AI-based Models for Predicting the Risk of Developing Diabetes

Sotiris Batsakis
Hellenic Mediterranean University, Greece
and University of Huddersfield, UK
Email: s.batsakis@hud.ac.uk

Bakhtiar Amen
City University of London, UK
Email: bakhtiar.amen@city.ac.uk

Grigoris Antoniou
Leeds Beckett University
Leeds, UK
Email: G.Antoniou@leedsbeckett.ac.uk

Steven Howard
Locala Health & Wellbeing
Batley, UK
steven.howard@locala.org.uk

Paul Rhys-Vivian
Locala Health & Wellbeing
Batley, UK
paul.rhysvivian@locala.org.uk

Gemma Gardner
Locala Health & Wellbeing
Batley, UK
gemma.gardner@locala.org.uk

***Abstract*—Diabetes is a serious condition affecting hundreds of millions of people worldwide. Diabetic risk prediction is an important task and preliminary risk assessments are crucial for clinicians for achieving prioritization of diabetes cases and for examining cases with higher risk first In this work we collect and analyse podiatry data related to diabetes prediction and develop and evaluate Machine Learning models using the created dataset for providing preliminary assessments of diabetes risk and help clinicians on the task of prioritizing cases according to the predicted risk.**

I. INTRODUCTION

Diabetes is a serious condition that affects various parts of the body including feet, legs, liver, kidneys and eyes [1]. It is estimated that more than 400 million people worldwide have diabetes making it one of the main contemporary health risks and the total health expenditure related to diabetes exceeds $700 billion [2]. The rising number of people with diabetes and the severity of the condition has led to the adoption of Artificial Intelligence methods and Machine Learning (ML) tools in particular in order to create prediction models for diagnosing diabetes [3], [4]. In general, there are different types of diabetes complications, however, the primarily aim of this work is to develop AI-based solution for an early detection of developing diabetes based on podiatry data and specifically on various clinical exams (non-invasive risk assessment). Traditionally, this type of assessment is performed by healthcare or podiatrist experts using paper-based screening assessment tools. At the conclusion of the assessment, an expert can determine whether the patient has a low, medium, or high chance of developing foot diabetes risk. However, it is important to note that these clinical examinations are time-consuming to complete. Therefore, an early detection and prediction of developing foot diabetes risk is shown to be the most effective method and beneficial for number of reasons ; (a) preventing patient from Diabetes Foot Ulcers (DFUs) and or leg amputation [5] , (b) on-time decision-making and prioritising a high risk patient, (c) improving waiting times in the long waiting lists of patients to be examined by healthcare experts.

To accomplish the aforementioned goal, in this work we propose a novel solution based on utilising different machine learning (ML) models to automatically classify and predict the risk of developing diabetic foot (i.e., low, medium and high risk classification). The patients' data is collected by the Locala Health and Well-being organisation[1] in partnership with UK's National Health Service (NHS). For the empirical result, we test and compare various ML models under different settings and criteria and report the best results obtained.

II. RELATED WORK

Over the past years, many model-centric AI approaches, in particular, ML models have been used to support automated diagnosis and prediction of various types of diabetes cases including [6] for type II diabetes,[7] analysing the open PIMA Indians UCI dataset, [8] using physical examination data , and [9], [10] applied on the same dataset as [7]. In [11] ML methods were applied for type II diabetes diagnosis, the work in [12] was based on the PIMA Indians dataset and in [13] three open datasets were analyzed including the PIMA Indians dataset. Notice that the performance of the above-mentioned papers is not typically comparable due to their use on different dataset and in case of a Pima Indians datasets the reported performance was on similar levels across different works. In this work we create a new dataset with a focus on podiatry data and the common ML methods proposed in previous work have bee applied on the new dataset.

Compared to existing work the current paper has the following contributions (a) a new dataset is proposed, instead of using existing datasets, (b) data is collected during initial assessments instead of detailed exams, thus contributing to preliminary prioritization of cases, (c) this work focuses on podiatry data which is not the case of the vast majority of existing work and (d) emphasis is given not only to achieving

[1]https://www.locala.org.uk/

high accuracy but also minimizing the number of critical (false negative) diagnostic errors which have more dangerous consequences related to false positive diagnostic errors. To the best of our knowledge this is the first related work emphasizing this important issue.

III. METHODOLOGY

This section describes the methodology for developing the proposed AI-based prediction models and the dataset used for the analysis. All patient records in the dataset were anonymized thus, having no information that can potentially lead to identification of individuals using the electronic health records.

A. Dataset Construction

Data from clinical practice became gradually available in two batches during this work and various prediction models were also developed gradually. The first batch of data consisted of two cohorts; first data cohort contained 96 cases with multiple rows per patient (one row per exam) for a total of 11296 rows. The second cohort of the first batch consisted of 11206 rows for 100 distinct cases. The objective of the entire process was to predict diabetes risk given the data which consists of demographic (e.g. age, ethic group) and medical data such as preliminary exam data. The second batch made available later during the analysis process, was also prepared in two stages. The first cohort of the second batch contained 491 cases in one row per patient format and 54413 rows when multiple rows per patient format is used, each row containing a single exam. The second cohort of the second batch at the final stage contained 748 cases in single row per case format (containing most recent exams) which also included the cases of the first cohort of the second batch and also all data from the first batch. Since the final dataset is a superset of the preliminary versions, only experiments on the final version are included in this work and preliminary experiments with subsets of the final data are not included, although these experiments provided insights for the presented analysis on the full dataset. .

The final dataset consists of 748 instances, 418 cases of male patients and 330 female patients. 60 cases are characterized as high risk (risk level 1), 420 as medium risk (risk level 2) and 232 are low risk (risk level 3) while for 36 cases the risk classification was missing and they were not considered further in the experimental process. During the dataset construction several issues where identified: many features have a lot of missing data (e.g. 81% missing data for HbA1c and 66% missing data for the neuropathy feature presented in the following) while details for several features are missing too, e.g., alcohol use corresponds to present, while in the past the use between two persons that non longer drink may be very different, thus, finer distinctions when collecting data may be helpful for developing more accurate models. The dataset features are presented in the following.

B. Dataset Features

The developed dataset variables set consists of patient's age on referral (high risk cases correlate with older age), gender, ethnicity, housing status, referral primary reason, smoking history (e.g., smoker, ex-smoker and non-smoker), alcoholBIscore (values 0 to 12), alcohol referral and HbA1c scores (blood sugar level test with values 36 to 160). Importantly, higher values of HbA1c correlate with higher risk. Other features are Neuropathy (e.g., sciatica value indicated low or medium risk), right/left foot (R/L) dorsalis pulse which is a categorical feature (one value for each foot) with possible values being monophasic, bi-phasic and tri-phasic pulse. Triphasic pulse typically indicates low or medium risk. Additional features are R/L posterior tibial pulses: categorical feature with possible values being monophasic, bi-phasic and tri-phasic pulse, again tri-phasic values correlates with low or medium risk, R/L monofilament sensation (having values normal and abnormal, R/L vibration, R/L Diabetic foot risk (with values of low, medium and high risk). However, since overall risk is not available in the final dataset version, the higher value of the last two features (R/L foot risk) is used as proxy of the overall risk assessment value. History of foot ulcer indicates high risk, limb amputation indicates high risk, cracked skin correlates with high risk, and dry skin correlates (not strongly) with higher risk.

More features were included into the dataset in addition to the above features, but these features, specifically patient ID, GP practice center and dates for each exam were removed during the preprocessing because they are not correlated with the risk level and also for anonymization purposes.

C. Analysis Process

The overall architecture of the developed framework comprised four phases The first phase corresponds to collecting anonymized clinical data containing podiatric exams and simple demographic features. The second phase represents the data preprocessing task to clean, filter and select related features. The third and fourth phase are building ML models for predicting patients with a higher risk of developing diabetes causing foot problems and then detecting these patients during a preliminary assessment and prioritize their cases for further examinations.

IV. EXPERIMENTS

The first stage of the analysis involved data preparation, specifically dataset construction, removal of features not related to the target feature (risk level) such as patient ID, examination center and dates of exams and target feature construction (i.e., the target feature-overall risk level- was constructed by combining the left and right foot risk, specifically by selecting the higher risk level among these).

The construction of Machine Learning models for risk prediction was done using the WEKA machine learning tool[14] supporting numerous ML algorithms that were compared during the experiment process. Data imputation and normalization are handled by WEKA algorithm implementation, also for

all algorithms tested the default hyperparameters were used and the performance evaluation protocol was 10-fold cross validation.

The algorithms evaluated were rule based such as ZeroR (selecting the mode in the dataset as prediction and ignoring the input features) being the baseline, OneR and JRIP. Decision Tree algorithms such as J48 and Hoeffding Tree along with Random Forest are also evaluated. Logistic Regression, SMO (SVM implementation of WEKA) and Artificial Neural Networks (ANN) were included in the experimental process along with the Naive Bayes classifier and meta-classifiers (AdaBoost and Bagging). Performance metrics are precision (percentage of correctly classified instances for all risk levels) and recall for high risk cases, which corresponds to the percentage of high risk cases that were correctly classified. The experimental procedure included both interpetable algorithms such as ZaroR, OneR, JRIP, J48, Hoeffding Tree, Logistic Regression and Naive Bayes and non interpretable algorithms such as Random Forest, SMO, ANN, AdaBoost and Bagging. The results of the experiments with the aforementioned algorithms are presented in Table I.

TABLE I
SUMMARY OF INITIAL EXPERIMENTAL RESULTS WITH EQUAL COST FOR ERROR TYPES

Algorithm	Precision	Recall(High Risk)
ZaroR	0.59	0.00
OneR	0.64	0.00
JRIP	0.67	0.18
J48	0.65	0.00
Hoeffding Tree	0.62	0.15
Random Forest	0.69	0.02
Logistic Regression	0.68	0.03
SMO	0.68	0.08
ANN	0.65	0.25
Naive Bayes	0.68	0.35
AdaBoost	0.64	0.00
Bagging	0.68	0.00

The top performance was achieved using Random Forest (RF) in terms of total accuracy but RF does not produce an interpretable model which is an important requirement of deployed AI systems, especially in critical applications such as medical diagnosis. Among the interpetable algorithms there was not an algorithm clearly dominating the rest of interpetable algorithms. In terms of total precision Logistic Regression and Naive Bayes were the top performing interpetable algorithms followed closely by JRIP.

Besides total accuracy another issue that is of great importance is the type of errors and not just the overall accuracy. Specifically classification errors for medium and low risks cases are not as important as errors of high risk cases and false negative errors of high risk cases in particular, because in such a case a high risk patient is not given the corresponding high priority required given the existing health risks. When examining false negative errors for high risk cases the total number of errors (among the 60 high risk cases) were (per algorithm): ZeroR=60, OneR=60, JRIP=49, J48=60, Hoeffding Tree=51, Random Forest=59, Logistic Regression=58, SMO=45, ANN=45, Naive Bayes=39, AdaBoost=60 and Bagging=60. This indicated that all algorithms,including those achieving best overall accuracy, failed to minimize the number of critical errors making all aforementioned models initially evaluated on the basis of total accuracy unsuitable for the patient prioritization task. This is illustrating by the very low recall for high risk cases of all algorithms, with Naive Bayes having the best performance, but still low, for high risk cases recall.

For dealing with this important issue the cost sensitive meta classifier of WEKA was used, combined with all classifiers of Table I but assigning higher costs for false negative errors in high risk cases. The total accuracy is again reported along with True Positive Rate (Recall) for high risk cases which is maximized when False Negative Rate is minimized. The results obtained when cost of classifying high risk cases as medium and low risk are 20 and 40 times higher than other types of errors (first cost sensitive configuration) are summarized in Table II.

TABLE II
SUMMARY OF ERROR TYPE SENSITIVE RESULTS, FIRST CONFIGURATION

Algorithm	Precision	Recall(High Risk)
ZaroR	0.08	1.0
OneR	0.12	0.93
JRIP	0.25	0.92
J48	0.45	0.65
Hoeffding Tree	0.08	1.0
Random Forest	0.49	0.58
Log. Regression	0.11	0.97
SMO	0.40	0.53
ANN	0.63	0.23
Naive Bayes	0.47	0.70
AdaBoost	0.34	0.75
Bagging	0.49	0.47

Some of the classifiers in the cost sensitive setting such as the baseline ZeroR and Hoeffding Tree minimized the critical errors by classifying all cases as high risk, thus having very low total accuracy and consequently being of no practical use. Similar results were obtained by OneR, Logistic Regression and JRIP, while on the other extreme ANN still has a very high rate of critical errors. J48, Random Forest, SMO, Naive Bayes, AdaBoost and Bagging produced more balance models by having higher total accuracy (still bellow, although close, to 50%) and a recall of high risk cases above 50% (with the exception of Bagging). Among those Naive Bayes and J48 are interpetable, which is an important requirement for medical diagnostic systems, thus being the most promising for developing prioritization systems based on risk estimation for diabetes.

Using a second configuration with 10 and 15 cost ratio of high risk classification errors (second configuration) instead of 20 and 40, leads to the results of Table III.

Using the second configuration ZeroR and Hoeffding Tree still didn't work, while OneR, J48. ANN, SMO, AdaBoost and Bagging produced similar results as with the first con-

TABLE III
SUMMARY OF ERROR TYPE SENSITIVE RESULTS, SECOND CONFIGURATION

Algorithm	Precision	Recall(High Risk)
ZaroR	0.08	1.0
OneR	0.21	0.81
JRIP	0.35	0.78
J48	0.44	0.68
Hoeffding Tree	0.08	1.0
Random Forest	0.62	0.38
Log. Regression	0.41	0.70
SMO	0.52	0.48
ANN	0.65	0.25
Naive Bayes	0.55	0.58
AdaBoost	0.32	0.75
Bagging	0.47	0.53

figuration. The remaining algorithms produced models with increased total accuracy compared to that of the first configuration but reduced recall for high risk cases (increased number of critical errors). Thus these algorithms (JRIP, Random Forest, Logistic Regression and Naive Bayes) can be adjusted using the cost ratio of various types of errors as parameter to produce a desired prioritization policy representing a trade off between total accuracy and minimization of critical errors. Naive Bayes for example can be used for deriving a model using the second configuration that achieves total accuracy exceeding 50% while having recall for high risk cases also above 50%.

The proposed model using Naive Bayes and the second configuration can achieve better prioritization than random First Come First Served basis by having total accuracy exceeding 50% and identifying correctly the majority of high risk cases, but bigger datasets and improved models are still required for large scale deployment for such a system.

V. CONCLUSIONS

In this work, diabetes risk prediction models were developed using podiatry data for patient prioritization. During the dataset construction several issues where identified related to missing data and missing distinctions in some categorical features, issues resulting to reducing the accuracy of the developed models. Still the experimental results are promising since proposed models can achieve performance better than a random first come first served policy and further improvements are expected when more data are collected. Another issue identified during the analysis is the type of errors to minimize and corresponding experiments on cost sensitive classifiers have been conducted. It is expected that in future large scale deployments of such diagnostic systems the developed models will be based on cost sensitive classifiers, instead of models that maximize total accuracy without special treatment of various types of errors.

Directions for future work can be summarized as follows: (a) improving the dataset by including more cases as they become available and by making finer distinctions in various features during the data collection process, gradually increasing the predictive performance of machine learning based models and (b) combining models generated from data using machine learning with knowledge based models based on rules created with the help of clinicians and then combining these two models in a hybrid setting.

REFERENCES

[1] G. Roglic, "Who global report on diabetes: A summary," *International Journal of Noncommunicable Diseases*, vol. 1, no. 1, pp. 3–8, 2016.

[2] E. L. Feldman, B. C. Callaghan, R. Pop-Busui, D. W. Zochodne, D. E. Wright, D. L. Bennett, V. Bril, J. W. Russell, and V. Viswanathan, "Diabetic neuropathy," *Nature reviews Disease primers*, vol. 5, no. 1, p. 41, 2019.

[3] K. De Silva, W. K. Lee, A. Forbes, R. T. Demmer, C. Barton, and J. Enticott, "Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis," *International journal of medical informatics*, vol. 143, p. 104268, 2020.

[4] T. Howard, R. Ahluwalia, and N. Papanas, "The advent of artificial intelligence in diabetic foot medicine: A new horizon, a new order, or a false dawn?" *The International Journal of Lower Extremity Wounds*, vol. 22, no. 4, pp. 635–640, 2023.

[5] D. G. Armstrong, A. J. Boulton, and S. A. Bus, "Diabetic foot ulcers and their recurrence," *New England Journal of Medicine*, vol. 376, no. 24, pp. 2367–2375, 2017.

[6] A. Dagliati, S. Marini, L. Sacchi, G. Cogni, M. Teliti, V. Tibollo, P. De Cata, L. Chiovato, and R. Bellazzi, "Machine learning methods to predict diabetes complications," *Journal of diabetes science and technology*, vol. 12, no. 2, pp. 295–302, 2018.

[7] M. Maniruzzaman, N. Kumar, M. M. Abedin, M. S. Islam, H. S. Suri, A. S. El-Baz, and J. S. Suri, "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm," *Computer methods and programs in biomedicine*, vol. 152, pp. 23–34, 2017.

[8] H. Yang, Y. Luo, X. Ren, M. Wu, X. He, B. Peng, K. Deng, D. Yan, H. Tang, and H. Lin, "Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators," *Information Fusion*, vol. 75, pp. 140–149, 2021.

[9] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of diabetes using machine learning algorithms in healthcare," in *2018 24th international conference on automation and computing (ICAC)*. IEEE, 2018, pp. 1–6.

[10] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76 516–76 531, 2020.

[11] A. Bernabe-Ortiz, P. Perel, J. J. Miranda, and L. Smeeth, "Diagnostic accuracy of the finnish diabetes risk score (findrisc) for undiagnosed t2dm in peruvian population," *Primary care diabetes*, vol. 12, no. 6, pp. 517–525, 2018.

[12] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima indians diabetes mellitus classification based on machine learning (ml) algorithms," *Neural Computing and Applications*, vol. 35, no. 22, pp. 16 157–16 173, 2023.

[13] L. Ismail, H. Materwala, M. Tayefi, P. Ngo, and A. P. Karduck, "Type 2 diabetes with artificial intelligence machine learning: methods and evaluation," *Archives of Computational Methods in Engineering*, pp. 1–21, 2022.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.