalanced data, the decisions of the classifiers tend to incline with the majority class [12], which likely leads to poor performance on the minority class (i.e. positive class). To address the imbalance of the original dataset, resampling techniques were applied on the dataset before the construction and training of the classifiers. The resampling techniques used in this study were Random Under-sampling and Synthetic Minority Over-sampling Technique (SMOTE). In random under-sampling, instances in the majority class were randomly selected and removed until a balanced distribution of data is achieved [4]. On the other hand, in SMOTE, the algorithm samples the minority class instances and their corresponding neighbors in the feature space to generate synthetic data which have features based on the respective instances and neighbors [10]. The imbalanced-learn library were used in this study to apply these resampling methods.

The features in the dataset were then scaled via standardization method using the scikit-learn library in Python. This sets the mean of each feature to 0, and then scales the features to unit variance [17]. The dataset was split into training and test sets with split ratio of 80%:20%.

3.3 Classifier Construction and Evaluation

This study developed three classifiers: (1) artificial neural networks (ANN), (2) random forest classifier, and (3) XGBoost classifier for prediction of prediabetes/diabetes from patient data. A fraction of the training set were used to train these models, and the remaining fraction (i.e. validation set) were then used to tune the models and verify validity.

The artificial neural network model used for prediction of diabetes cases were built using Keras library. Table 1 shows the ANN model hyperparameters used in this study. The validation set for this model constitutes 30% of the training set.

Hyperparameter	Value
Network Structure	2 Hidden + 1 Output Layers
Hidden Layer 1	16 nodes, LeakyReLU(0.01)
Hidden Layer 2	16 nodes, LeakyReLU(0.01)
Output Layer	1 node, Sigmoid
Optimizer	Adam
Loss Function	Cross-entropy
Regularizer	L2(0.001)
Callbacks	Early Stopping, LR Decay

Table 1: ANN model hyperparameters used for prediction of prediabetes/diabetes cases.

The random forest models and XGBoost models were constructed in Python using the scikit-learn and xgboost libraries, respectively. The hyperparameters tuned for the random forest model were (1) the number of estimators, (2) maximum depth of tree, and (3) minimum samples required for node split. On the other hand, the hyperparameters tuned for the XGBoost model were (1) the number of estimators, (2) learning rate, (3) maximum depth of tree, (4) subsample fraction, and (5) fraction of features used per tree. For both ensemble models, tuning was done via grid search with 3-fold cross-validation. Tables 2 and 3 show the grid of hyperparameters tested for tuning of random forest and XGBoost models. The best hyperparameters were selected based on the corresponding area under the Receiver Operating Characteristics curve (ROC-AUC).

Hyperparameter	Values		
'n_estimators'	50	100	200
'max_depth'	None	10	20
'min_samples_split'	2	5	10

Table 2: Hyperparameter grid used for tuning the random forest model for prediction of prediabetes/diabetes cases.

Hyperparameter	Values		
'n_estimators'	50	100	200
'learning_rate'	0.01	0.1	0.2
'max_depth'	3	5	7
'subsample'	0.8	1.0	-
'colsample_bytree'	0.8	1.0	-

Table 3: Hyperparameter grid used for tuning the XGBoost model for prediction of prediabetes/diabetes cases.

The classifiers used in this paper were evaluated based on the models' accuracy, sensitivity, specificity, F1 score, and area under the receiver operating characteristic curve (ROC) on the test datasets. The sensitivity score or the true positive rate (TPR) indicates how many diseased cases were correctly detected as positive, while the specificity or the true negative rate (TNR) indicates how many healthy cases were correctly identified as negative. Furthermore, the learning curves of the artificial neural networks on the test sets were also plotted to analyze the training behavior of the ANN classifier.

3.4 Feature Importance

The Gini importance, also known as mean reduction in impurity, was used to rank the most important features for prediction of diabetes cases from patient health data. This importance score indicates how frequent the feature x_i is selected for node split, as well as its overall degree of contribution for the classification problem [13]. A tuned random forest model trained on the under-sampled dataset was used for this purpose. The under-sampled dataset were selected to exclude synthetic data which may affect the importance values of the features. The 5 most important and least important features were selected, and their relevance in diabetes risk and prediction were analyzed.

4. EXPERIMENTAL RESULTS

4.1 ANN Learning Curves

The learning curves of the artificial neural network models in this study describes the trends in models loss values on the training and validation sets per epoch. Figures 1 and 2 shows the learning curves of ANN models on the under-sampled dataset and over-sampled dataset, respectively. The curves show that training finished as soon as the validation losses were detected to plateau, which indicates that further training may no longer improve classification performance.

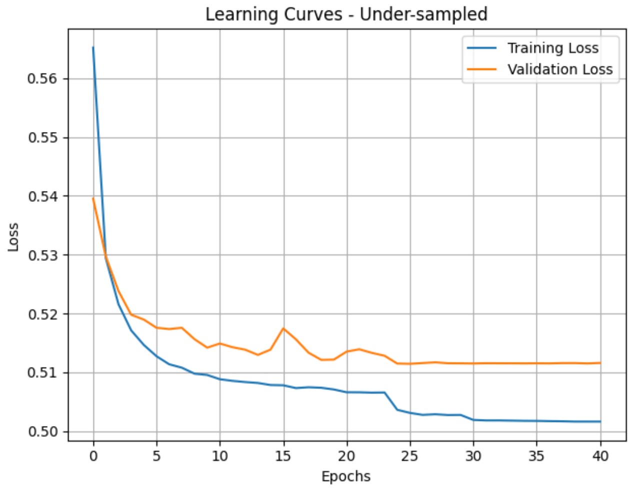

Figure 1: Learning curves of the ANN model on the under-sampled dataset.

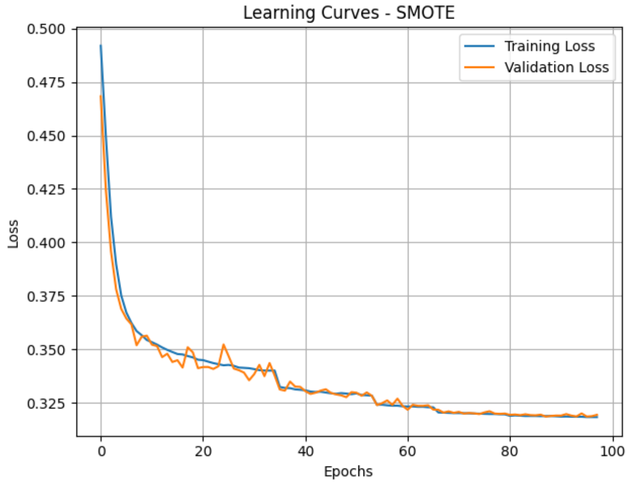

Figure 2: Learning curves of the ANN model on the over-sampled dataset.

4.2 RF and XGBoost Models Tuning

The best set of hyperparameters for the random forest and XGBoost models were determined via grid search and validated via 3-fold cross-validation. Tables 4 and 5 enumerate the best selected hyperparameters for the random forest and XGBoost models. For both models, maximizing the number of estimators resulted in better performance in prediction of prediabetes/diabetes cases.

Hyperparameter	Under-sampled	SMOTE
‘n_estimators’	200	200
‘max_depth’	10	None
‘min_samples_split’	2	5

Table 4: Best set of hyperparameters for the random forest model.

Hyperparameter	Under-sampled	SMOTE
‘n_estimators’	200	200
‘learning_rate’	0.1	0.2
‘max_depth’	3	5
‘subsample’	0.8	1.0
‘colsample_bytree’	1.0	0.8

Table 5: Best set of hyperparameters for the XGBoost model.

4.3 Classifier Performance

After training and validation of the three selected models, their classification performance on the test dataset were evaluated using accuracy (ACC), sensitivity (TPR), specificity (TNR), F1 score, and ROC-AUC. Table 6 compares the classification performance measures of ANN, RF, and XGBoost models on the test sets, while Figure 3 shows the respective ROC curves for each model. All three classifiers trained on the under-sampled dataset performed similarly with 75% accuracy and 76% F1 scores, while the random forest trained on the over-sampled dataset performed best with 92% accuracy and 92% F1 score.

Model	ACC	TPR	TNR	F1	AUC
ANN-Under	75%	80%	70%	76%	0.82
RF-Under	75%	80%	69%	76%	0.82
XGB-Under	75%	80%	70%	76%	0.83
ANN-SMOTE	85%	81%	89%	84%	0.93
RF-SMOTE	92%	88%	96%	92%	0.97
XGB-SMOTE	92%	86%	97%	91%	0.97

Table 6: Comparison of classification performance of ANN, RF, and XGBoost models on the test sets.

4.4 Feature Importances

The features used in this study for diabetes prediction were ranked using the Gini importance. Figure 4 shows the features ranked in descending importance order. From this figure, the feature importance scores appeared to be dominated by only a few features. The top 5 features found to be relevant for prediction of diabetes are ’GenHlth’, ’HighBP’, ’BMI’, ’Age’, and ’HighChol’, while the bottom 5 features are ’AnyHealthCare’, ’NoDocbcCost’, ’Veggies’, ’Fruits’, and ’Smoker’. The top 5 most important features found in this study contributed to 77% of total importances of 21 features.

5. ANALYSIS AND DISCUSSION OF RESULTS

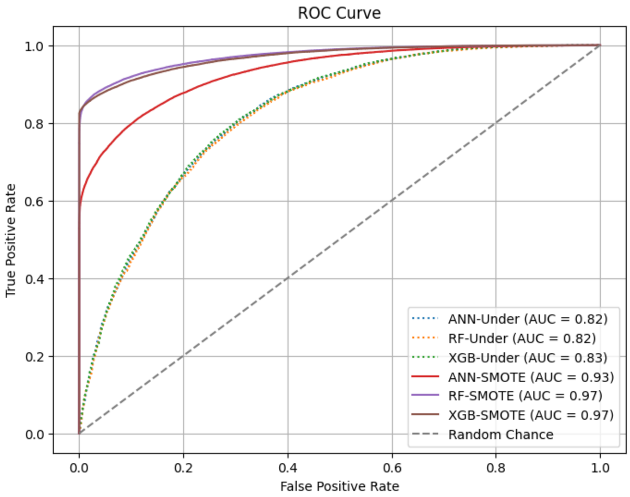

Figure 3: Receiver operating characteristic curves for ANN, RF, and XGBoost models.

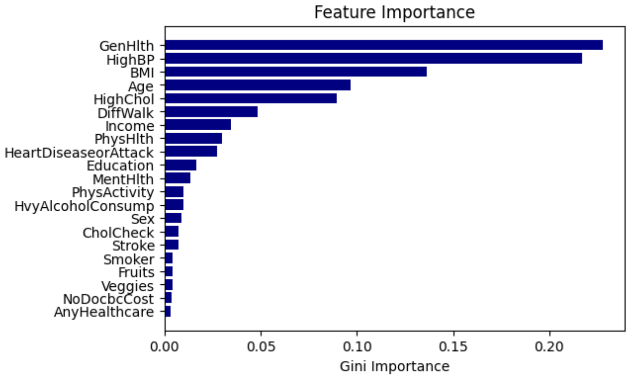

Figure 4: Feature importances (Gini) for diabetes prediction ranked in descending order.

5.1 Analysis of ANN Learning Curves

The trends in learning curves shown in Figures 1 and 2 indicate that the models have learned the patterns in the datasets adequately without over-fitting on the training set. The rapid decrease in training and validation losses during the early stages of training followed by plateau of validation loss in later stages implies that the model have converged to a solution which minimizes classification errors. The problem of over-fitting was prevented in this study using regularization (L2) and callback techniques (early stopping and learning rate decay).

In Figure 1, training and validation losses of the model on under-sampled data exhibit a gap approximately equal to 0.01 on the final stage of training (epoch 40). This suggests that the ANN model performs slightly better on the training set than on the test set. This behavior was expected as the model learned directly from the training set. Nevertheless, the observed loss gap is desirably minimal, which suggests that the ANN model trained on the under-sampled dataset can generalize well on unseen data.

On the other hand, a very minimal gap was observed between the training and validation losses of the ANN model on over-sampled dataset as shown in Figure 2. Furthermore, this model achieved superior performance compared to the model trained on the under-sampled dataset, as evidenced by its lower loss values. The validation loss is only slightly higher than the training loss, but the training loss may be inflated due to L2 regularization, implying that the actual loss gap could be higher.

5.2 Comparison of Model Performance

As shown in Table 6 and Figure 3, models trained on the over-sampled dataset (via SMOTE) outperformed those that were trained on the randomly under-sampled dataset. This performance difference can be attributed to the higher number of instances from the minority class after over-sampling, which preserves most information from the original dataset. The under-sampled dataset, on the other hand, may potentially lose relevant information [4] especially for highly imbalanced original dataset. Nevertheless, all models were adequate for prediction of diabetes cases with $> 80\%$ sensitivity and > 0.80 ROC area under the curve.

The models trained on the under-sampled dataset demonstrated similar performance metrics, indicating that the data limitations, such as low data complexity or the presence of noise, may have hindered the models' ability to capture an optimal decision threshold. However, while the inherent noise in the dataset can adversely affect performance measures, it may also contribute to the robustness of these models. In contrast, models trained on low-noise data may perform poorly on real-world environments, which are generally regarded as noisy environments [27].

In the over-sampled dataset, the random forest model performed best among the constructed models. This result can be attributed to the capability of RF models to prevent over-fitting and enhance generalization especially with higher number of estimators. Both RF and XGBoost models performed better than the ANN model, which suggests that tree ensembles captured more relevant and complex patterns than neural networks in this specific dataset. This can also indicate that the ANN model used for this dataset may be too simplistic, and adding more neurons and layers could potentially enhance the performance of the model.

5.3 Analysis of Feature Importance

Figure 4 shows that the importances of features in predicting diabetes cases were dominated by 'GenHlth', 'HighBP', 'BMI', 'Age', and 'HighChol'. The description for these features are as follows:

1. '**GenHlth**' - Self-reported rating (1-5) of overall health status
2. '**HighBP**' - Score that indicates if a person has high blood pressure (0 or 1)
3. '**BMI**' - Person's body mass index
4. '**Age**' - Person's age
5. '**HighChol**' - Score that indicates if a person has high cholesterol levels (0 or 1)

These features (except 'GenHlth') can be considered as objective factors since these can be verified through measurements, tests, and document checks. This makes these features more reliable for diabetes prediction among others in this dataset. In fact, these features are generally accepted as risk factors for development of type 2 diabetes [15]. On the other hand, the importance of the 'GenHlth' feature can be due to the reported experiences of the people with diabetes who participated in the data collection. People who reported low scores in their general health status possibly experience illnesses and other recurrent symptoms, which likely indicate that they have underlying health conditions such as diabetes.

The least important features found in this study are as follows:

1. '**AnyHealthcare**' - Score that indicates if a person has health care coverage (e.g. health insurance) (0 or 1)
2. '**NoDocbcCost**' - Score that indicates if a person has ever felt the need to see a doctor but could not due to the cost (0 or 1)
3. '**Veggies**' - Score that indicates if a person eats vegetables once or more per day (0 or 1)
4. '**Fruits**' - Score that indicates if a person eats fruits once or more per day (0 or 1)
5. '**Smoker**' - Score that indicates if a person has ever smoked at least 100 cigarettes in total (0 or 1)

Although these features exhibit low importance scores in this study, these may still play a role to the development of diabetes. For instance, one of the least important features in this study ('Smoker') was found to be linked to development of type 2 diabetes [22]. The low importance scores of these factors can be attributed to their subjective nature. This introduces biases due to potential over-reporting or under-reporting of participants with regards to their lifestyle and habits. Also, using binary scoring for some of these features (e.g., 'Veggies', 'Fruits', and 'Smoker') may have limited their importances, and that using the frequency scores (i.e. how many times per day) instead of binary scores could possibly improve correlation of these features with diabetes risk. If these biases and scoring limitations were reduced, these 5 least important features could become more robust predictors of diabetes, which ultimately results in model performance improvements.

6. CONCLUSION AND FUTURE WORK

This study has successfully constructed machine learning models such as artificial neural network models, random forest models, and XGBoost models for prediction of prediabetes/diabetes cases from patient health indicators. These models were trained on the CDC Diabetes Health Indicators dataset, balanced via random under-sampling and SMOTE. All three models predicted prediabetes/diabetes cases on the under-sampled dataset with 75% accuracy, 80% sensitivity, 70% specificity, 76% F1 score, and 0.82-0.83 ROC-AUC. On the over-sampled dataset, the random forest classifier performed best in predicting prediabetes/diabetes cases with 92% accuracy, 88% sensitivity, 96% specificity, 92% F1 score, and 0.97 ROC-AUC. All models constructed were found to be adequate for diabetes prediction.

Feature importance rankings were successfully determined using the Gini importance score from the random forest classifier trained on the under-sampled dataset. The 5 most important features are general health status, body mass index, age, high cholesterol, and high blood pressure. Most of these features are widely accepted risk factors for development of type 2 diabetes.

Future work relevant to this study can augment additional data which includes surveillance data collected from the period of 2016 onwards to ensure the relevance and applicability of the findings to current health trends. Reduction of biases and using frequency scores instead of binary scores for other features may also improve the performance of the constructed models.

7. REFERENCES

[1] Abraham, A. Artificial neural networks. *Handbook of measuring system design* (2005).

[2] Ali, Z., Abduljabbar, Z., Tahir, H., Sallow, A., and Almufti, S. Exploring the power of extreme gradient boosting algorithm in machine learning: a review. *Academic Journal of Nawroz University 12* (05 2023), 320–334.

[3] American Diabetes Association. Classification and diagnosis of diabetes: Standards of medical care in diabetes—2022. *Diabetes Care 45*, Supplement_1 (Dec. 2021), S17–S38.

[4] Batista, G., Prati, R., and Monard, M.-C. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations 6* (06 2004), 20–29.

[5] Bonora, E., and Tuomilehto, J. The pros and cons of diagnosing diabetes with a1c. *Diabetes Care 34*, Supplement_2 (Apr. 2011), S184–S190.

[6] Breiman, L. Bagging predictors. *Machine learning 24* (1996), 123–140.

[7] Breiman, L. Random forests. *Machine learning 45* (2001), 5–32.

[8] Chen, T., and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Aug. 2016), KDD '16, ACM, p. 785–794.

[9] Fowler, M. J. Microvascular and macrovascular complications of diabetes. *Clinical Diabetes 26*, 2 (Apr. 2008), 77–82.

[10] Ghorbani, R., and Ghousi, R. Comparing different resampling methods in predicting students' performance using machine learning techniques. *IEEE Access 8* (2020), 67899–67911.

[11] Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A., and Rueckert, D. Random forest-based similarity measures for multi-modal classification of alzheimer's disease. *NeuroImage 65* (Jan. 2013), 167–175.

[12] Kotsiantis, S., Kanellopoulos, D., and Pintelas, P. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering 30* (11 2005), 25–36.

[13] Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., and Hamprecht, F. A. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics 10*, 1 (July 2009).

[14] Mohan, S., Thirumalai, C., and Srivastava, G. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access 7* (2019), 81542–81554.

[15] Public Health Agency of Canada. Diabetes: Prevention and risk factors - Canada.ca — canada.ca. `https://www.canada.ca/en/public-health/services/chronic-diseases/diabetes/prevention-risk-factors.html`. [Accessed 13-12-2024].

[16] Sapra, A., and Bhandari, P. Diabetes. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), Jan. 2024.

[17] Scikit-Learn. StandardScaler — scikit-learn.org. `https://scikit-learn.org/1.5/modules/generated/sklearn.preprocessing.StandardScaler.html`. [Accessed 11-12-2024].

[18] Sharmila, A., and Geethanjali, P. Dwt based detection of epileptic seizure from eeg signals using naive bayes and k-nn classifiers. *IEEE Access 4* (2016), 7716–7727.

[19] Sowers, J. R. Diabetes mellitus and vascular disease. *Hypertension 61*, 5 (May 2013), 943–947.

[20] Tarwidi, D., Pudjaprasetya, S. R., Adytia, D., and Apri, M. An optimized xgboost-based machine learning method for predicting wave run-up on a sloping beach. *MethodsX 10* (2023), 102119.

[21] Teboul, A. Cdc diabetes health indicators. `https://archive.ics.uci.edu/dataset/891`, 2017. [Accessed 13-11-2024].

[22] US Food and Drug Administration. How Smoking Can Increase Risk for and Affect Diabetes — fda.gov. `https://www.fda.gov/tobacco-products/health-effects-tobacco-use/how-smoking-can-increase-risk-and-affect-diabetes`. [Accessed 13-12-2024].

[23] Wang, H., Liu, Y., Zhou, B., Li, C., Cao, G., Voropai, N., and Barakhtenko, E. Taxonomy research of artificial intelligence for deterministic solar power forecasting. *Energy Conversion and Management 214* (2020), 112909.

[24] World Health Organization. Diabetes — who.int. `https://www.who.int/health-topics/diabetes`. [Accessed 13-11-2024].

[25] Wu, Y.-c., and Feng, J.-w. Development and application of artificial neural network. *Wireless Personal Communications 102*, 2 (Dec. 2017), 1645–1656.

[26] Yadav, S. S., and Jadhav, S. M. Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big Data 6*, 1 (Dec. 2019).

[27] Zeng, C., Xu, Y., and Tian, J. Analyze the robustness of classifiers under label noise. *arXiv preprint arXiv:2312.07271* (2023).

[28] Zhang, J., Zhang, Z., Zhang, K., Ge, X., Sun, R., and Zhai, X. Early detection of type 2 diabetes risk: limitations of current diagnostic criteria. *Frontiers in Endocrinology 14* (Nov. 2023).