Machine Learning for Risk Stratification of Diabetic Foot Ulcers using Biomarkers

Kyle Martin[0000−0003−0941−3111], Ashish Upadhyay[0000−0003−0124−8879], Anjana Wijekoon[0000−0003−3848−3100], Nirmalie Wiratunga[0000−0003−4040−2496], Stewart Massie[0000−0002−5278−4009]

School of Computing, Robert Gordon University, Aberdeen, Scotland
{k.martin3, a.upadhyay, a.wijekoon1, n.wiratunga, s.massie}@rgu.ac.uk
https://rgu-repository.worktribe.com/tag/1437329/artificial-intelligence-reasoning-air

Abstract. Development of a Diabetic Foot Ulcer (DFU) causes a sharp decline in a patient's health and quality of life. The process of risk stratification is crucial for informing the care that a patient should receive to help manage their Diabetes before an ulcer can form. In existing practice, risk stratification is a manual process where a clinician allocates a risk category based on biomarker features captured during routine appointments. We present the preliminary outcomes of a feasibility study on machine learning techniques for risk stratification of DFU formation. Our findings highlight the importance of considering patient history, and allow us to identify biomarkers which are important for risk classification.

Keywords: Diabetic Foot Ulceration · Machine Learning · Biomarkers

1 Introduction

Diabetic Foot Ulcers (DFUs) are a severe complication of Diabetes Mellitus. It is estimated that 15% of diabetic patients will develop a DFU during their lives [12]. Development of a DFU can cause a sharp decline in health and quality of life, often leading to further infection, amputation and death [10, 3]. Predicting the likelihood of DFU formation is based on risk stratification.

Risk stratification is crucial for informing the level and regularity of care that a patient should receive. Improper treatment of a DFU can exacerbate patient condition and lead to further health complications, impacting quality of life and increasing cost of treatment [3]. Given medical knowledge of biological markers which act as patient features, clinicians leverage their domain expertise to manually allocate a risk category which describes the likelihood of developing a DFU [1]. Biological markers (henceforth 'biomarkers', as per domain terminology) are physiological features captured during routine medical examinations which contribute to a clinician's understanding of patient condition and their capability to effectively stratify future risk. For example, concentration of albumin in the blood is a recognised indicator of Diabetes and its complications [5]. Of specific interest are biomarkers which describe the onset of peripheral neuropathy (i.e. damaged nerve endings in the extremities [2]).

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_11

Risk stratification is therefore an expertise-driven task with good potential for automation using machine learning. However, few existing works have used patient health records for this purpose. The authors in [11] produced a review of journal papers describing applications of machine learning algorithms for the diagnosis, prevention and care of DFUs. They surveyed 3,769 papers (reduced to 37 after application of inclusion and exclusion criteria) and found the majority of ML algorithms have been applied to thermospectral or colour images of the foot (29 papers). Only a single reviewed paper examined patient health records. In [4] the authors collated a dataset of 301 patient records from a hospital in India and trained a decision tree for the purposes of explaining amputations caused by DFUs. Key differences to our work include: (a) our work is on a much larger dataset - we apply machine learning algorithms to health records for approximately 27,000 individual patients; (b) our work is targeted towards risk stratification of ulcer formation (predictive) whereas [4] aimed to explain amputation decisions due to DFUs (retrospective); and (c) the works are applied in the context of different health systems.

In [8] the authors describe a method of predicting DFU formation (and subsequent amputation) using national registry data for 246,000 patients in Denmark. Their dataset is formed from socio-economic features, with some knowledge of concurrent health conditions. Though their results are not comparable to those we present here, due to differences in features and task, they emphasise the limitations of high-level data. In their results, they stated that knowledge of DFU history is an important indicator of recurrence. Our findings mirror this, hence the use of historical data to augment the biomarker dataset.

In this paper we present three contributions. Firstly, we compare 3 machine learning algorithms for risk stratification using a large dataset of health records extracted from SCI-Diabetes. As part of this experimentation, we provide a novel comparison of recent and historical features to identify their impact on decision-making. Finally, we identify the contributory power of each feature in our dataset using mutual information values. Results indicate that risk labels are highly dependant on features for detection of peripheral neuropathy.

We structure this paper in the following manner. In Section 2 we formalise our methodology and evaluation by introducing the dataset and machine learning task. In Section 3 we describe the results of our evaluation, while in Section 4 we provide some discussion. Finally, in Section 5 we present our conclusions.

2 Methods

The Scottish Care Information - Diabetes (SCI-Diabetes) platform is operated by the National Health Service (NHS) Scotland and contains digital health records of patients with Diabetes. The SCI-Diabetes platform has been used since 2014 to allow sharing of records across multi-disciplinary care teams (i.e. endocrinologist, diabeteologist, etc) to facilitate treatment of Diabetes patients in Scotland. Records within SCI-Diabetes store information captured during routine healthcare appointments, including results of medical tests and procedures. We describe

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_11

these features as biomarkers. Biomarker values are recorded for a patient from the time they were first diagnosed with Diabetes until their most recent examination, and are used to predict the progression of associated health concerns. This includes our primary interest; risk of DFU formation in either foot.

We have been given access to a subset of electronic health records from the entire scope of SCI-Diabetes[1]. The SCI-Diabetes Biomarker dataset (henceforth simply the Biomarker Dataset) contains data describing biomarker features for 30,941 unique patients. The dataset contains a mix of recent and historical biomarker information, and the length of patient history varies based upon the number of appointments they have attended or tests they have taken. These biomarkers are of two types: numerical, where values are continuous numbers; and categorical, where values are discrete categories. Each patient is also allocated a risk classification, provided by clinicians based upon National Institute for Health and Care Excellence (NICE) guidelines [6].

To our knowledge, patient SCI-Diabetes records have not previously been used to stratify risk of DFU development using machine learning algorithms. However, researchers have applied machine learning to patient records within SCI-Diabetes to predict whether a patient has Type 1 or Type 2 Diabetes [7]. They found that a simple neural network model could outperform clinicians on this task. This suggests that SCI-Diabetes is a promising data source for us to explore risk stratification using biomarker data.

2.1 Recent and Historical Biomarker Datasets

We derive a machine learning task to classify the risk of a patient developing a DFU based on their biomarker data. As a first step, a data quality assessment was performed and anomalies were corrected. Categorical data was standardised across the dataset. Missing numerical values were infrequent, and only occurred in patients with multiple appointments, so were imputed using the mean of that biomarker value for that patient. Finally, missing class labels were removed - of the initial 30,941 patients, 4,621 have no recorded risk status and thus are dropped. The remaining 26,320 patients are divided into three risk classes: 19,419 Low risk, 4,286 Moderate risk and 2,615 High risk patients respectively.

We differentiate between using only recent biomarkers (basing risk prediction only upon a patient's most recent appointment) and using historical biomarkers (incorporating knowledge of a patient's medical history) to create two distinct datasets. This allows us to compare whether additional historical knowledge allows improved risk stratification for DFU formation. To create the Recent Biomarker dataset, we extracted the biomarker values recorded at a patients' most recent appointment. This resulted in a dataset with 26,320 instances (one for each patient), each of which had 37 features. To create the Historical Biomarker dataset, the Recent dataset was augmented with additional features calculated from historical biomarkers. Records of patient appointments were grouped using their CHI number, then aggregated using the following techniques:

[1] Access provided by NHS Data Safe Haven Dundee following ethical approval and completion of GDPR training.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_11

- **Numerical:** Calculating the mean of values from first examination to second most recent examination and subtract this from the most recent examination value. Use of an average allows us to capture a baseline for the patient and standardise variable lengths of patient histories. Using a difference measure allows us to explicitly capture feature changes between appointments.
- **Categorical:** Using the second most recent examination value as the new feature.

When historical features were unavailable (i.e. a patient only had a single appointment), we reused recent feature values. The resulting dataset had 26,320 instances and 47 features - the 37 features of the Recent dataset, and 10 additional features created through our aggregation of historical biomarkers (broken down into 4 numerical and 6 categorical features). Not all features were suitable for aggregation, as some were unalterable (such as Diabetes Type).

We highlight that we have removed a feature describing whether a patient has previously developed a foot ulcer. While we acknowledge the literature suggests that DFU history is an important feature for risk prediction [8], we removed this feature from our task because it is linked with the class label. In SCI-Diabetes, a patient is always allocated a High risk label if they have a history of DFU.

2.2 Machine Learning for Risk Stratification

We now have a multi-class classification problem where the goal is to identify whether a patient is at Low, Moderate or High risk of developing a DFU. We apply three different machine learning algorithms for this purpose:
Logistic Regression (LR) estimates the probability of an event or class using a logit function. For multi-class classification, the prediction based on a one-vs-all method (i.e. the probability of belonging to one class vs belonging to any other class, repeated for each class in the dataset).
Multi-Layer Perceptron (MLP) is a neural network formed of an input layer, multiple hidden layers and an output layer. The goal is to model the decision boundary of different classes by learning an easily separable representation of the data. Within each hidden layer, data is transformed by applying a set of weights and biases to the output of the previous layer (i.e. creating a new representation of the data). The final layer converts this representation to a probability distribution across the range of possible classes (i.e. modelling the decision boundary).
Random Forest (RF) classifier is an ensemble of multiple decision tree classifiers. Decision trees learn by inferring simple decision rules to classify data. Random forest learns an uncorrelated set of decision trees where the output of the forest is an improvement over any individual tree it contains.

These algorithms were selected as they compliment the feature composition of the dataset. RF is well suited for categorical data, as inferred rules are inherently categorical and previous work suggests decision trees perform well in this context [7]. However, conversion from numerical to categorical data reduces granularity. Therefore we also evaluated LR and MLP algorithms, which are well suited towards numerical data.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_11

Table 1. Comparison of ML Models for Risk Stratification of DFU.

Model	Recent		Historical	
	Accuracy (%)	F1 Score (%)	Accuracy (%)	F1 Score (%)
Random Forest	82.57	63.00	**82.98**	**64.34**
Multi-Layer Perceptron	79.28	53.39	78.91	54.65
Logistic Regression	81.12	58.68	79.80	48.54

2.3 Understanding Important Biomarkers for Risk Stratification

Finally, it is desirable to understand biomarkers which contribute to allocation of risk classes. Mutual Information (MI) is a method to calculate dependency between two variables X and Y (see Equation 1).

$$MI(X;Y) = H(X) - H(X|Y) \tag{1}$$

Where $H(X)$ is the entropy for X and $H(X|Y)$ is the conditional entropy for X given Y. Greater MI values indicate greater dependency between two variables, whereas low values indicate independence. We calculate MI between each feature and label in both the Recent and Historical Biomarker datasets.

3 Results

Overall, the results (shown in Table 1) are very promising for this challenging problem, with a peak accuracy of 82.98% and Macro F1 Score of 64.14% respectively. We believe the difference indicates some overfitting to the majority class, suggesting the ML algorithms are capable of recognising low risk patients, but struggle to accurately identify Moderate and High risk patients. RF is the best performing ML algorithm, obtaining the highest accuracy and Macro F1 Score on both datasets. The difference in Macro F1 Score highlights that RF is more capable of correctly predicting minority classes, Moderate and High risk.

Next we examine whether risk stratification of DFU is more accurate if we consider patient history. In the Recent Biomarker dataset, we achieve a peak accuracy of 82.57% and Macro F1 Score of 63% using RF. Using the Historical Biomarker dataset demonstrates a slight increase to 82.98% and 64.14% respectively (the best performing set up from our experiments). However, results on MLP and LR are mixed. For example, the MLP algorithm demonstrates a minor reduction in accuracy on the Historical Biomarker dataset compared with the Recent Biomarker dataset (dropping from 79.28% to 78.91%), but an improved Macro F1 Score (increasing from 53.39% to 54.65%). This suggests a drop in performance on the majority class (Low risk), but an increase in performance in one of the minority classes (Moderate or High risk). The LR algorithm demonstrates a noticeable drop in performance when historical features are included.

Finally, as a proxy for feature importance we calculate MI between each biomarker feature the risk label (see Table 2).

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_11

Table 2. MI Values for Features in Recent and Historical Biomarker Datasets.

Feature	MI	
	Recent	Historical
Albumin Concentration	0.79	1.04
Angina	0.7	0.63
Body Mass Index	0	0
CVA Haemorrhagic	0.18	0.13
CVA Non-Haemorrhagic	0	0
CVA Summary	0.24	0.79
CVA Unspecified	0.73	1.03
Coronary Artery Bypass Surgery	0.66	0.05
Current Tobacco Nicotine Consumption Status	0	0
Diabetes Mellitius Sub Type	0.9	0.47
Diabetes Mellitius Sub Type 2	0.3	0.31
Diabetes Mellitius Type	0.29	0.54
Diastolic Blood Pressure	0.57	0.69
Estimated Glomerular Filtration Rate	2.68	2.78
HbA1c	0.08	0.44
Hypertension	0.41	0.96
Ischaemic Heart Disease	0.41	0.23
Maculopathy Left	0.16	0.97
Maculopathy Right	0.02	0.6
Maculopathy Summary	0.18	0.59
Monofilament Left Sites	16.97	16.37
Monofilament Right Sites	16.87	17.41
Myocardinal Infraction	0.57	0.92
Peripheral Pulses Left	8.34	8.3
Peripheral Pulses Right	7.95	7.86
Peripheral Vascular Disease	0.92	0.82
Protective Sensation Left	17.01	17.2
Protective Sensation Right	16.68	17.12
Protective Sensation Summary	19.38	18.94
Retinopathy Left	1.04	0.94
Retinopathy Right	1.21	0.66
Retinopathy Summary	0.93	0.65
Systolic Blood Pressure	0.57	0.27
Total Cholesterol	0.81	0.85
Transient Ischemic Attack	0.52	0
Triglyceride Level	0.44	0
Weight	0.08	0
Historical Numerical Features		
Historical Estimated Glomerular Filtration Rate	-	2.39
Historical Albumin Concentration	-	1.53
Historical Systolic Blood Pressure	-	1.07
Historical Diastolic Blood Pressure	-	0.73
Historical Categorical Features		
Historical Protective Sensation Left	-	16.18
Historical Protective Sensation Right	-	16.06
Historical Peripheral Pulses Left	-	7.8
Historical Peripheral Pulses Right	-	7.57
Historical Monofilament Left Sites	-	15.81
Historical Monofilament Right Sites	-	16.11

4 Discussion

Our results suggest that inclusion of historical features make algorithms more capable of recognising Moderate and High risk patients. However, we observed evidence of all algorithms overfitting, which we suspected was due to class imbalance in the dataset. We applied downsampling to reduce the number of samples in the Low risk class to 5,000 (making class sizes comparable across the three classes). In almost all scenarios, downsampling resulted in decreased accuracy and F1 Score. The drop in F1 Score highlights that additional instances within the Low risk class is contributory to ML algorithms' ability to learn this task.

By applying MI, we find the most important features for model decision-making are derived from tests of foot health (such as protective sensation, pe-

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_11

ripheral pulses, and monofilament tests), many of which are directly related to tests for peripheral neuropathy. However, these features are mostly categorical in nature - for example, monofilament feature values are calculated based on clinical test thresholds. It would be useful to capture more granular detail about these features, for improved risk stratification. Interestingly, there are several features which are not traditionally associated with foot health which also contribute, specifically albumin concentration, retinopathy and estimated glomerular filtration rate. We suspect that unusual recordings of these features indicate further complications of Diabetes, which would be linked with increasing risk of DFU formation. A recent study suggesting links between retinopathy and peripheral neuropathy [9] supports this idea. Biomarkers such as smoking status or BMI (which can lead to complications in other aspects of Diabetes) are not so relevant to DFU formation. Finally, we note large dependency values using the historical knowledge we have captured from relevant categorical features, and low dependency values on historical numerical features. Even several features which previously showed low importance scores when only considered using knowledge from the most recent appointment (such as systolic and diastolic blood pressure) demonstrated a noticeable increase. This supports the outcome of our experiments; that an effective risk stratification system should be based on evolving knowledge of the patient and their history.

There are several limitations to our study. SCI-Diabetes captures data for patients within Scotland, so our results may not generalise to other healthcare systems, though we highlight that our findings overlap with [8] despite this. We have only tried a subset of possible machine learning algorithms, selected as they compliment the feature composition of the dataset. It would be desirable to compare more algorithms on this task. Finally, while we have performed some initial experimentation to address dataset imbalance and subsequent overfitting of trained algorithms, further strategies in the literature could be investigated (both at the training and evaluation stages of model development). Despite these limitations, we believe our results show good potential for risk stratification of DFU formation using ML algorithms.

5 Conclusions

In this paper, we presented a comparative study of machine learning algorithms for risk stratification of diabetic foot ulceration. Our results have indicated the empirical value of examining historical biomarker features of a patient to stratify this risk. Finally, we have highlighted the importance of biomarker indicators of peripheral neuropathy as contributing to risk categorisation of a patient.

In future work, we plan to incorporate historical and recent biomarkers into a single time-series record for a given patient. This should allow more granular prediction of the evolution of DFU formation risk. Another interesting aspect is addressing the dataset imbalance by examining cost by error class. Finally, we plan to improve capability to explain model decision-making by combining fea-

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_11

ture importance and counterfactual explainer algorithms to identify contributing factors to DFU formation and which biomarkers to target for treatment.

Acknowledgements This work was funded by SBRI Challenge: Delivering Safer and Better Care Every Time for Patients with Diabetes. The authors would like to thank NHS Data Safe Haven Dundee for providing access to SCI-Diabetes, and Walk With Path Ltd as a partner on this project.

References

1. Dhatariya, K., Levy, N., Kilvert, A., Watson, B., Cousins, D., Flanagan, D., Hilton, L., Jairam, C., Leyden, K., Lipp, A., et al.: Nhs diabetes guideline for the perioperative management of the adult patient with diabetes. Diabetic Medicine **29**(4), 420–433 (2012)
2. Dros, J., Wewerinke, A., Bindels, P.J., van Weert, H.C.: Accuracy of monofilament testing to diagnose peripheral neuropathy: a systematic review. The Annals of Family Medicine **7**(6), 555–558 (2009)
3. Guest, J.F., Fuller, G.W., Vowden, P.: Diabetic foot ulcer management in clinical practice in the uk: costs and outcomes. International Wound Journal **15**(1), 43–52 (2018). https://doi.org/https://doi.org/10.1111/iwj.12816, https://onlinelibrary.wiley.com/doi/abs/10.1111/iwj.12816
4. Kasbekar, P.U., Goel, P., Jadhav, S.P.: A decision tree analysis of diabetic foot amputation risk in indian patients. Frontiers in endocrinology **8**, 25 (2017)
5. Li, L., Liu, B., Lu, J., Jiang, L., Zhang, Y., Shen, Y., Wang, C., Jia, W.: Serum albumin is associated with peripheral nerve function in patients with type 2 diabetes. Endocrine **50**(2), 397–404 (2015)
6. National Institute for Health and Care Excellence: Diabetic foot problems: prevention and management (NICE Guideline NG19) (2015)
7. Sainsbury, C., Muir, R., Osmanska, J., Jones, G.: Machine learning (neural network)-driven algorithmic classification of type 1 or type 2 diabetes at the time of presentation significantly outperforms experienced clinician classification. In: Diabetic Medicine. vol. 35, pp. 168–168. Wiley (2018)
8. Schäfer, Z., Mathisen, A., Svendsen, K., Engberg, S., Rolighed Thomsen, T., Kirketerp-Møller, K.: Towards machine learning based decision support in diabetes care: A risk stratification study on diabetic foot ulcer and amputation. Frontiers in Medicine **7**, 957 (2020)
9. Sharma, V.K., Joshi, M.V., Vishnoi, A.A., et al.: Interrelation of retinopathy with peripheral neuropathy in diabetes mellitus. Journal of Clinical Ophthalmology and Research **4**(2), 83 (2016)
10. Snyder, R.J., Hanft, J.R.: Diabetic foot ulcers–effects on qol, costs, and mortality and the role of standard wound care and advanced-care therapies. Ostomy/wound management **55**(11), 28–38 (2009)
11. Tulloch, J., Zamani, R., Akrami, M.: Machine learning in the prevention, diagnosis and management of diabetic foot ulcers: a systematic review. IEEE Access **8**, 198977–199000 (2020)
12. Yazdanpanah, L., Nasiri, M., Adarvishi, S.: Literature review on the management of diabetic foot ulcer. World journal of diabetes **6**(1), 37 (2015)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_11