*informatics*

MDPI

*Article*

# A review on trending Machine Learning techniques for type 2 diabetes.

**Firstname Lastname [1,†,‡], Firstname Lastname [2,‡] and Firstname Lastname [2,*]**

[1] Affiliation 1; e-mail@e-mail.com
[2] Affiliation 2; e-mail@e-mail.com
* Correspondence: e-mail@e-mail.com; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials) +xx-xxxx-xxx-xxxx (F.L.)
† Current address: Affiliation 3.
‡ These authors contributed equally to this work.

**Abstract:** A single paragraph of about 200 words maximum. For research articles, abstracts should give a pertinent overview of the work. We strongly encourage authors to use the following style of structured abstracts, but without headings: (1) Background: place the question addressed in a broad context and highlight the purpose of the study; (2) Methods: describe briefly the main methods or treatments applied; (3) Results: summarize the article's main findings; (4) Conclusions: indicate the main conclusions or interpretations. The abstract should be an objective representation of the article, it must not contain results which are not presented and substantiated in the main text and should not exaggerate the main conclusions.

**Keywords:** keyword 1; keyword 2; keyword 3 (List three to ten pertinent keywords specific to the article; yet reasonably common within the subject discipline.)

## 0. How to Use this Template

The template details the sections that can be used in a manuscript. Note that the order and names of article sections may differ from the requirements of the journal (e.g., the positioning of the Materials and Methods section). Please check the instructions on the authors' page of the journal to verify the correct order and names. For any questions, please contact the editorial office of the journal or support@mdpi.com. For LaTeX-related questions please contact latex@mdpi.com.

## 1. Introduction

The introduction should [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16] briefly place the study in a broad context and highlight why it is important. It should define the purpose of the work and its significance. The current state of the research field should be reviewed carefully and key publications cited. Please highlight controversial and diverging hypotheses when necessary. Finally, briefly mention the main aim of the work and highlight the principal conclusions. As far as possible, please keep the introduction comprehensible to scientists outside your particular field of research. Citing a journal paper. Now citing a book reference or other reference types . Please use the command for the following MDPI journals, which use author–date citation: Administrative Sciences, Arts, Econometrics, Economies, Genealogy, Humanities, IJFS, Journal of Intelligence, Journalism and Media, JRFM, Languages, Laws, Religions, Risks, Social Sciences, Literature.

## 2. Diabetes

Maybe some details about diabetes

## 3. Machine Learning Background

Maybe some details about Machine Learning Theory.

**Citation:** Lastname, F.; Lastname, F.; Lastname, F. Title. *Journal Not Specified* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

## 4. Relevant Sections

### *4.1. Related Work*

Here we will review the two referenced review paper.

### *4.2. Machine Learning applications in diabetes*

As mentioned before, the applications of Statistical Analysis and Machine Learning in healthcare and more specifically in diabetes condition have demonstrated a steady rise in the last two decades, since the development of corresponding programming frameworks have enabled the easy storage, collection, processing, analysis of the massively available data quantity and employment of statistical and Machine Learning models [17–19]. Regarding diabetes research field, the literature deals with the identification of diabetic people, early or long term (2-10 years) prognosis and diabetes complications prediction or identification. Considering the prevention of diabetes, the ultimate goal is the extraction of features (e.g markers) which are relevant to diabetes occurrence. Then, in case that these features are configurable, the patient could have available some suggestions to apply in his lifestyle or diet in order to minimize the risk of developing diabetes.

Our literature review is focused on relatively new research articles or systematic reviews which are related with the context of our article e.g prediction of diabetes mellitus or prediabetes utilizing demographic, anthropometric, biometric, laboratory, nutritional, medical history, etc. data as input features. The first mathematical approaches over diabetes issue consisted of statistical risk scores exploiting questionnaires filled by waves from the participants. Some of the famous ones risk scores are Leicester Risk Assessment Score[20] developed by Leicester University and FINDRISC [21] developed by University of Helsinki. The former utilizing a Logistic Regression model, take into account age, ethnicity, sex, first degree family history of diabetes, antihypertensive therapy or history of hypertension, waist circumference and BMI to predict current impaired glucose regulation or diabetes mellitus, achieving an AUC metric of 72% and the latter -also exploited Logistic Regression- uses gender, age, BMI, use of blood pressure medication, history of high blood glucose, physical activity, daily consumption of vegetables, fruits or berries and family history of diabetes to predict a 10-year development achieving an AUC metric of 86%. We can observe at a first glance two variances of diabetes studies. The Leicster Risk aims to identify the current health condition, while FINDRISC tries to predict a long term prevalence. There are also numerous researches that deal with deep learning and more specifically with image recognition for the classification of diabetic retinopathy, which is a typical complication and very well studied in the research field, using images from eye bulb as input [2,12]. Another diabetes complications studies utilizing Machine Learning and Deep Learning include neuropathy and nephropathy [2,12]. Apart from classification problems there are also regression methods which are exploited for the prediction of Fasting Plasma Glycose or HbA1c levels, i.e. biomarkers that are the best indicators of abnormal glycose regulation and consequently diabetes mellitus presence [2,3,12].

Delving more into literature that is more relevant with the purpose of this study we can observe an adequate quantity of high quality articles which will help to understand a principal methodology in order to identify or predict diabetes development. Next, the chosen papers will be clustered based on their purpose, their key methodologies will be in a more detailed context described and also each other compared for advantages and disadvantages.

The current-state detection of diabetes, in the sense that the class variable and the independent features values are registered the same time is studied in [3,4,6,8,10,11,13–16]. In [4] the dataset used is PIMA from UCI repository [22], containing 768 records of healthy (500) and diabetic (268) Arizonan women over 21 years old with target variable the diabetes presence. First, during the feature selection procedure, methods like information gain, gain ratio, gini index,ANOVA, $\chi^2$ test, an extension of Relief, correlation, fast correlation and filter subset evaluation where employed. Glucose levels, BMI, diabetes pedigree function and age was identified as the best features on average from the aforementioned

techniques. Then, a variety of models was trained and tested on the different feature subsets derived from the feature selection techniques using 10 fold cross validation. The models probed were GAMBoost, regularized logistic regression, penalized multinomial regression, Bayesian generalized linear model, penalized logistic regression, generalized linear model, sparse distance weighted discrimination, generalized boosted regression model and Naive Bayes. The results showed that there is not a particular model that yields the highest metrics (Accuracy, Kappa Statistics, AUC, Sensitivity, Specificity, Log loss) simultaneously. Generalized additive model using LOESS yeld the best score in Friedman test, achieving AUC 85.36% and Sensitivity, Specificity 86%, 60% respectively. They concluded that the aforementioned feature subset and Machine Learning model could assist physicians and researchers to predict T2D, however this model should be assessed in bigger datasets for detecting new potentially crucial features and compared with other high performance models. In [6], researchers trained a GBM , a Random Forest and a Logistic Regression over a dataset containing 13,309 records from healthy and patients. The input features were Age, Gender, FPG, BMI, Triglycerides, Systolic pressure and LDL. First, the dataset was split in 80% training set and 20% testing set. Then, a misclassification cost matrix was constructed with a false negative-false positive ratio equal to 3/1 and zero cost for correct predictions. This cost matrix was used along with AUC as objective functions in order to tune the hyperapameters of the models using 10-fold cross validation. Due to the class imbalance the cut-off point of decision boundary was adjusted such that the misclassification cost is minimized. After this adjustment, each model with the tuned hyperapameters was trained on the entire training set and finally evaluated in the testing set. The best model was GBM with tuned hyperapameters number of trees $= 257$, interaction depth $= 2$, min samples per leaf $= 75$, learning rate $=$ 0.126 and threshold=0.24 achieving AUC 84.7%, misclassification rate 18.9%, Sensitivity 71.6% and Specificity 83.7%. Similarly, the Logistic Regression achieved values of 84%, 19.6%, 73,4% and 82.3%, respectively. For the Random Forest classifier the AUC value was 83.4%. They summarized suggesting the incorporation of such models to online programms for further assistance of physicians during patient assessments.

As mentioned before, apart from classification problems, Machine Learning can be applied to diabetes through Regression for estimation of predictive biomarkers such as FPG (Fasting Plasma Glycose) and revelation of factors that relate with the FPG variability. To this end, [3] utilize models of three conceptually different families such as boosting, bagging and linear regression, because each family has a different capability to detect hidden patterns and important features. The dataset initially consisted of 27,050 adults EHRs (Electronic Health Records) with no prior diabetes diagnosis between 2014 and 2017. A first propose of the study is to compare the models performance against FINDRISC, thus records that have missing values in any of the features that included also in FINDRISC, were dropped. Assuming normal distribution of the features, outlier detection took part using the formula $\overline{X} \pm (3 \times \text{SD})$ and each outlier value was marked as missing. Then, records and features tha had 50% or greater percent of missing values where dropped. Finally, the remaining missing values were imputed with MICE method using Bayesian linear regression for numerical values, logistic regression for binary values and polytomous regression for variables with more than two possible values. The preprocessing stage yields a final dataset of 3,723 records, 58 features and the FPG target variable. These features can be grouped in the following four groups: lipid profile lab results (HDL, LDL, total cholesterol and triglycerides), social determinants of health (consumption of alcohol, smoking, dietary habits, stress), cardiovascular variables (blood pressure measurements, atrial fibrillation history) and history of other health conditions (stroke, hypertension, colon cancer). As the final step before piping the data into the Machine Learning models, is the partition of the data into 6-months intervals (T6, T12, T18, T24, T30) according to the submission date of each record, thus 5 subdatasets where created and each Machine Learning model was trained in each subdataset and validated using 100 times random sampling with replacement (bootstrap). In each run the remaining unselected samples

were used to test the model. Linear Regression performed with the lowest Root Mean Squared Error 0.838 (95% CI 0.814-0.862) trained on only 7 features which are common to the FINDRISC. Next, RF achieved a value of 0.842 (95% CI 0.818-0.866), LightGBM 0.846 (95% CI 0.821-0.871), Glmnet 0.859 (95% CI 0.834-0.884) and XGBoost 0.881 (95% 0.856-0.907). When the whole dataset where available for training and testing, RF performed the lowest RMSE at 0.745 (95% CI 0.733-0.757) followed by Glmnet at 0.747(95% CI 0.734-0.759), while XGBoost performed the worst value at 0.760 (95% CI 0.748-0.772). The regression capability of each model was measured through $R^2$ coefficient, to measure how well the regressor fits the actual FPG value given the input features. For only 6 month data available linear model performed the best with an average value of 0.310, while RF performed the best for 18 and 30 months data available achieving a mean value of 0.340 and 0.368 respectively. In contrary, XGBoost performed the worst in all three time points (6-18-30 months of available data) achieving a mean value of 0.241, 0.293 and 0.343 respectively, despite its general superiority. Apart from the regression capability, the authors tested the classification capability using the cut-off FPG value of 6.1 mmol/L. The best AUC value demonstrated by Glmnet at 0.836 on average through the five time points, while the worst performance model was again XGBoost yielding a value of 0.8142. In terms of AUPRC, linear regression performed best with an average value of 0.6948 through the five time points, while was the worst model with a value of 0.6576. Finally, the feature importance was assessed for every model through the five time point datasets using different metrics (because each model has different structure) like $\beta$-coefficient, permutation importance on MSE or on Accuracy and variance gain. Triglycerides levels was assessed as the most important feature on LightGBM, while the remaining three models have Hyperglycemia history. For the next lower-importance features, even if there are some differences in the ranking, Age, HDL cholisterol, LDL cholisterol, Total cholisterol, Systolic pressure, Diastolic pressure and weight are the in the top 10. They concluded that, the more data available, the better stability do models have, even if from this research none new FPG related feature revealed, apart from those already clinical derived. LightGBM performed the most stable results through the multiple evaluations.

The long term diabetes prediction in the sense that the class variable is filled many years after the features are filled, utilizing the baseline-followup method, is studied in [16,21]. In [21] the study examined a cohort of 7949 people with known and unknown family history of diabetes. The features involved were questionnaires about lifestyle, socioeconomic and psychosocial matters, along with measurements of plasma glucose and insulin in an oral glucose tolerance test (OGTT), glycosylated haemoglobin (HbA1c), blood pressure, weight, height and hip circumference. In the baseline study, T2D was diagnosed in 51 women and 66 men, and prediabetes in 219 women and 259 men. A 1$^{st}$ follow-up study was carried out 8–10 years later, and a 2$^{nd}$ follow-up about 20 years later, with at least 70% participation. At the 1$^{st}$ follow-up, they found 102 women and 171 men with T2D, and 399 women and 522 men with prediabetes, and at the 2$^{nd}$ follow-up 230 women and 326 men with T2D and 615 women and 522 men with prediabetes. Those with diagnosed T2D at the baseline and the 1$^{st}$ followup were not called to follow-up later. The dataset was partitioned in 3 sets a training set, a validation set and a test set. The classifier utilized was Random Forest to predict the individual diabetes type 2 development after 10 years of the measurement. SHAP TreeExplainer was used to build an interpretable Machine Learning model, in order to find factors that relate with high or low diabetes risk. Hyperparameter tuning using 5 fold cross validation into the validation set took part in order to find the best hyperparameters set for the Random Forest, having as objective function a combination of AUC and robustness. This function is defined as:

$$S_l = AUC_l^{val} \cdot (1 - \Theta_l) \tag{1}$$

Where

$$\Theta_l = \mu(\sigma(X'_{ijkl})) \tag{2}$$

and

$$X'_{ijkl} = \frac{X_{ijkl} - \mu(X_{ijkl})}{\sigma(X_{ijkl})} \tag{3}$$

. $X_{ijkl}$ is a tensor of SHAP values per person i, feature j, cross validation split k and parameter set l. Then $X'_{ijkl}$ is the standardized tensor with zero mean and unit variance. Finally, $\Theta_l$ is the mean variance of o SHAP values per hyperapameters set l. The best hyperparameters set was: number of estimators $= 120$, min samples leaf $= 125$, max depth $=$ 4 and number of models $= 30$, achieving a robustness value of $S = 0.630$ and a value of AUC at 0.779. According to the SHAP values analysis, the features that increase the risk are: family history of diabetes, high waist-hip ratio, high BMI, increased Systolic pressure, increased Diastolic pressure, low physical activity and male gender. On the other hand the features that decrease the risk are: exercise, higher socioeconomic strata and lower age. Also with the help of a SHAP force plot, personalized risk profiles where extracted in order to assess the individual risk score, which is called *output value* and revealed the features that have the largest impact in the individual risk score. Finally, they suggest this method to be probed in the primary health care in order to improve diabetes care. In [5] a dataset of 500,000 records from Hanaro medical foundation containing diagnostic results and questionnairies, was utilized to conduct a multi classification experiment for prediction of prediabetic, diabetic and normal people in the following 1 year. People included in the dataset are those who had at least two years of continuous annual medical check-ups during the follow-up period, however those who had been diagnosed with diabetes, hyperlipidimia, hypertension or took medication for those diseases were excluded. During the preprocessing step, exclusion of records with at least one null value, employment of majority undersampling and Synthetic Minority Oversampling techniques, ANOVA and chi-square test, as well as Recursive Feature Elimination using an impurity metric to rank feature importance, were conducted. The selected features were fasting plasma glucose (FPG), body mass index (BMI), Gamma glutamyl transpeptidase (gamma-GTP), triglycerides, sex, age, uric acid, hemoglobin A1c (HbA1c), smoking, drinking, physical activity, and family history. Smoking status was divided into "currently smoking regularly", "never smoked" and "had quit smoking". Physical activity indicates the number of days the subject has engaged in physical exercise such as running, hillwalking, climbing stairs, jump roping for a minimum of 20 min. Family history with diabetes considers only parents and siblings diagnosed with T2D and drinking indicates the number of days the subject consumed alcoholic drinks. Then, RF, XGBoost and SVM were trained and tested on the dataset. First, two 10-fold cross validation grid searches for hyperaparemeter tunning were employed. The first in a more general grid, but the second focused on the best - more specific neighborhoods of good parameters values derived from the first. In addition, a Logistic Regression, a Stacking classifier, a Soft Voting classifier and a confusion matrix based classifier were employed. For the testing phase, 10-fold cross validation repeated 10 times and the average value for each metric such as accuracy, precision, recall, F1-score, MCC and KC was calculated. The results shown none significant difference between the models performance. All metrics were in the range 0.71-0.75. Then RF, XGBoost and SVM, were trained using data from two years, three years and four years before, which where added each one to the previous set. The accuracy performance showed significant improvement for every model. As a final test, the 12-features set was compared to a set of very classic diabetes predictors such as FPG, HbA1c, BMI, age, and sex into the different four datasets, which where previously introduced . The 10-repetion evaluation of accuracy through the different timepoint datasets showed significant superiority of the 12-features subset over the 5-features ( e.g. using data from the last four years the accuracy values were 0.81 for the former feature set and 0.77 for the latter set). The authors concluded that clinicians must pay attention to the changes in gamma-GTP, uric acid, and triglycerides over the years, as well as these models can work as decision support systems for practioners and diabetes educators. In [7], researchers evaluated a variety of single and ensemble models on ELSA dataset to predict type 2 diabetes occurrence. The dataset contains a variety of biometric,

anthropometric, hematological, lifestyle, sociodemographic and performance index variables. A number of different feature selection techniques were employed such as LASSO, correlation and Greedy stepwise. The selected method was Greedy stepwise with Naive Bayes and after the addition of some extra features the final dataset consisted of 34 input features and the class variable indicating the diabetes condition of the person. Random undersampling was conducted in order to have diabetic distribution per age group similar to real life. This yeld a final dataset of 2,331 records. For the evaluation of the models the procedure consists of creating 10 datasets from the existing using stratified train/test split with proportion 70/30 respecitively. Logistic Regression, Naive Bayes, Decision Tree, Random Forest, Artificial Neural Network, Deep Neural Network and three ensembles of Random Forest and Logistic Regression, namely Stacking, Voting and Weighted Voting, were employed. Due to class imbalance the method of adjusting threshold were conducted for each model having as objective function the $J$, Younden Index, which is the sum of specificity and sensitivity. For the Weighted Soft Voting, a biobjective optimization problem was solved in order to calculate the best weights for RF and LR, which maximize the sensitivity and AUC. Indeed the Weighted classifier produced the best results in terms of AUC with a value of 0.884. In addition, the Sensitivity, Specificity +PV, -PV, +LR, -LR, were 0.856, 0.798, 0.449, 0.967 and 4.245 respectively. Finally, the concluded that due to the superiority of ensemble models can be embedded into recommendation systems to prevent patients from development of diabetes. In [8] a physical, 138,000-records dataset containing 14 features such as age, pulse rate, breathe, left systolic pressure, right systolic pressure, left diastolic pressure, right diastolic pressure, height, weight, physique index, fasting glucose, waistline, LDL and HDL was used. Five subdatasets were created with random sampling in order to train the models five times and then calculate the average performance into a independent testing set using 5 fold cross validation. The models which were trained are Random Forest, Decision Tree and a Neural Network. The models were evaluated in different subsets of features using feature selection techniques like PCA, mRMR, without fasting glycose and only fasting glycose. Using all features every model achieved the best accuracy, while excluding fasting glycose trained the worst models. In the first case the accuracy of RF, Decision Tree and Neural network was 0.8084, 0.7853 and 0.7841 respectively. Using only fasting glycose as input feature trains the models still better than PCA and mRMR techniques, yielding accuracy at 0.7597, 0.761 and 0.7572 for each model respectively. They concluded that fasting blood glycose is a very good predictor of diabetes, however adding more features gives better performing models. As feature work they propose the extraction of indicators importance and the classification of the speicific diabetes type.

Materials and Methods should be described with sufficient details to allow others to replicate and build on published results. Please note that publication of your manuscript implicates that you must make all materials, data, computer code, and protocols associated with the publication available to readers. Please disclose at the submission stage any restrictions on the availability of materials or information. New methods and protocols should be described in detail while well-established methods can be briefly described and appropriately cited.

Research manuscripts reporting large datasets that are deposited in a publicly available database should specify where the data have been deposited and provide the relevant accession numbers. If the accession numbers have not yet been obtained at the time of submission, please state that they will be provided during review. They must be provided prior to publication.

Interventionary studies involving animals or humans, and other studies require ethical approval must list the authority that provided approval and the corresponding ethical approval code.

This is an example of a quote.

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.

### *4.3. Subsection*

#### 4.3.1. Subsubsection

Bulleted lists look like this:

- First bullet;
- Second bullet;
- Third bullet.

Numbered lists can be added as follows:

1. First item;
2. Second item;
3. Third item.

The text continues here.

### *4.4. Figures, Tables and Schemes*

All figures and tables should be cited in the main text as Figure 1, Table 1, Table 2, etc.


**Figure 1.** This is a figure. Schemes follow the same formatting. If there are multiple panels, they should be listed as: (**a**) Description of what is contained in the first panel. (**b**) Description of what is contained in the second panel. Figures should be placed in the main text near to the first time they are cited. A caption on a single line should be centered.

**Table 1.** This is a table caption. Tables should be placed in the main text near to the first time they are cited.

| Title 1 | Title 2 | Title 3 |
|---|---|---|
| Entry 1 | Data | Data |
| Entry 2 | Data | Data |

**Table 2.** This is a wide table.

| Title 1 | Title 2 | Title 3 | Title 4 |
|---|---|---|---|
| Entry 1 | Data | Data | Data |
| Entry 2 | Data | Data | Data [1] |

[1] This is a table footnote.

Text.
Text.

### 4.5. Formatting of Mathematical Components

This is the example 1 of equation:

$$a = 1, \tag{4}$$

the text following an equation need not be a new paragraph. Please punctuate equations as regular text.

This is the example 2 of equation:

$$a = b + c + d + e + f + g + h + i + j + k + l + m + n + o + p + q + r + s + t + u + v + w + x + y + z \tag{5}$$


**Figure 2.** This is a wide figure.

Please punctuate equations as regular text. Theorem-type environments (including propositions, lemmas, corollaries etc.) can be formatted as follows:

**Theorem 1.** *Example text of a theorem.*

The text continues here. Proofs must be formatted as follows:

**Proof of Theorem 1.** Text of the proof. Note that the phrase "of Theorem 1" is optional if it is clear which theorem is being referred to. □

The text continues here.

## 5. Discussion

Authors should discuss the results and how they can be interpreted from the perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

## 6. Conclusions

This section is not mandatory, but can be added to the manuscript if the discussion is unusually long or complex.

## 7. Future Directions

This section is not mandatory, but may be added if there are patents resulting from the work reported in this manuscript.

**Author Contributions:** For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.", please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

**Funding:** Please add: "This research received no external funding" or "This research was funded by NAME OF FUNDER grant number XXX." and and "The APC was funded by XXX". Check carefully that the details given are accurate and use the standard spelling of funding agency names at https://search.crossref.org/funding, any errors may affect your future funding.

**Institutional Review Board Statement:** In this section, you should add the Institutional Review Board Statement and approval number, if relevant to your study. You might choose to exclude this statement if the study did not require ethical approval. Please note that the Editorial Office might ask you for further information. Please add "The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving humans. OR "The animal study protocol was approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving animals. OR "Ethical review and approval were waived for this study due to REASON (please provide a detailed justification)." OR "Not applicable" for studies not involving humans or animals.

**Informed Consent Statement:** Any research article describing a study involving humans should contain this statement. Please add "Informed consent was obtained from all subjects involved in the study." OR "Patient consent was waived due to REASON (please provide a detailed justification)." OR "Not applicable" for studies not involving humans. You might also choose to exclude this statement if the study did not involve humans.

Written informed consent for publication must be obtained from participating patients who can be identified (including by the patients themselves). Please state "Written informed consent has been obtained from the patient(s) to publish this paper" if applicable.

**Data Availability Statement:** In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Please refer to suggested Data Availability Statements in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics. If the study did not report any data, you might add "Not applicable" here.

**Acknowledgments:** In this section you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

**Conflicts of Interest:** Declare conflicts of interest or state "The authors declare no conflict of interest." Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funders in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results must be declared in this section. If there is no role, please state "The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results".

**Sample Availability:** Samples of the compounds ... are available from the authors.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| MDPI | Multidisciplinary Digital Publishing Institute |
| DOAJ | Directory of open access journals |
| TLA | Three letter acronym |
| LD | Linear dichroism |

## Appendix A

### *Appendix A.1*

The appendix is an optional section that can contain details and data supplemental to the main text—for example, explanations of experimental details that would disrupt the flow of the main text but nonetheless remain crucial to understanding and reproducing the research shown; figures of replicates for experiments of which representative data are shown in the main text can be added here if brief, or as Supplementary Data. Mathematical proofs of results not central to the paper can be added as an appendix.

**Table A1.** This is a table caption.

| Title 1 | Title 2 | Title 3 |
|---|---|---|
| Entry 1 | Data | Data |
| Entry 2 | Data | Data |

## Appendix B

All appendix sections must be cited in the main text. In the appendices, Figures, Tables, etc. should be labeled, starting with "A"—e.g., Figure A1, Figure A2, etc.

## References


1. Lama, L.; Wilhelmsson, O.; Norlander, E.; Gustafsson, L.; Lager, A.; Tynelius, P.; Wärvik, L.; Östenson, C.G. Machine learning for prediction of diabetes risk in middle-aged Swedish people. *Heliyon* **2021**, *7*, e07419. https://doi.org/https://doi.org/10.1016/j.heliyon.2021.e07419.
2. Kavakiotis, I.; Tsave, O.; Salifoglou, A.; Maglaveras, N.; Vlahavas, I.; Chouvarda, I. Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal* **2017**, *15*, 104–116. https://doi.org/https://doi.org/10.1016/j.csbj.2016.12.005.
3. Kopitar, L.; Kocbek, P.; Cilar, L.; Sheikh, A.; Stiglic, G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific reports* **2020**, *10*, 1–12.
4. Howlader, K.; Satu, M.; Awal, M.; Islam, M.; Shariful Islam, S.M.; Quinn, J.; Moni, M.A. Machine learning models for classification and identification of significant attributes to detect type 2 diabetes. *Health Information Science and Systems* **2022**, *10*. https://doi.org/10.1007/s13755-021-00168-2.
5. Deberneh, H.M.; Kim, I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. *International Journal of Environmental Research and Public Health* **2021**, *18*. https://doi.org/10.3390/ijerph18063317.
6. Lai, H.; Huang, H.; Keshavjee, K.; Guergachi, A.; Gao, X. Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocrine Disorders* **2019**, *19*. https://doi.org/10.1186/s12902-019-0436-6.
7. Fazakis, N.; Kocsis, O.; Dritsas, E.; Alexiou, S.; Fakotakis, N.; Moustakas, K. Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction. *IEEE Access* **2021**, *9*, 103737–103757. https://doi.org/10.1109/ACCESS.2021.3098691.
8. Zou, Q.; Qu, K.; Luo, Y.; Yin, D.; Ju, Y.; Tang, H. Predicting Diabetes Mellitus With Machine Learning Techniques. *Frontiers in Genetics* **2018**, *9*. https://doi.org/10.3389/fgene.2018.00515.
9. De Silva, K.; Lim, S.; Mousa, A.; Teede, H.; Forbes, A.; Demmer, R.T.; Jönsson, D.; Enticott, J. Nutritional markers of undiagnosed type 2 diabetes in adults: Findings of a machine learning analysis with external validation and benchmarking. *PLOS ONE* **2021**, *16*, 1–21. https://doi.org/10.1371/journal.pone.0250832.
10. Dinh, A.; Miertschin, S.; Young, A.; Mohanty, S. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making* **2019**, *19*. https://doi.org/10.1186/s12911-019-0918-5.
11. Zhang, L.; Wang, Y.; Niu, M.; Wang, C.; Wang, Z. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study. *Scientific Reports* **2020**, *10*. https://doi.org/10.1038/s41598-020-61123-x.

12. Fregoso-Aparicio, L.; Noguez, J.; Montesinos, L.; García-García, J. Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetology & Metabolic Syndrome* **2021**, *13*. https://doi.org/10.1186/s13098-021-00767-9.
13. Xiong, X.l.; Zhang, R.; Bi, Y.; Zhou, W.h.; Yu, Y.; Zhu, D.l. Machine Learning Models in Type 2 Diabetes Risk Prediction: Results from a Cross-sectional Retrospective Study in Chinese Adults. *Current Medical Science* **2019**, *39*, 582–588. https://doi.org/10.1007/s11596-019-2077-4.
14. Rufo, D.D.; Debelee, T.G.; Ibenthal, A.; Negera, W.G. Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM). *Diagnostics* **2021**, *11*. https://doi.org/10.3390/diagnostics11091714.
15. Morgan-Benita, J.A.; Galván-Tejada, C.E.; Cruz, M.; Galván-Tejada, J.I.; Gamboa-Rosales, H.; Arceo-Olague, J.G.; Luna-García, H.; Celaya-Padilla, J.M. Hard Voting Ensemble Approach for the Detection of Type 2 Diabetes in Mexican Population with Non-Glucose Related Features. *Healthcare* **2022**, *10*. https://doi.org/10.3390/healthcare10081362.
16. Dritsas, E.; Trigka, M. Data-Driven Machine-Learning Methods for Diabetes Risk Prediction. *Sensors* **2022**, *22*. https://doi.org/10.3390/s22145304.
17. Frank, E.; Hall, M.A.; Holmes, G.; Kirkby, R.; Pfahringer, B.; Witten, I.H., Weka: A machine learning workbench for data mining. In *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*; Maimon, O.; Rokach, L., Eds.; Springer: Berlin, 2005; pp. 1305–1314.
18. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
19. Seabold, S.; Perktold, J. statsmodels: Econometric and statistical modeling with python. In Proceedings of the 9th Python in Science Conference, 2010.
20. Gray, L.J.; Taub, N.A.; Khunti, K.; Gardiner, E.; Hiles, S.; Webb, D.R.; Srinivasan, B.T.; Davies, M.J. The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. *Diabetic Medicine* **2010**, *27*, 887–895, [https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1464-5491.2010.03037.x]. https://doi.org/https://doi.org/10.1111/j.1464-5491.2010.03037.x.
21. Lindstrom, J.; Tuomilehto, J. The Diabetes Risk Score: A practical tool to predict type 2 diabetes risk. *Diabetes Care* **2003**, *26*, 725–731, [https://diabetesjournals.org/care/article-pdf/26/3/725/665299/dc0303000725.pdf]. https://doi.org/10.2337/diacare.26.3.725.
22. Dua, D.; Graff, C. UCI Machine Learning Repository, 2017.