ABSTRACT

Objectives The prevalence of diabetes has increased globally, leading to a significant disease burden and financial cost. Early prediction is crucial to control its prevalence.

Design A prospective cohort study.

Setting National representative study on Irish.

Participants 8504 individuals aged 50 years or older were included.

Primary and secondary outcome measures Surveys were conducted to collect over 40,000 variables related to social, financial, health, mental and family status. Feature selection was performed using logistic regression. Different machine/deep learning algorithms were trained, including distributed random forest, extremely randomised trees, a generalised linear model with regularisation, a gradient boosting machine and a deep neural network. These algorithms were integrated into a stacked ensemble to generate the best model. The model was tested using various metrics, such as the area under the curve (AUC), log loss, mean per classification error, mean square error (MSE) and root MSE (RMSE). The Shapley Additive exPlanations (SHAP) method was used to interpret the established model.

Results After 2 years, 105 baseline features were identified as major contributors to diabetes risk, including sex, low-density lipoprotein cholesterol and c-reactive protein. The SHAP algorithm provided insights into the decision-making process of the model.

Conclusions These findings could help physicians in the early identification of high-risk patients and implement targeted interventions to reduce diabetes incidence.

INTRODUCTION

Diabetes is a prevalent global health challenge that is expected to affect over 693 million individuals by 2045. Type 1 and type 2 diabetes, the two main types of diabetes, are clinically relevant, and they tend to affect individuals aged over 64 years in developed countries. Ageing has been shown to be closely associated with diabetes, with an increase in the number of senescent cells in various tissues being the most convincing evidence of this link. This disease imposes a considerable financial and health burden on national health and social care systems, with a cost of over US$1.31 trillion in 2015. Therefore, it is essential to establish an accurate clinical approach based on physical examination parameters to predict the risk of diabetes in the future.

Several research studies have attempted to predict the risk of diabetes. In one such study conducted in Tehran, the lipid accumulation product, a composite index of waist circumference and triglyceride level, was identified as a strong predictor of diabetes. However, no single risk factor can fully evaluate the risk of diabetes. Recently, machine learning technology has been used to assist in clinical decision-making. For example, an Artificial Neural Networks (ANN) and support vector machine have been applied to predict...
diabetes with a high degree of accuracy. A previous study compared the prediction performances of multiple machine learning models, including logistic regression, random forest, support vector machine and eXtreme Gradient Boosting, and reported that some of these models performed well. However, machine models are considered ‘black box’ models, which lack interpretability, and, therefore, there is a need for models that are transparent and enhance user trust, particularly in high-stakes fields such as medicine, where the consequences of flawed decision-making can be severe.

The Irish Longitudinal Study on Ageing (TILDA) is a comprehensive, longitudinal study that focuses on ageing in Ireland. This research is nationally representative and involves two waves of demographically representative data, with participants aged 50 years and above who were selected using a geographic cohort-based RANSAM sampling system. We used TILDA data to create an ensemble learning model that provides an interpretation for predicting the risk of both type 1 and type 2 diabetes based on the clinical parameters of older individuals. In the present study, many popular machine learning models were used to train our prediction model, including distributed random forest (DRF), extremely randomised trees (XRT), a regularised generalised linear model (GLM), gradient boosting machine (GBM) and deep neural network (DNN). By applying these powerful models, we could highly enhance the performance of prediction model.

METHODS

Participants

The TILDA collected data from Irish community-dwelling residents aged ≥50 years using a geographic cluster-based random sample matching system for data conduction. TILDA gathered population-representative data during wave 1 (October 2009–July 2011) and wave 2 (April 2012–January 2013). To ensure compliance with a previous study, only anonymised TILDA data were used, so no additional ethical clearance or informed consent was required. After the exclusion of participants who reported diabetes during baseline or were lost to follow-up in wave 2 (n=3118), the sample size for further analysis was 5386 participants.

Measures

Demographic information such as age, sex, smoking status and marital status was collected during the wave 1 survey of the participants in the study. Diabetes data were obtained through self-report in both waves. Participants were asked if a doctor had ever informed them that they had diabetes. A total of 44,622 clinical parameters were recorded during the wave 1 survey, which comprised 11 sections. These sections included variables on disability, functional impairment, medication, mental health and well-being, cognitive function, physical activity, income, socioeconomic status, social status, screening tests and health behaviours. Data were obtained from the participants through these sections. The raw data and derived variables from wave 1 survey are available at tilda.tcd.ie.

Feature selection

In the first wave of data, a substantial amount of data was censored, which posed a challenge for further feature selection and modelling. To avoid any potential bias and maintain sample size, a previous study recommended that censored data for each candidate feature be eliminated independently, enabling a more accurate assessment of the contribution of each feature to the outcome. Accordingly, logistic regression was conducted between each baseline feature and diabetes reported in the second wave, with censored data being independently removed in each model. Details and codes of all the features included were provided in the online supplemental material S1 Code Book. Variables with a p value of <0.05 and ORs of <0.6 or >1.5 were identified as the most important predictors for diabetes.

Modelling and performance testing

Following feature selection, feature matrices were extracted from wave 1 data, and features with censored data of over 500 participants were excluded, resulting in a final sample size of 5295. The participants were randomly split into training and test sets in an 8:2 ratio. Many popular machine learning models were used, including DRF, XRT, a regularised GLM, GBM and DNN. Thanks. ‘Stacking’ is a machine learning technique that combines multiple models, often referred to as base learners or weak learners, to produce more accurate and robust predictions than any single model alone. It operates on the principle that a group of diverse models working together can provide better results than a single model, as their individual strengths and weaknesses can complement each other. Therefore, a stacked ensemble approach was applied, integrating the models based on fivefold validation in the training cohort. To balance the class distribution, the minority classes were oversampled, and hyperparameters were optimised using grid search. The maximum number of ensemble models was 20, with the metric for early stopping being the area under the receiver operating characteristic curve. The pipeline’s execution time was about 1.5 hours. The performance of the model was assessed using multiple indicators, including the area under the curve (AUC), log loss, mean per classification error, mean square error (MSE), and root MSE (RMSE) in the test cohort. The ensemble methods, which combine predictions of multiple machine learning models, were employed to enhance the overall prediction accuracy by assembling a variety of powerful and diverse models.

Model interpretability

The SHapley Additive exPlanations (SHAP) is a game theory-based approach that can be used to interpret the output of any deep learning models. The SHAP method
connects optimal credit allocation with local explanations by employing classic Shapley values and their related extensions. In this study, the established model was interpreted using SHAP.

**Statistical analyses**

The data were gathered using R software (V.4.1.0) and analysed using the same software. Model training was carried out using the H2O package in R. Further information about the algorithms used for model training is available at h2o.ai.

**Patient and public involvement**

None.

**RESULTS**

Demographics of the participants at baseline

Categorisation of the participants was based on their report of diabetes in the wave 2 survey, which had a prevalence of 6.55% after 2 years. Demographic data of the Irish participants aged ≥250 years are provided in online supplemental materials. At baseline, the median age of participants without diabetes was 59 years, while those with diabetes was 64 years. A statistically significant difference was observed between the two groups, suggesting that the risk of diabetes increases with age. Additionally, sex was identified as a risk factor, whereas smoking and marital status did not impact the risk of diabetes after 2 years.

**Feature selection**

The TILDA study provided a comprehensive assessment of social and medical factors affecting diabetes risk. Independent logistic regression analysis of 105 baseline features identified significant contributors to diabetes risk after 2 years, as presented in figure 1. Sex was a significant factor, with an OR of 0.500 (0.399–0.625), consistent with the baseline demographic analyses. Protective factors included family finances, sandwich generation, private medical care, house benefits, receiving payments, mental health, state services, eye diseases, low-density lipoprotein cholesterol and child benefit. Hazardous factors included using dietician services, taking five or more medications, antihypertensive use, congestive heart failure and cirrhosis. These results highlight the complex nature of diabetes, influenced by various physical, psychosomatic and social factors.

The ensembled model performed well in terms of predicting diabetes risk after 2 years

We developed a set of machine learning models that predict whether a patient will develop diabetes after 2 years, using selected features. All features selected during the feature selection process were used in model training. We used the grid search algorithm to fit 20 different models, including XRT, DNN, DRF, GBM and GLM, and performed hyperparameter tuning for each model. This resulted in 22 models for assessment, including an ensemble model stacking the better-performing models and a final ensemble model with all 20 original models. We evaluated the performance of each model in an independent test set and selected the best model based on combined performance (figure 2). The stacked ensemble of all models was identified as the best model, with an AUC of 0.854, log loss of 0.187, mean per classification error of 0.267, RMSE of 0.229 and MSE of 0.052. We used a calibration plot to assess the agreement between predictions and observations, and the bootstrap method with 1000 simulations to calculate the robust AUC of the best model in the training and test sets. The best model showed favourable discrimination in both the training and test sets, with an AUC of 0.99 and 0.85, respectively (figure 3A,B). Calibration curve is graphical representations used to assess the reliability and performance of a predictive model, particularly in the context of probabilistic predictions. They display the relationship between the observed outcomes and the predicted probabilities, allowing for the evaluation of how well the predicted probabilities align with the actual outcomes. The calibration curve analysis also demonstrated high consistency between predicted outcomes and actual outcomes of diabetes after 2 years in both the training and independent test sets (figure 3C,D). In conclusion, our proposed best model is reliable and accurate in predicting the risk of diabetes after 2 years (figures 2 and 3A–D).

**Figure 1** Logistic regression identified top protective and hazardous factors for diabetes risk in older adults after 2 years.
in log loss decrease. Ultimately, the model was stopped early after 116 iterations, providing optimal performance. RMSE and deviance also exhibited a comparable trend of rapid decline during early iterations, followed by a gradual decline after 90 iterations (figure 3F,G). Additionally, AUC experienced a rapid increase in the initial training phase, and then gradually reached equilibrium, indicating the effectiveness of the learning process of the best model (figure 3H).

**Interpreting the decision-making process of the best model**

The capability of machine learning and deep learning models to fit data distribution in high-dimensional space accurately comes at the expense of being difficult to explain their decision-making process. Such models are known as ‘black box’ models, limiting their clinical application by not improving user trust. To enhance the model’s clinical potential, the SHAP algorithm, which is a game theory-based algorithm, was used to provide individual and global interpretations of the model’s predictions.

Our findings revealed that the 20 model components of the best model had a significant positive correlation among their predictions, as shown in figure 4A. Furthermore, variable importance was assessed across the model components, with medication (scaled importance 1.000), long-term illnesses (0.072) and sex (0.053) being significant predictors of the best model’s output, which was consistent with the results of the logistic regression analysis. Our results indicated that cardiac problems, such as heart attack and abnormal cardiac rhythm, significantly impacted the risk of diabetes. Moreover, health issues, including disability and self-felt health status, were strongly linked to diabetes, while lipid metabolism was also found to influence the onset of diabetes, as shown in figure 4B.

After exploring the best model, we aimed to elucidate the impact of the top features on the model’s output using the SHAP algorithm. Due to the unavailability of SHAP for interpreting the ensemble model, we used...
Figure 3  Performance and learning curves of the best model. (A) and (B) presented the AUCs of the best model in the training and test set. (C) and (D) established the calibration plots for the best model in the training and test set, respectively. (E)–(H) exhibited the changes of logloss, RMSE, deviance and AUC during the model training process, respectively. AUC, area under the curve; MSE, mean square error; RMSE, root mean square error.
the GBM model with the best performance for feature interpretation (figure 5A). Our findings reaffirmed that variables such as taking medications, sex and long-term illnesses significantly influenced the risk of diabetes after 2 years, alongside other factors such as dietician services, health status, ageing, high cholesterol and mental health. Furthermore, we estimated the contribution of the top features in the top models using the SHAP algorithm. From the dependence plots of SHAP (figure 5B–F), the number of medications, sex, long-term illnesses, availing dietician services and the number of regular medications were identified as the most important factors for DNN-1, DRF and GBM-1. In contrast, the best model weight-averaged every prediction to output the stacked prediction. An increase in the number of medications, long-term illnesses and the number of regular medications led to an increase in the risk of diabetes.

**DISCUSSION**

The present study successfully developed a robust and precise model that exhibits good generalisation and can serve as a feasible tool for predicting diabetes risk after 2 years in a clinical setting. The proposed model employed automatic hyperparameter optimisation and ensemble deep/machine learning techniques and was validated using a large sample size of Irish residents aged ≥50 years. The study performed a comprehensive analysis of over 40,000 parameters, including social, medical, financial and health-related factors, to investigate their impact on diabetes risk after 2 years. With the aid of this model, healthcare professionals can make accurate predictions of diabetes risk and implement tailored interventions to manage the identified risk factors.

The study’s significant finding is the creation of a stacked model, combining various state-of-the-art deep
learning and machine learning models that were optimised with hyperparameters. This new approach accurately predicts the risk of diabetes over 2 years. Previous research explored the risk factors and prediction of diabetes. For instance, Choi et al used an artificial neural network and support vector machine to predict diabetes with an AUC of 0.731. However, our model outperforms this research with an AUC of 0.85 in the independent test set. The superior performance of our model is due to the fact that Choi et al’s study used the cross-sectional Korea National Health and Nutrition Examination Survey (KNHANES), which contained less than 1000 parameters. The lack of a prospective nature and survey depth hindered its performance. Likewise, a previous study that employed the KNHANES dataset failed to enhance the models’ performance, with an AUC of <0.75, despite using several advanced machine learning models and a stacking method. The proposed best model, on the other hand, accurately predicted diabetes risk over 2 years, using various baseline parameters, including social, financial, health, mental and familial factors to improve its performance.

The interpretation of machine learning models is the most essential factor in facilitating their clinical application, as their decision-making processes are too complex to know and thus highly compromise the clinician and patients’ trust. Generally, high-stakes fields such as medicine are especially in need of interpretable models given the fatal consequences that can result from following a model prediction without understanding its decision process, which may be potentially flawed. With the assistance of SHAP, we took extensive and deep insights into our model’s decision-making process, and enabled clinicians to make the precision interventions for the identified risk factors of patients. Interestingly, we found that sex, age and alcohol intake increased diabetes risk after 2 years, which is highly agreed with previous findings. We also identified many novel factors that contributed importantly to model’s output, such as taking medicines, which could enlighten further studies on diabetes.

The accuracy of our diabetes risk prediction model was significantly improved by using hyperparameter optimisation, a variety of machine learning models and a stacking technique. While many statistical models have been developed to offer practical insights, machine and deep learning models are considered state-of-the-art. The XRT algorithm is a machine learning technique that creates an extremely randomised tree-growing process by integrating randomisation of a random subspace with a fully random selection of the cut point. In terms of

Figure 5  Model interpretation. (A) Summary SHAP plot for GBM-1. (B)-(F) exhibited how the top contributor affects models’ prediction. GBM, gradient boosting machine; SHAP, SHapley Additive exPlanations.
efficiency and accuracy, XRT is highly competitive when compared with the current state-of-the-art randomisation algorithm. XRT has been successfully applied in a variety of domains, including antibacterial peptide prediction. The optimisation of the solubility of oxaprazin, an anti-inflammatory drug, automated brain tumour detection and segmentation, and protein–protein interaction sites. DRF is another machine learning algorithm that generates a forest of regression or classification trees, each of which is a weak learner established on a subset of rows and columns. The final prediction of DRF is made by averaging the outputs from all the trees. GBM is a forward learning ensemble algorithm that employs increasingly refined approximations to make accurate predictions. In our study, a DNN, which is a multilayer feedforward artificial neural network, was trained using back propagation and stochastic gradient descent. All of these machine and deep learning models were used to develop our diabetes risk prediction model, resulting in high accuracy and robustness, as expected.

To enhance the predictive performance of our proposed model, it is important to address some of the limitations of this study. Validation of the model with a larger sample size and well-designed clinical trials is necessary, along with the exploration of additional deep learning methods, such as neural networks with self-attention. Unfortunately, the original questionnaires’ design prevented the prediction of diabetes subtype. Moreover, the present study used data from a single country, and investigation in different regions with varying lifestyle factors could verify the identified features remain consistent. Although the study focuses on patients in a high-risk age group, it would be beneficial to assess diabetes risks for patients at earlier ages, emphasising disease-prevention research. These limitations will be addressed in our upcoming studies as planned.

CONCLUSION
The study presented an ensemble model that accurately predicts diabetes risk after 2 years, using state-of-the-art machine learning techniques and identifying features from a national longitudinal study on ageing in Ireland. The elucidation of the decision-making process increased the model’s potential for clinical use, enabling clinicians to identify older individuals at high risk of developing diabetes. The findings of this study can facilitate further research on risk or protective factors for diabetes.

Author affiliations
1Department of Endocrinology, People’s Hospital of Wanning, Wanning, Hainan Province, China
2Department of Industrial Design, Hubei University of Technology, Wuhan, Hubei Province, China
3Department of Urology, The Second Affiliated Hospital of Hainan Medical University, Haikou, Hainan, China
4Department of Endocrinology, The First Affiliated Hospital of Hainan Medical University, Haikou, Hainan, China
5Department of Urology, The First Affiliated Hospital of Hainan Medical University, Haikou, Hainan, China

Contributors XX: conceptualisation, XM and XX: methodology, formal analysis and writing—original draft preparation. XM and JY: data curation. ZC, JY and HZ: writing—review and editing. XM: visualisation. ZC: project administration. ZC and JY: funding acquisition. HZ and XX: responsible for the overall content as the guarantors.

Funding This work was supported by the Natural Science Foundation of Hainan province (grant number: 822QN472) and Graduate Innovation research project of Hainan province (grant number: Qhyb2021-57).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants. According to the certificates from Hainan Medical University Ethics Committee, we only included anonymised data from TILDIA; therefore, no further ethical approval and written informed consent were required. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information. All data generated were included in the present study.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iD Zhifei Che http://orcid.org/0000-0003-2459-0853

REFERENCES
9 Bozorgmanesh M, Haraldeh F, Azizi F. Diabetes prediction, lipid accumulation product, and adiposity measures; 6-year follow-up; Tehran lipid and glucose study. Lipids Health Dis 2010;9:45.
13 Yao Q, Jin W, Li Y. Associations between fear of falling and activity restriction and late life depression in the elderly population: findings from the Irish longitudinal study on ageing (TILDA). J Psychosom Res 2021;146.
Supplementary Table 1

Table. Demographics of Irish participants aged 50 and above on baseline.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>No diabetes</th>
<th>Diabetes</th>
<th>p</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n = 5033)</td>
<td>(n = 353)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years)</td>
<td>59.00 (53.00-67.00)</td>
<td>64.00 (57.00-71.00)</td>
<td>1.08E-15</td>
<td>loge W = 13.40</td>
</tr>
<tr>
<td></td>
<td>67.00</td>
<td>71.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>2355 (0.47)</td>
<td>225 (0.64)</td>
<td>7.19E-10</td>
<td>χ² = 37.97</td>
</tr>
<tr>
<td>Female</td>
<td>2678 (0.53)</td>
<td>128 (0.36)</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Smoke</td>
<td>Yes</td>
<td>2740 (0.54)</td>
<td>221 (0.60)</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>2293 (0.46)</td>
<td>142 (0.40)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>3725 (0.74)</td>
<td>257 (0.73)</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>1308 (0.26)</td>
<td>96 (0.27)</td>
<td></td>
</tr>
</tbody>
</table>

*All skewed continuous variables were described by “median (interquartile range)”*

*and all categorical variables were described by “frequency [constituent ratio (%)]”.*