Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240491

Balancing Acts: Tackling Data Imbalance in Machine Learning for Predicting Myocardial Infarction in Type 2 Diabetes

Berk OZTURK[a,1], Tom LAWTON[a,b], Stephen SMITH[a] and Ibrahim HABLI[a]
[a] *University of York, York, YO10 5GH, UK*
[b] *Bradford Teaching Hospitals NHS Foundation Trust, Bradford, BD9 6RJ, UK*

Abstract. Type 2 Diabetes (T2D) is a prevalent lifelong health condition. It is predicted that over 500 million adults will be diagnosed with T2D by 2040. T2D can develop at any age, and if it progresses, it may cause serious comorbidities. One of the most critical T2D-related comorbidities is Myocardial Infarction (MI), known as heart attack. MI is a life-threatening medical emergency, and it is important to predict it and intervene in a timely manner. The use of Machine Learning (ML) for clinical prediction is gaining pace, but the class imbalance in predictive models is a key challenge for establishing a trustworthy deployment of the technology. This may lead to bias and overfitting in the ML models, and it may cause misleading interpretations of the ML outputs. In our study, we showed how systematic use of Class Imbalance Handling (CIH) techniques may improve the performance of the ML models. We used the Connected Bradford dataset, consisting of over one million real-world health records. Three commonly used CIH techniques, Oversampling, Undersampling, and Class Weighting (CW) have been used for Naive Bayes (NB), Neural Network (NN), Random Forest (RF), Support Vector Machine (SVM), and Ensemble models. We report that CW overperforms among the other techniques with the highest Accuracy and F1 values of 0.9948 and 0.9556, respectively. Applying the most appropriate CIH techniques for the ML models using real-world healthcare data provides promising results for helping to reduce the risk of MI in patients with T2D.

Keywords. Type 2 diabetes, heart attack, machine learning, dataset, class imbalance

1. Introduction

Type 2 Diabetes (T2D) is a lifelong health condition in which the body is unable to regulate glucose levels, due to the body's resistance to insulin or the pancreas not producing enough insulin [1]. As of 2021, nearly 250 million adults have lived with T2D, and its prevalence is increasing steadily [2]. T2D may occur at any age, and if not managed, continues to progress [3]. The progression of T2D may cause a range of serious comorbidities with one of the most critical being Myocardial Infarction (MI), known as heart attack [4]. MI is a medical emergency that occurs when the blood flow to the heart is suddenly blocked, posing life-threatening risks for the patient [5]. MI in T2D patients may occur at any time and is challenging to predict its occurrence in advance [3].

[1] Corresponding Author: Berk Ozturk,Tel: +44 751 337 5115.
E-mail: berk.ozturk@york.ac.uk (B. Ozturk), tom.lawton@bthft.nhs.uk (T. Lawton), stephen.smith@york.ac.uk (S. Smith), ibrahim.habli@york.ac.uk (I. Habli).

Studies show that early interventions may reduce or prevent the risk of MI in patients with T2D [6]. Making credible predictions is crucial for early interventions, and there are various clinical methodologies to predict and manage the risk of MI [7]. In the UK, the National Health Service (NHS) and National Institute for Health and Care Excellence (NICE) have published guidelines for early diagnosis and treatment methods [5]. However, early prediction and effective management of MI risk remain challenging, and the use of supplementary tools can potentially be beneficial in overcoming this challenge.

Since the early 2010s, AI's promising results in healthcare have increased its use in predicting MI among T2D patients using various Machine Learning (ML) models. [8]. Despite its potential, leveraging AI in healthcare poses clinical, technical and organizational challenges, particularly with real-world health datasets [8]. Data imbalance, where certain classes are over or underrepresented, is inevitable in these datasets, potentially leading to biased or overfitted ML models [9]. Addressing this issue, our study investigates class imbalance using different techniques and develops a predictive AI-based model for the risk of developing MI in patients with T2D.

2. Methods

2.1. Data Preprocessing

In our study, the RStudio-Caret Package has been employed for data preprocessing, class imbalance handling, ML implementation, and result evaluations with default parameters [2]. We used the Connected Bradford (CB) dataset, a pseudonymized dataset linking primary and secondary healthcare records for over 1 million people around Bradford, UK. We constructed a subset in a table format, consisting of 1,058,139 patient entries as rows and over 14,000 columns as features [9].

We filtered the CB dataset for patients with T2D, generating an output column that consists of MI-related entries based on OpenCodelists, a collection of clinical codes (ctv3 codes) that classifies patients as having certain conditions or demographic properties [2]. We searched for ctv3 codes that identified MI and marked patients in our dataset accordingly. However, as the resulting dataset was too large to handle and train ML models, we removed all the variables with more than 10% missing data for the specified features. This resulted in a dataset with 69,075 rows (unique patients) and 24 variables consisting of various demographic data and biomarkers (see Figure 1). Any remaining missing quantitative values were imputed via bagged trees. Bagged trees generate multiple imputed datasets, each containing estimations for missing values, and then combine the results for the final estimations [4]. This imputation method can be used when the input and output variables are numeric or categorical. Imputed categorical variables retain their categorical nature. Then, the qualitative values were encoded, because all the non-numeric values have to be converted to numeric values by zeros and ones before training ML models [1]. Lastly, the input variables in the new dataset were scaled between zero and one, to prevent the optimization processes from being dominated by features with larger scales [8].

2.2. Handling Class Imbalance & ML Implementation

After completing the data preprocessing steps, we had a highly imbalanced output column with ratios of 90% and 10%, respectively, signifying that the majority of the

patients in the dataset have a ctv3 code representing MI at some point in their data, leading to data imbalance. Therefore, we used the three most common data imbalance handling methods, Class Weighting (CW), Oversampling, and Undersampling. For the CW, the output classes have been weighted with 0.1 and 0.9 respectively, based on their representation ratio. For Oversampling, the class "No" has been duplicated to achieve class balance. Lastly, randomly selected data from class "Yes" has been removed from the dataset to balance the two classes for Undersampling. As a result, we had three new datasets, and the main dataset with no Class Imbalance Handling (CIH) technique. Then, the most common ML models for the risk prediction of MI in patients with T2D were used with the default hyperparameters [7]: Naive Bayes (NB) with classification threshold set to 0.5, Neural Network (NN) with a single-hidden-layer architecture, one-third the number of inputs, and 100 iterations, Random Forest (RF) with 500 trees and one terminal node, and Support Vector Machine (SVM) with radial kernel. The four datasets, including one with no data imbalance handling method, were split into 80% for training and 20% for testing. In addition, randomly selected 20% of the training set was used for validation, and it was used for hyperparameter tuning, employing tune-length of 5 to optimize performance. Also, k-fold cross-validation with 5 folds was implemented to mitigate bias and overfitting. Next, for the Ensemble model, Generalized Linear Model (GLM) with logistic regression was used, and four ML models were ensembled to enhance the performance of the predictive models [3].

3. Results

Table 1 shows the Accuracy and F1 values, default performance metrics in R-Caret [2], for both ML models before CIH (Pre-CIH) and after CIH techniques. Accuracy is the performance value of the closeness of the predicted value to the known value [2]. The F1 is another metric to evaluate the performance of the ML-based classification models, particularly in dealing with imbalanced datasets [10]. The F1 is the harmonic mean of the accuracy of the positive predictions and the ability of the ML models to detect all the positives, and it is a metric used to see the model's ability to minimize the False Negatives and False Positives [10].

Table 1. Results of performance metrics of each CIH method for each ML model

	Pre-CIH		**Class Weighting**		**Oversampling**		**Undersampling**	
ML	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**
NB	0.9762	0.8893	0.9848	0.9051	0.9066	0.8365	0.9437	0.8950
NN	0.9714	0.9036	0.9772	0.9087	0.9594	0.8211	0.9596	0.8775
RF	0.9885	0.9145	0.9917	0.9265	0.9812	0.9263	0.8612	0.9006
SVM	0.9779	0.8566	0.9801	0.8588	0.8582	0.8144	0.8963	0.8254
Ensemble	0.9892	0.9224	0.9948	0.9556	0.9854	0.9280	0.9568	0.9388

Figure 1 shows the importance level of input variables in the predicted class. These importance levels represent the Shapley value of each feature for presenting the relative impact of each feature we measure on the eventual output of the ML model by comparing the relative effect of the inputs against the average. These are calculated for the most performant method, the Ensemble model using the CW. In this model, increased sodium and diastolic blood pressure have the most negative impact, and increased potassium or being of Pakistani ethnicity has the most positive impact on the predicted risk of MI.

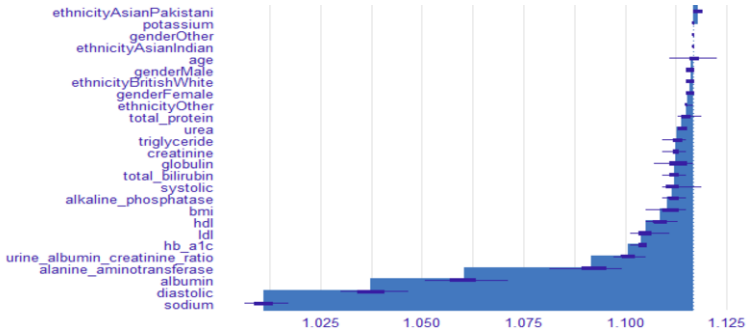

Figure 1. The importance level of each input variable

4. Discussion

In Table 1, it is seen that CW has the highest Accuracy and F1 values among the used ML models with 0.9948 and 0.9556, respectively. Higher Accuracy may provide an impression that the ML model classifies the patient's MI with high performance. However, a high Accuracy value may not be sufficient to make decisions about the patient's MI risk. Hence, a higher F1 showing that the ML model minimizes the risk of misclassifying (i.e. labeling a patient "No MI risk" despite MI risk or labeling a patient "Yes" with no actual MI risk) may help to provide more convincing results for clinical use. This also potentially prevents missing or overestimating the risk of MI, which may lead to incorrect interventions. In addition, the potential of obtaining improved results when using Ensemble ML models is another noticeable outcome of this study.

In Figure 1, it is important to note that each variable in the dataset may have a different impact on the output, and this visual information may give better insight for understanding the reasons behind the decisions made by ML-based models. This also provides an opportunity to make interactions between the clinical users and the ML in the decision-making stage for interventions.

5. Conclusions

MI is a serious and complex clinical condition in the CB dataset. It is important to make accurate predictions so that appropriate interventions can be made to prevent the risk of developing MI in patients with T2D before the event occurs. However, it has been noticed that healthcare datasets may have imbalanced classes because of the nature of the real-world data, and this challenges ML-based models to make appropriate predictions. When CIH methods are used, the performance of the predictive ML models can potentially improve. In addition, the performance of these models may be further improved by using the Ensemble model, and the outcomes can become more explainable by visualizing the importance of the features used for output prediction. Because there is no established way to develop ML models with extremely imbalanced data, investigating the most relevant preprocessing, CIH, and ML methods for developing ML models is crucial [11]. As further work, our goal is to improve the robustness and human-centric explainability of our ML model and importantly develop a Safety Case with clinical hazard analysis dedicated to the clinical workflow and setting [8, 12]. We believe this

will contribute to developing trustworthy ways for safe and predictive ML models for clinical decision-making in complex healthcare settings.

Acknowledgment

This study is based on data from Connected Bradford (REC 18/YH/0200 & 22/EM/0127). The data is provided by the citizens of Bradford and district, and collected by the NHS, DfE and other organizations as part of their care and support. The interpretation and conclusions contained in this study are those of the authors alone. The NHS, DfE and other organizations do not accept responsibility for inferences and conclusions derived from their data by third parties. This work was supported by the Assuring Autonomy International Programme, a partnership between Lloyd's Register Foundation and the University of York, and the UKRI project (EP/W011239/1) "Assuring Responsibility for Trustworthy Autonomous Systems".

References

[1] Sudharsan B, Peeples M, Shomali M. Hypoglycemia prediction using machine learning models for patients with type 2 diabetes. Journal of diabetes science and technology. 2014 Oct 14;9(1):86-90.

[2] Ozturk B, Lawton T, Smith S, Habli I. Predicting Progression of Type 2 Diabetes Using Primary Care Data with the Help of Machine Learning. InCaring is Sharing–Exploiting the Value in Data for Health and Innovation 2023 May 18 (pp. 38-42). IOS Press. doi: 10.3233/SHTI230060

[3] Alonso-Morán E, Orueta JF, Esteban JI, Axpe JM, González ML, Polanco NT, Loiola PE, Gaztambide S, Nuño-Solinis R. The prevalence of diabetes-related complications and multimorbidity in the population with type 2 diabetes mellitus in the Basque Country. BMC public health. 2014 Dec;14:1-9.

[4] Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific reports. 2020 Jul 20;10(1):11981.

[5] Moore A, Bell M. XGBoost, A Novel Explainable AI Technique, in the Prediction of Myocardial Infarction: A UK Biobank Cohort Study. Clinical Medicine Insights: Cardiology. 2022 Nov;16:11795468221133611. doi: 10.1177/1179546822113361

[6] Mani S, Chen Y, Elasy T, Clayton W, Denny J. Type 2 diabetes risk forecasting from EMR data using machine learning. InAMIA annual symposium proceedings 2012 (Vol. 2012, p. 606). American Medical Informatics Association.

[7] Ryan Conmy P, Ozturk B, Lawton T, Habli I. The Impact of Training Data Shortfalls on Safety of AI-based Clinical Decision Support Systems. InInternational Conference on Computer Safety, Reliability, and Security 2023 Sep 11 (pp. 213-226). Cham: Springer Nature Switzerland.

[8] Davahli MR, Karwowski W, Fiok K, Wan T, Parsaei HR. Controlling safety of artificial intelligence-based systems in healthcare. Symmetry. 2021 Jan 8;13(1):102. doi: 10.3390/sym13010102

[9] Sohal K, Mason D, Birkinshaw J, West J, McEachan RR, Elshehaly M, Cooper D, Shore R, McCooe M, Lawton T, Mon-Williams M. Connected Bradford: a whole system data linkage accelerator. Wellcome Open Research. 2022;7. doi: 10.12688/wellcomeopenres.17526.2

[10] Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, De Cata P, Chiovato L, Bellazzi R. Machine learning methods to predict diabetes complications. Journal of diabetes science and technology. 2018 Mar;12(2):295-302. doi: 10.1177/1932296817706375

[11] Hawkins R, Paterson C, Picardi C, Jia Y, Calinescu R, Habli I. Guidance on the assurance of machine learning in autonomous systems (AMLAS). arXiv preprint arXiv:2102.01564. 2021 Feb 2.

[12] Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bulletin of the World Health Organization. 2020 Apr 4;98(4):251. doi: 10.2471/BLT.19.237487