Machine-learning algorithm to non-invasively detect diabetes and pre-diabetes from electrocardiogram

Anoop R Kulkarni,1,2 Ashwini A Patel,2 Kanchan V Pipal,2 Sujeet G Jaiswal,2 Manisha T Jaisinghani,2 Vidya Thulkar,2 Lumbini Gajbhiye,2 Preeti Gondane,2 Archana B Patel,2 Manju Mamtani,2,3 Hemant Kulkarni

ABSTRACT

Objectives Early detection is of crucial importance for prevention of type 2 diabetes and pre-diabetes. Diagnosis of these conditions relies on the oral glucose tolerance test and haemoglobin A1c estimation which are invasive and challenging for large-scale screening. We aimed to combine the non-invasive nature of ECG with the power of machine learning to detect diabetes and pre-diabetes.

Methods Data for this study come from Diabetes in Sindhi Families in Nagpur study of ethnically endogenous Sindhi population from central India. Final dataset included clinical data from 1262 individuals and 10461 time-aligned heartbeat records recorded digitally. The dataset was split into a training set, a validation set and an independent test set (8892, 523 and 1046 beats, respectively). The ECG recordings were processed with median filtering, band-pass filtering and standard scaling. Minority oversampling was undertaken to balance the training dataset before initiation of training. Extreme gradient boosting (XGBoost) was used to train the classifier that used the signal-processed ECG as input and predicted the membership to ‘no diabetes’, pre-diabetes or type 2 diabetes classes (defined according to American Diabetes Association criteria).

Results Prevalence of type 2 diabetes and pre-diabetes was 30% and 14%, respectively. Training was smooth and quick (convergence achieved within 40 epochs). In the independent test set, the DiaBeats algorithm predicted the classes with 97.1% precision, 96.2% recall, 96.8% accuracy and 96.6% F1 score. The calibrated model had a low calibration error (0.06). The feature importance maps indicated that leads III, augmented Vector Left (aVL), V4, V5 and V6 were most contributory to the classification performance. The predictions matched the clinical expectations based on the biological mechanisms of cardiac involvement in diabetes.

Conclusions Machine-learning-based DiaBeats algorithm using ECG signal data accurately predicted diabetes-related classes. This algorithm can help in early detection of diabetes and pre-diabetes after robust validation in external datasets.

INTRODUCTION

The global epidemic of diabetes continues to burgeon. Estimates suggest that the...
world-wide prevalence of diabetes in adults (>18 years of age) was 9.3% (463 million people affected) in the year 2019.2 Two aspects related to these estimates are of critical importance. First, the current estimates of diabetes prevalence around the globe are likely to be an underestimate of the true diabetes burden.3 Second, diabetes is frequently preceded by pre-diabetes—a condition that is more tenable to preventive measures and effective in diabetes prevention.4 Thus, early detection of pre-diabetes and diabetes is of utmost importance in a concerted effort to reduce global diabetes burden.5 6 It is notable that the definitions for diabetes and pre-diabetes rely heavily on the measurement of blood glucose (ideally in the setting of an oral glucose tolerance test, OGTT) and haemoglobin A1c (HbA1c) concentrations which are both invasive and challenging to roll out as screening methods especially in low-resource settings.7 9

The cardiovascular system is an early target of the diabetic process. Even in the absence of a frank cardiac disease and appreciable dysglycaemia, there continue to be perturbation in the autonomic control of cardiac function in pre-diabetes and diabetes.10–12 Further, the diabetic process is marked with structural changes as seen in diabetic cardiomyopathy and cardiac fibrosis.13 For example, it has been posited14 that hyperglycaemia, reactive oxygen species, accumulation of advanced glycation end-products (AGE) and the abnormalities in neurohormonal regulation together contribute to a degenerative process characterised by cardiac fibrosis. These structural and functional changes in the cardiovascular system can be detected very early during pre-diabetes/diabetes progression even when dysglycaemic symptoms have not yet surfaced.15

Since cardiac autonomic neuropathy and cardiac fibrosis can both alter the electrical properties of heart and since electrocardiography (ECG) is non-invasive, we reasoned that ECG can be used for early detection of pre-diabetes and diabetes. Previous attempts in this direction include using electronic health records and natural language processing to predict diabetes,16 using deep learning (DL) techniques to investigate the predictive ability of heart rate signals in diabetes detection,17 ECG-based detection of impending hypoglycaemic events in critical care18 and using raw ECG to predict diabetes.19 More recently, the IRGNet architecture20 was developed to predict pre-diabetes (defined as presence of impaired fasting glucose or impaired glucose tolerance or both) which achieved an accuracy of 85% to predict pre-diabetes. However, a comprehensive effort to predict the spectrum of diabetic affliction (including latent diabetes defined on the basis of HbA1c concentrations) is currently lacking.

The aim of this paper is to demonstrate how machine learning (ML) techniques can be used to harness the screening potential of ECG to predict pre-diabetes and type 2 diabetes. Here, we present our results that demonstrate the screening potential of ECG when combined with ML to predict pre-diabetes and diabetes in a field study of ethnically endogenous, high diabetes risk population in India. Using data from this field study, we compared the classification performance of several ML and DL architectures with a goal to predict the diabetes status. We also added interpretive value to the classification to gain insights into the potential ECG features that can differentiate individuals based on diabetes status. Finally, we also investigated whether concomitant comorbidities and demographic conditions can influence the classification performance of our ML model.

METHODS
Study participants
Data for this study come from the Diabetes in Sindhi Families in Nagpur (DISFIN) study. The DISFIN study focused on the determination of the genetic basis of type 2 diabetes and other metabolic traits within the families of an ethnically endogenous population (Sindhis) with high risk of type 2 diabetes. Detailed inclusion and exclusion criteria for this observational study are described in its registration page on ClinicalTrials.gov (study id NCT03918525). Briefly, families with at least one known case of type 2 diabetes and living in the Jaripatka, Mecosabagh or Khamla areas of Nagpur (where there is maximum density of Sindhi population) were enrolled in the study. Detailed personal and family history, clinical assessment, diet history, anthropometry and laboratory investigations were carried out on all consenting participants.

Definition of type 2 diabetes and pre-diabetes
As a part of the full protocol DISFIN participants underwent ECG, fasting plasma glucose (FPG), fasting plasma insulin, HbA1c and random blood glucose (RBG) and fasting serum C-peptide measurements. Type 2 diabetes was defined in accordance with the definition of the American Diabetes Association criteria.21 Specifically, type 2 diabetes was considered present when one or more of the following criteria were met: self-reported diabetes on anti-diabetic medication, FPG ≥7.0 mmol/L, HbA1c ≥6.5% or RBG ≥11.1 mmol/L. Pre-diabetes was defined as presence of one or more of the following criteria in the absence of type 2 diabetes: FPG 5.6–6.99 mmol/L or HbA1c 5.7%–6.49%.

Electrocardiography
All the study participants were offered the standard, 12-lead ECG. All ECGs were recorded using the BPL Medical Technologies’ digital ECG device (Cardiart 6208) with following specifications: recording duration – 10s, ≥12bit analogue to digital converters, 0.05–150 Hz bandwidth and a sampling frequency of 1000 Hz. All recordings were saved as dicom files and used for analyses in a deidentified fashion.
**Data preprocessing**

Figure 1 shows the overall analytical pipeline and the preprocessing steps (orange bordered box in figure 1) used in this study. The unit of analyses for this study was a heartbeat. Raw ECG signals extracted from the dicom files were first subjected to median filtering (kernel size 11), then bandpass filtering (within the range of 5–40 Hz) and then standard scaled. These preprocessed signals were then passed through a QRS-detector algorithm that identified all the R peaks. To ensure that all beats extracted with have full set of available features, we trimmed the detected R peaks list to exclude the first and last (potentially incomplete) beats. These beats were then aligned such that the R peak always represented the centre of a 0.8 s wide window. These extracted beats were used for training and validation.

**Training the ML algorithm**

The aligned beats were then down sampled to a frequency of 125 Hz. Thus, the resulting beat-level waveforms had a total of 100 unique features per lead. These features were horizontally concatenated to generate a set of 1200 features (12 leads × 100 features) per heartbeat. This feature set was used for training and validation of the ML algorithm. As shown in figure 1, the full feature set of extracted beats was split into a training set (85%), a validation set (5%) and an independent test set (10%).

As a first step, we examined the comparative predictive performance of five ML techniques—component-wise boosting, multilayer perceptron, probabilistic neural network, random forest and extreme gradient boosting (XGBoost) algorithm. These ML algorithms have been commonly used in clinical classification tasks. This comparison was done in a 10-fold cross-validation scenario restricted to the training set only for a prefixed number of 10 iterations. From these techniques, the best performing techniques (based on minimum misclassification error, low Brier score and high classification confidence) was chosen for further training. For comparison, we used three metrics (misclassification error, Brier score and predictive confidence). Based on these analyses, we chose a technique that provided least misclassification error, lowest Brier score and highest confidence.

Training was then undertaken by extensive hyperparameter tuning (using grid search) with a goal to minimise the multiclass log loss and maximise the multiclass classification accuracy. Training was continued till the loss and accuracy metrics plateaued (with patience level of 5). Best model with the least loss and maximum accuracy was chosen as the final model.

**Validation of the best model**

The best ML model was evaluated during training against the validation set. The validation performance of the model was evaluated using four metrics: precision (comparable to positive predictive value in epidemiology), recall (comparable to sensitivity in epidemiology), accuracy and F1 score (summarised using the macro method). These analyses were based off a confusion matrix that cross-tabulated the actual
and predicted classes. The predictive performance of the best ML model was then evaluated in a similar fashion in the independent test set that was not used for any analyses prior to this step. To determine the generalisability of the best ML model, we generated a calibration curve and estimated the expected calibration error (ECE) using established methods.22 23 Finally, the results were interpreted in the light of the gain-derived feature importance (expressed as F score)24 to mark specific locations within a beat that provided maximum discriminatory information.

Comparison with DL models
We also compared the classification performance of the XGBoost model with that of other DL architectures based on long short-term memory (LSTM), 1-dimensional convolution neural networks (1D-CNN), the deepECG model and a time-series transformer-based classifier. The details of the rational and architecture of these models are provided in online supplemental note 1. Briefly, the DL models used for comparison were based on combinations of LSTM, 1D-CNN and transformer blocks and reflected a wide spectrum of parameter space (ranging from a minimum of 24 579 to a maximum of 433 523 parameters). All the DL models used the same dataset as that used by the DiaBeats algorithm and used the same optimiser (Adam), learning rate (0.0001), stopping criteria (loss patience of 5) and input and output layers (online supplemental note 1).

Analytical platform
All ML-related analyses were conducted using Python V.3.7 within an Anaconda environment.25 Following python libraries were used for specific analytical tasks: Scipy (for signal pre-processing),26 Numpy (for beat representation and vectorised analyses),27 WFDB (for R peak detection and beat alignment, https://pypi.org/project/pywfdb/), xgboost (for XGBoost training, https://pypi.org/project/xgboost/), Scikit learn (for creating confusion matrices, performance metrics and feature importance metrics)28 and, netcal (for creating confusion matrices, performance metrics) for inclusion in this study. We aimed to include ECG recordings that were minimally influenced by artefacts like person movement, electrical interference and poor recording in one or more leads. We, thus, restricted this dataset using a global criterion of the total voltage range (all leads, all times) between −2 and +2 mV. Using this criterion, we finally included clinical data and ECGs for 1262 individuals. A comparison of the included and excluded individuals based on study characteristics is shown in table 1). Overall, the excluded participants had a higher proportion of males, were leaner, more hypertensive and had a higher prevalence of known cardiac conditions as compared with the included participants.

The included participants were on an average ~48 years of age and were predominantly (61%) females (table 1). Since the primary DISFIN study was designed to enrol families with at least one known case of type 2 diabetes and since the study focused on endogenously related families, the prevalence of both type 2 diabetes and pre-diabetes was high in this study (~30% and 14%, respectively). Also, the prevalence of insulin resistance was 35%. Similarly, the coexistence of other metabolic conditions like hypertension (51%), obesity (~40% based on BMI≥30 Kg/m²) and dyslipidaemia (36%) was also high. From 10s ECG recordings of these participants, we extracted a total of 10 461 heart beats as shown in figure 1. The distribution of target classes in these 10 461 beats was as follows: no diabetes 5744 (54.91%); pre-diabetes 1398 (13.36%) and type 2 diabetes 3319 (31.73%) which was very similar to the distribution of the number of participants in these classes.

Choice of best performing ML technique
Using the training set only and training under a 10-fold cross-validation scenario, we compared the predictive performance of six candidate ML methods mentioned earlier. The results of these analyses are shown in online supplemental figure 1 and indicated that XGBoost was associated with least misclassification error (0.07), lowest Brier score (0.07) and highest confidence (0.82). We, therefore, used XGBoost as the ML classifier in all the following analyses.
Training and predictive performance of DiaBeats

Within the randomised training set of 8892 beats (85% of all beats, figure 1) the prevalence of no diabetes, pre-diabetes and type 2 diabetes classes was 54.87%, 13.56% and 31.57%, respectively. Thus, there existed a class imbalance with pre-diabetes as the least represented class. Therefore, we first used the SMOTE algorithm to generate a balanced dataset that yielded approximately equal instances of each class. The resulting balanced training dataset contained a total of 14,637 beats equally distributed with respect to the target classes. This balanced dataset retained all the instances from the original training set and added the oversampled nearest neighbours for the two minority classes to make up equal number of instances per target class. Hyperparameter grid search provided the following estimates which were used for further training: learning rate (eta) 0.1, maximum tree depth 16, subsample

Table 1 Characteristics of the study participants

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Included (N=1262)</th>
<th>Excluded (N=112)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>47.54 (13.15)</td>
<td>51.22 (14.85)</td>
</tr>
<tr>
<td>Females</td>
<td>780 (61.81%)</td>
<td>43 (38.39%)</td>
</tr>
<tr>
<td>Self-reported metabolic conditions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>309 (24.48%)</td>
<td>29 (25.89%)</td>
</tr>
<tr>
<td>High blood pressure</td>
<td>362 (28.68%)</td>
<td>43 (38.39%)</td>
</tr>
<tr>
<td>High cholesterol</td>
<td>64 (5.07%)</td>
<td>2 (1.79%)</td>
</tr>
<tr>
<td>Heart disease</td>
<td>68 (5.39%)</td>
<td>18 (16.07%)</td>
</tr>
<tr>
<td>Body mass index (BMI) (kg/m²)</td>
<td>29.16 (5.52)</td>
<td>26.77 (5.49)</td>
</tr>
<tr>
<td>Waist circumference (WC) (cm)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>95.58 (11.62)</td>
<td>90.22 (12.83)</td>
</tr>
<tr>
<td>Females</td>
<td>93.17 (12.59)</td>
<td>91.40 (14.91)</td>
</tr>
<tr>
<td>Comorbidities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>378 (29.95%)</td>
<td>37 (33.04%)</td>
</tr>
<tr>
<td>Pre-diabetes</td>
<td>170 (13.47%)</td>
<td>12 (10.71%)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>649 (51.43%)</td>
<td>69 (62.16%)</td>
</tr>
<tr>
<td>Obesity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indian cut-off (BMI≥25 kg/m²)</td>
<td>989 (78.37%)</td>
<td>70 (62.50%)</td>
</tr>
<tr>
<td>Universal cut-off (BMI≥30 kg/m²)</td>
<td>500 (39.62%)</td>
<td>31 (27.68%)</td>
</tr>
<tr>
<td>Central obesity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males (WC≥90 cm)</td>
<td>341 (70.75%)</td>
<td>36 (52.17%)</td>
</tr>
<tr>
<td>Females (WC≥85 cm)</td>
<td>579 (74.23%)</td>
<td>28 (65.12%)</td>
</tr>
<tr>
<td>All</td>
<td>780 (61.81%)</td>
<td>64 (57.14%)</td>
</tr>
<tr>
<td>Dyslipidaemia</td>
<td>450 (35.66%)</td>
<td>29 (25.89%)</td>
</tr>
<tr>
<td>Insulin resistance</td>
<td>447 (35.42%)</td>
<td>42 (37.50%)</td>
</tr>
<tr>
<td>Estimated variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOMA-IR</td>
<td>3.54 (2.40)</td>
<td>3.75 (3.13)</td>
</tr>
<tr>
<td>HOMA-β (%)</td>
<td>88.78 (22.62)</td>
<td>86.42 (25.50)</td>
</tr>
<tr>
<td>Total body fat (%)</td>
<td>37.42 (6.65)</td>
<td>33.59 (8.08)</td>
</tr>
<tr>
<td>Laboratory investigations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fasting plasma glucose (mmol/L)</td>
<td>5.70 (2.36)</td>
<td>5.92 (2.78)</td>
</tr>
<tr>
<td>Fasting plasma insulin (mIU/L)</td>
<td>13.81 (5.77)</td>
<td>13.76 (6.06)</td>
</tr>
<tr>
<td>HbA1c (%)</td>
<td>6.08 (1.32)</td>
<td>6.16 (1.34)</td>
</tr>
<tr>
<td>Serum C-peptide (ng/mL)</td>
<td>2.66 (1.18)</td>
<td>4.22 (16.89)</td>
</tr>
<tr>
<td>Total serum cholesterol (mg/dL)</td>
<td>161.97 (31.00)</td>
<td>155.30 (27.74)</td>
</tr>
<tr>
<td>Serum triglycerides (mg/dL)</td>
<td>123.20 (45.38)</td>
<td>112.60 (33.15)</td>
</tr>
<tr>
<td>Serum HDL (mg/dL)</td>
<td>54.87 (5.71)</td>
<td>55.26 (6.62)</td>
</tr>
<tr>
<td>Males</td>
<td>54.74 (5.54)</td>
<td>55.73 (5.86)</td>
</tr>
<tr>
<td>Females</td>
<td>82.33 (27.26)</td>
<td>77.02 (24.92)</td>
</tr>
<tr>
<td>Serum LDL (mg/dL)</td>
<td>24.81 (9.42)</td>
<td>22.52 (6.63)</td>
</tr>
</tbody>
</table>

HbA1c, haemoglobin A1c; HDL, high density lipoprotein; HOMA-IR, homeostatic model assessment - insulin resistance; HOMA-β, homeostatic model assessment - β cell function; LDL, low density lipoprotein; VLDL, very low density lipoprotein.
fraction 0.8, L2 regularisation (lambda) 2 and booster type gbtree.

Training of the XGBoost-based DiaBeats model was smooth and rapid (as indicated by the loss curves shown in figure 2A). There was a monotonic decrease in the multiclass log loss for both the training and the validation sets with the loss for the latter set staying consistently higher throughout the training process. Convergence was achieved within 40 epochs. Conversely, the predictive accuracy of the training set reached near perfection during the first 16 epochs but that for the validation set stabilised after 34 epochs at ~98%. The predictive performance of the DiaBeats model is shown in figure 2B,C in the validation as well as the independent test set. The confusion matrices (figure 2B) showed that both in the validation and the test sets the DiaBeats algorithm discriminated among the target classes effectively as indicated by aggregation of the numbers along the diagonal of the confusion matrix and very few beats incorrectly classified off-diagonally. Also, the proportion of correctly classified beats within each target class was consistently high (>93% in the validation set and >95% in the independent test set. These confusion matrices distilled into impressive performance metrics for the DiaBeats model as shown in figure 2C. The precision, recall, accuracy and F1 scores were very high for both the validation and the independent test sets. Indeed, when we dichotomised the dataset as diabetes or pre-diabetes versus ‘no diabetes’, the predictive accuracy of the DiaBeats algorithm in the independent test set was 99.58% (95% CI 99.36% to 99.81% (online supplemental figure 2).

Lastly, when we examined the classification confidence using calibration curve, we observed that the ECE of the original, uncalibrated prediction was 10.36% (figure 2D, cyan curve). After temperature scaling, the revised estimates of the classification confidence provided an ECE estimate of 6.32% (>4% improvement). Hence, we used these revised and scaled probability estimates as the basis of predictions from the DiaBeats model.

Feature importance
As a final interpretive step, we investigated the contribution of the specific ECG features to the overall model prediction. For this, we used the F-score for the gain metric and compared the feature importance across all the 1200 features used in the DiaBeats model. We found (figure 3A) that the highest importance of 21.72 was for feature #550 (located in lead augmented Vector Foot (aVF)) which rapidly decreased to <10 after the top 14th feature (feature ID #633 located in lead V2). After that there was a more gradual decrease in the feature importance estimates. Of the top 25 most important features, there were five each located in leads V3 and V4, followed by three each in leads V2 and aVF; two each in leads II, III, augmented Vector Right (aVR) and V5; and one in lead augmented Vector Left (aVL). These results indicated that major discriminatory information for the DiaBeats model came from the chest leads (especially V2 and V3), augmented limb leads and lead III. Indeed, the mean F score for all features located in each lead (figure 3B) was
highest for lead V2 (3.69), followed by lead V3 (3.68) but least for leads I, aVL and II (2.42, 2.66 and 2.70, respectively). To further visually inspect the location of feature importance, we created an average heartbeat for all study beats belonging to the ‘no diabetes’ class. We plotted this average beat lead-wise and color-coded the plots based on the estimated feature importance. We found (figure 3C) that the region of QRS complex in all chest leads and in lead III and the T waves in leads V2, V3 and V5 were especially important for the classification problem studied here.

**Comparison of DiaBeats algorithm with DL frameworks**

Lastly, we compared the classification performance of our algorithm with that of five DL models that have been previously used for time series classification. Details of these five models are provided in online supplemental note 1 and online supplemental figures 5–9. We found (online supplemental table 1) that the accuracy of the XGBoost-based DiaBeats algorithm was higher than that of each of the five DL models in all three datasets (training, validation and test). The computationally complex algorithm based on transformers achieved the lowest classification accuracy (~75% in the test set) but took a long time to converge (~7 hours on a GPU-enriched hardware). The results shown in online supplemental table 1 demonstrated that depth and model complexity did not improve the classification accuracy and that the XGBoost classifier outperformed all the candidate ML (online supplemental figure 1) and DL (online supplemental table 1) models studied.

**DISCUSSION**

In this study, we have demonstrated the potential use of an ML-driven approach using ECG to detect diabetes and pre-diabetes at the level of a single heartbeat. Our algorithm (DiaBeats) was highly accurate in both the validation set as well as the independent test set. The limitations of OGTT and HbA1c in population screening for diabetes are well known. In comparison, there are clear benefits in using ECG as a gatekeeper for diabetes/pre-diabetes. First, ECG is non-invasive. Second, ECG is relatively inexpensive. For example, based on a survey of laboratory practices in India (www.medifee.com) the average end-user price of OGTT and HbA1c estimation is currently ₹760 (₹321 and ₹439, respectively) while that of a standard ECG measurement is ₹214. Third, while the DiaBeats algorithm is specifically trained to detect diabetes/pre-diabetes, an ECG can yield a far richer information spectrum about coexisting cardiovascular conditions, and therefore, simultaneously act as a first-level screen for the overall cardiovascular health as well. Fourth, in this era of a global pandemic, telemedicine is likely to become a norm. In that context, ECG is more suited to remote assessment than invasive methods like OGTT and HbA1c estimation.

Of note, the use of ECG for detection of diabetes and pre-diabetes has a strong biological rationale. The need for evaluation of cardiac status in type 2 diabetes has been long recognised. Diabetic cardiomyopathy is characterised by stimulation of the connective tissue growth factors, fibrosis, accumulation of AGE and overall stiffness of the heart muscle. Therefore, ECG can be of value early during diabetes/pre-diabetes to detect the subtle, subclinical, concomitant and characteristic cardiac involvement. In our study, the best predicted (with highest confidence) illustrative examples from each target class showed (figure 4) that in the pre-diabetes and diabetes patients there were clearly detectable deep S waves in lead III; tall R waves in leads aVF and V2–V3; T wave abnormalities in leads V2, V3 and V5; and higher heart rate as compared with the best predicted person with ‘no diabetes’. These observations are in line with those described as biologically characteristic of diabetes without concomitant cardiac disease.

Our results need to be considered in the light of the published literature on similar detection strategies for diabetes/pre-diabetes. Automation attempts for ECG annotation with respect to cardiac abnormalities, especially arrhythmias, have been aplenty but use
of ECG as a predictor of diabetes in the context of ML or DL has been infrequent and very recent. Our study differs from the previous attempts in this direction in the following ways. First, previous studies have relied on the fact that heart rate variability (HRV), an indicator of autonomic regulation of the heart function, is affected in diabetes. While this is certainly true, the cardiac affliction in diabetes is also a result of the structural damage to myocardium. Our study used ECG in such way that information from both the structural and functional abnormalities could be included in training. For example, we did not explicitly extract HRV-related features (time and frequency domain) and used that as input to training as previous studies have done. Rather, we extracted beats and used the time-aligned features of these beats as inputs for training. This approach permitted the inclusion of both HRV-related and HRV-unrelated beat morphology to be used for model training. Second, all previous studies have focused on a sequence of heart beats rather than a single beat as a unit of analysis as used in this study. Consequently, the speed of diagnosis is greatly improved in our approach. Third, the accuracy of diabetes detection has varied in previous studies within the range of 86%–95%. Two recent studies in this regard are also important. The study by Porumb et al did use heartbeat as the unit of analysis but was designed to detect hypoglycaemic events in the context of continuous glucose monitoring while the second study used DL techniques to achieve an accuracy of 77% for detection of diabetes. Our estimate of 97% accuracy in the independent test set is a further improvement as compared with the previous studies. Last few years have seen an upsurge in the concept of using ML methods to detect pre-diabetes. For example, Wang et al developed the IGRNet model—an extensive DL model—to predict pre-diabetes from ECG images with high accuracy. Also, Bent et al have recently published a proof of concept study to demonstrate how the wearable smartwatches can be combined with ML methods to remotely monitor HbA1c and interstitial glucose concentrations. Similarly, Zhu et al have developed an impedance based DS21 tool to detect diabetes. Also, a large study (n=27779) by Kengne et al investigated the ability of 12 non-invasive risk score models to predict type 2 diabetes and found the accuracy to lie in the range 0.76–0.81. Using the rich waveform data recorded by digital ECG machines, our study was able to detect pre-diabetes with a very high accuracy both in the validation and independent test sets thereby providing a potential candidate for population-level screening. Together, our study provides an ECG-driven ML model that is fast and accurate in predicting both diabetes and pre-diabetes.

Notably, our XGBoost-based DiaBeats classifier was the best performing algorithm as compared other ML models or DL models. Algorithmically, XGBoost represents an ensemble approach based on decision-tree analytical framework. This algorithm has several advantages like hardware optimisation and speed, a parallelised tree-building process, tree pruning using a depth-first approach and, most importantly, L1 and L2 regularisation to reduce or avert overfitting. The value of this increasingly popular algorithm in clinical classification tasks is demonstrated in two recent, elegant reviews. An important observation from our analysis was the performance superiority of XGBoost over mainstream DL models (online supplemental table 1). Regardless of the depth and complexity of the DL models considered in this study XGBoost was faster and more accurate as compared with other DL models for time series classification task undertaken in this study. A similar observation has recently been reported for time series forecasting also. The classification superiority of XGBoost, however, should not be considered as a generalisation of its algorithmic capability in comparison to the DL models studied here. It is conceivable that the tabular representation of the preprocessed beat data, mitigation of the potential long-range time series features by clipping the ECG signals into individual heartbeats and the inherent nature of the classification problem may have made DL models somewhat redundant for this classification task. Future studies need to evaluate the formal algorithmic aspects of the ML/DL models that provide the best results for a given problem and dataset.

Our study has limitations. First, by way of design, the DISFIN study enrolled high-risk individuals and therefore the prevalence of metabolic conditions shown in table 1 cannot be considered reflective of a general population prevalence. This is important...
since, if the DiaBeats algorithm is to be used for screening purposes, it will likely be used in scenarios with much lower prevalence rates for diabetes and pre-diabetes than that in the DISFIN study. By (1) balancing the training dataset, (2) applying strict L2 regularisation, (3) calibration using temperature scaling and (4) then testing the predictive performance in an independent test we ensured that the DiaBeats model was as generalisable as possible. The consistency of results between the validation and test sets also points towards a strong internal validity. However, future studies need to test the predictive performance of the DiaBeats algorithm in other populations and across a range of other sampling frequencies and devices used to record ECG signals. Until such robust external validation on independent and unseen datasets is established, the results of the current study should only be considered as supportive evidence for the hypothesis. Second, 88.2% of the type 2 diabetes patients in this study were on antidiabetic medications. In the independent test set, the DiaBeats model had a comparable predictive accuracy of 96.51% (95% CI 94.27% to 98.75%) for beats from patients on medication and 89.47% (95% CI 82.57% to 96.37%) for beats from patients not on medications. These results do not appear to indicate potential confounding due to medication. However, larger studies with greater representation of patients not only on anti-diabetic but also on other important drugs (eg, antihypertensive, statins, aspirin) needs to be evaluated. From a screening point of view, however, this confounding is not very concerning since ECG analysis can be easily combined with medication history to improve screening performance. Third, this is an observational study with no data on conversion rates (from ‘no diabetes’ to pre-diabetes and from pre-diabetes to diabetes). Therefore, direct impact of early screening using ECG on these outcomes cannot be determined from our study. Fourth, in an observational study of this nature it is challenging to tease out a highly specific association of the ECG features with type 2 diabetes/pre-diabetes. In our study, there was a demonstrable gradient in the prevalence of coexisting metabolic conditions (online supplemental figure 3) with pre-diabetes and type 2 diabetes. Even though the very high value of precision of the DiaBeats algorithm are indicative of a specific association with type 2 diabetes and pre-diabetes, we conducted sensitivity analyses to examine if the presence of the comorbidities and age and gender influenced DiaBeats’ classification performance. We found (online supplemental figure 4) that age, gender and coexisting metabolic conditions did not significantly influence the classification performance as indicated by overlapping 95% CIs. Still, some concomitant extraction of features related to other coexisting metabolic conditions cannot be ruled out. Fifth, we did not have the resources to quantify cardiac fibrosis (using circulating biomarkers or contrast enhanced MRI) and therefore our interpretation of the potential of the DiaBeats algorithm to correlate with the degree of cardiac fibrosis could not be proved in this study. However, it needs to be remembered that the goal of the DiaBeats algorithm was not to diagnose cardiac fibrosis but to detect the cardiac changes (structural and functional) to aid in predicting the type 2 diabetes status. Sixth, we considered the possibility that preprocessing individual beats by taking a window of 0.8 s around the R peaks may result in trimming off of some usual features in individuals with lower heart rates. We thus also repeated the entire set of analyses using a beat window size of 1 s around the R peaks. The results (online supplemental figure 10) indicated that the classification performance of DiaBeats was better with a beat window size of 0.8 s. However, alternative approaches that account for heart rate in the choice of the window size can also be considered.

In conclusion, early detection is a cornerstone of diabetes prevention strategies, especially in high-risk populations. The invasive and expensive nature of OGTT and HbA1c assessments can be an impediment to large-scale screening for diabetes. In theory, our study provides a relatively inexpensive, non-invasive and accurate alternative which can be used as a gatekeeper to effectively detect diabetes and pre-diabetes early in its course. Nevertheless, adoption of this algorithm into routine practice will need robust validation on external, independent datasets.
REFERENCES
Digital health

41 Falcão-Pires I, Leite-Moreira AF. Diabetic cardiomyopathy: understanding the molecular and cellular basis to progress in diagnosis and treatment. Heart Fail Rev 2012;17:325–44.
AI + ECG heart trace can accurately predict diabetes and pre-diabetes

Could be used to screen for the disease in low resource settings, say researchers

An artificial intelligence (AI) algorithm, derived from the features of individual heartbeats recorded on an ECG (electrocardiogram), can accurately predict diabetes and pre-diabetes, suggests preliminary research published in the online journal *BMJ Innovations*.

If validated in larger studies, the approach could be used to screen for the disease in low resource settings, say the researchers.

An estimated 463 million adults around the world had diabetes in 2019. And picking up the disease in its early stages is key to preventing subsequent serious health problems. But diagnosis relies heavily on the measurement of blood glucose.

This is not only invasive but also challenging to roll out as a mass screening test in low resource settings, point out the researchers.

Structural and functional changes in the cardiovascular system occur early on even before indicative blood glucose changes, and these show up on an ECG heart trace.

The researchers therefore wanted to see if machine learning (AI) techniques could be used to harness the screening potential of ECG to predict pre-diabetes and type 2 diabetes in people at high risk of the disease.

They drew on participants in the Diabetes in Sindhi Families in Nagpur (DISFIN) study, which looked at the genetic basis of type 2 diabetes and other metabolic traits in Sindhi families at high risk of the disease in Nagpur, India.

Families with at least one known case of type 2 diabetes and living in Nagpur, which has a high density of Sindhi people, were enrolled in the study.

Participants provided details of their personal and family medical histories, their normal diet, and underwent a full range of blood tests and clinical assessments. Their average age was 48 and 61% of them were women.

Pre-diabetes and diabetes were identified from the diagnostic criteria specified by the American Diabetes Association.

The prevalence of both type 2 diabetes and pre-diabetes was high: around 30% and 14%, respectively. And the prevalence of insulin resistance was also high—35%—as was the prevalence of other influential coexisting conditions—high blood pressure (51%), obesity (around 40%), and disordered blood fats (36%).

A standard 12-lead ECG heart trace lasting 10 seconds was done for each of the 1262 participants included. And 100 unique structural and functional features for each lead were
combined for each of the 10,461 single heartbeats recorded to generate a predictive algorithm (DiaBeats).

Based on the shape and size of individual heartbeats, the DiaBeats algorithm quickly detected diabetes and prediabetes with an overall accuracy of 97% and a precision of 97%, irrespective of influential factors, such as age, gender, and coexisting metabolic disorders.

Important ECG features consistently matched the known biological triggers underpinning cardiac changes that are typical of diabetes and pre-diabetes.

The researchers acknowledge that the study participants were all at high risk of diabetes and other metabolic disorders, so unlikely to represent the general population. And DiaBeats was slightly less accurate in those taking prescription meds for diabetes, high blood pressure, high cholesterol, etc.

Nor were data available for those who became pre-diabetic or diabetic, making it impossible to determine the impact of early screening.

“In theory, our study provides a relatively inexpensive, non-invasive, and accurate alternative [to current diagnostic methods] which can be used as a gatekeeper to effectively detect diabetes and pre-diabetes early in its course,” they conclude.

“Nevertheless, adoption of this algorithm into routine practice will need robust validation on external, independent datasets,” they caution.
SUPPLEMENTARY MATERIALS FOR

DiaBeats: A MACHINE-LEARNING ALGORITHM TO NONINVASIVELY DETECT DIABETES AND PREDIABETES FROM ELECTROCARDIOGRAM

Anoop R. Kulkarni 1,2, Ashwini A. Patel 1, Kanchan V. Pipal 1, Sujeet G. Jaiswal 1, Manisha T. Jaisinghani 1, Vidya Thulkar 1, Lumbini Gajbhiye 1, Preeti Gondane 1, Archana Patel 1, Manju Mamtani 1,3, Hemant Kulkarni 1,3

1 Lata Medical Research Foundation, Nagpur, India
2 Innotomy Consulting, Bengaluru, India
3 M&H Research, LLC, San Antonio, Texas, USA

Corresponding author:

Dr. Hemant Kulkarni,
12023 Waterway Rdg
San Antonio, TX, USA 78249

Phone: (210) 602 5537
Fax: (210) 920 4434
Email: hemant.kulkarni@mnhresearch.com
Table of Contents

Supplementary Note 1: Choice of Deep Learning Architectures ............................................................ 3

Supplementary Figure 1. Comparison of predictive performance of candidate machine learning techniques ................................................................................................................................. 5

Supplementary Figure 2. Receiver operating characteristic curve to detect type 2 diabetes/prediabetes based on DiaBeats model .......................................................... 6

Supplementary Figure 4. Sensitivity analyses of the accuracy of the DiaBeats algorithm in the presence and absence of comorbidities and risk factors ............................................. 8

Supplementary Figure 5: Model architecture for the LSTM model (# parameters = 24,579) ........ 9

Supplementary Figure 6: Model architecture for the 1DCNN-Keras model (# parameters = 28,035) ............................................................................................................................... 10

Supplementary Figure 7: Model architecture for the DeepECG model (# parameters = 112,355) ................................................................................................................................. 11

Supplementary Figure 8: Model architecture for the Transformers model (# parameters = 222,899) ............................................................................................................................... 12

Supplementary Figure 9: Model architecture for the CNN16 model (# parameters = 433,523) .... 13

Supplementary Figure 10: Comparison of differing beat widow size on classification performance of DiaBeats .................................................................................................................. 14

Supplementary Table 1. Comparison of classification performance of deep learning architectures to predict diabetes and prediabetes ................................................................. 15
Supplementary Note 1: Choice of Deep Learning Architectures

**Rationale:** We aimed to compare the performance of the DiaBeats algorithm (XGBoost classifier) with deep learning architectures that are currently considered mainstream in the field of time series classification tasks. The chosen deep models represented a spectrum of deep learning methods (e.g. that included one-dimensional convolutional neural networks (1DCNN), long short-term memory (LSTM)), model complexity (number of layers in a model) and total number of model parameters.

**Selected deep learning models:** We selected the following five deep learning models for comparative performance evaluation:

- An LSTM-based model (hereinafter LSTM model),
- a 1DCNN time series classification model documented on the Keras platform ([https://keras.io/examples/timeseries/timeseries_classification_from_scratch/](https://keras.io/examples/timeseries/timeseries_classification_from_scratch/), hereinafter 1DCNN-Keras model),
- the DeepECG model (the Conv1D.py model from [https://github.com/ismorphism/DeepECG](https://github.com/ismorphism/DeepECG)),
- the time series transformers model for classification ([https://keras.io/examples/timeseries/timeseries_transformer_classification/](https://keras.io/examples/timeseries/timeseries_transformer_classification/), hereinafter Transformers model), and
- a deep, 16-layer 1D CNN-BatchNorm model (adapted from Kim and Pan\(^1\), hereinafter CNN16 model)

**Model architecture description:** All the models used an input shape of a 100 x 12 matrix that represented a 12 lead ECG for a single heartbeat downsampled to a sampling frequency of 125 Hz (from the original recording at a sampling frequency of 1000 Hz). Also, the last layer of all the models was a fully connected layer with three neurons and a softmax activation function.

The backbone of the LSTM model was a total of 64 LSTM cells followed by batch normalization and a fully connected layer with 64 relu-activated neurons (Supplementary Figure 5). The 1DCNN-Keras architecture was as described by the original author and comprised of 1DCNN layers and batch normalizations, followed by global average pooling and the final fully connected layer (Supplementary Figure 6). The DeepECG model was also a 1D CNN model with the same architecture as described by the original author (with the exception of the sizes of the kernels for the convolutional layers as appropriate for the input shape used here). The details of this model architecture are shown in Supplementary Figure 7. The Transformers model has gained popularity for time series classification tasks in the clinical settings (Song et al\(^2\), 2017; Wu et al\(^3\), 2021) and consists of an encoder-decoder core that uses the concept of multi-head attention. The transformers model we used had 4 heads with a head size of 256, a series of 4 transformer blocks and a total of 128 multilayer perceptron units (Supplementary Figure 8). Lastly, the CNN16 model comprised of serial combination of 1D CNN layers (gradually increasing in number of kernels with the depth of the network) and batch normalization layers.
culminating into the final fully connected output layer. The detailed architecture is shown in Supplementary Figure 9.

**Runtime model specifications:** All the models used a batch size of 32, sparse categorical crossentropy loss function and the Adam optimizer with a learning rate of 0.0001. Model performance was monitored using sparse categorical accuracy in the validation set. The training, validation and test sets used for training and evaluation were the same as used by the DiaBeats classifier. Best model performance for each model was used for comparative purpose (Supplementary Table 1).

**References:**

2. Song H, Rajan D, Thiagarajan JJ, Spanias A. Attend and Diagnose: Clinical time series analysis using attention models. arXiv 2017:1711.0.905v2 [stat.ML]
Supplementary Figure 1. Comparison of predictive performance of candidate machine learning techniques. These analyses were restricted to the unbalanced training set (number of beats = 8,892) and used a 10-fold cross-section scenario. Best method was chosen as the one that yielded least misclassification error, lowest Brier score and highest classification confidence. The machine learning techniques used were component-wise boosting (CWBoost), multilayer perceptron (MLP), probabilistic neural networks (PNN), random forest classifier (RF), support vector machines (SVM) and extreme gradient boosting (XGBoost). Each method was evaluated with three performance metrics: misclassification error (MCError), Brier score (Brier) and confidence (average softmax probability when predicting correctly). We aimed to select a method that yielded lowest values of MCError and Brier scores and highest value of confidence.
Supplementary Figure 2. Receiver operating characteristic curve to detect type 2 diabetes/prediabetes based on DiaBeats model. The target variable was a composite of type 2 diabetes and/or prediabetes. The predictor was a maximized softmax probability for the target class based on predictions given by the DiaBeats model. These analyses were restricted to the independent test set (1,046 beats).

![ROC Curve](image)

AUROC 0.9958 (95% CI 0.9936 - 0.9981)
Supplementary Figure 3. Coexistence of other metabolic conditions along with prediabetes and diabetes in the DISFIN study
Supplementary Figure 4. Sensitivity analyses of the accuracy of the DiaBeats algorithm in the presence and absence of comorbidities and risk factors. Diamonds and error bars represent the point and 95% confidence interval estimates for the predictive accuracy.
Supplementary Figure 5: Model architecture for the LSTM model (# parameters = 24,579)
Supplementary Figure 6: Model architecture for the 1DCNN-Keras model (# parameters = 28,035)
Supplementary Figure 7: Model architecture for the DeepECG model (# parameters = 112,355)
Supplementary Figure 8: Model architecture for the Transformers model
(# parameters = 222,899)
Supplementary Figure 9: Model architecture for the CNN16 model (# parameters = 433,523)
Supplementary Figure 10: Comparison of differing beat widow size on classification performance of DiaBeats

[Graph showing comparison of different beat window sizes for validation and test sets.]

BMJ Publishing Group Limited (BMJ) disclaims all liability and responsibility arising from any reliance placed on this supplemental material which has been supplied by the author(s).
Supplementary Table 1. Comparison of classification performance of deep learning architectures to predict diabetes and prediabetes.

<table>
<thead>
<tr>
<th>Model (Number of Parameters)</th>
<th>Metric</th>
<th>Dataset</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Training</td>
<td>Validation</td>
<td>Test</td>
<td></td>
</tr>
<tr>
<td>LSTM (24,579)</td>
<td>Loss</td>
<td>0.0855</td>
<td>0.2428</td>
<td>0.1663</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>0.9737</td>
<td>0.9465</td>
<td>0.9426</td>
<td></td>
</tr>
<tr>
<td>1DCNN-Keras (28,035)</td>
<td>Loss</td>
<td>0.0878</td>
<td>0.1941</td>
<td>0.1602</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>0.9701</td>
<td>0.9350</td>
<td>0.9503</td>
<td></td>
</tr>
<tr>
<td>DeepECG (112,365)</td>
<td>Loss</td>
<td>0.1134</td>
<td>0.1604</td>
<td>0.1438</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>0.9574</td>
<td>0.9426</td>
<td>0.9474</td>
<td></td>
</tr>
<tr>
<td>Transformers (222,899)</td>
<td>Loss</td>
<td>0.4932</td>
<td>0.6508</td>
<td>0.6044</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>0.7982</td>
<td>0.7467</td>
<td>0.7467</td>
<td></td>
</tr>
<tr>
<td>CNN16 (433,523)</td>
<td>Loss</td>
<td>0.0678</td>
<td>0.2044</td>
<td>0.1355</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>0.9768</td>
<td>0.9407</td>
<td>0.9465</td>
<td></td>
</tr>
</tbody>
</table>