*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:** Diabetes is a chronic disease which is characterized by elevated blood glycose levels, causing damages to human organs and tissues, worsen quality of life, and even having lethal implications. Last years, medicine have been in pace with data science, in the sense that its digital retransformation led to electronic health records (EHRs) collection of indivudials, patients and healthy ones. Thus, plenty of indivudial research articles as well as systematic reviews have been conducted in order to produce innovative findings and summarize the current development, respecitively. In this study, a detailed review is conducted in the context of T2D mellitus and Machine Learning, examining relatively new publications using tabular data, demonstrating the relevant use cases, the workflows during model building and the candidate predictors. Our findings showed that Gradient Boosting and Tree based models are the most successfull ones, SHAPley and Wrapper algorithms are quite popular feature interpretation and evaluation methods highlighting urinary results and dietary as asceding diabetes predictors apart from the classical invasive ones. These results could be offer easier management of the diabetes condition and open new research areas.

**Keywords:** Diabetes Mellitus; Machine Learning; Electronic Health Records; Non invasive decision support;

## 1. Introduction

Diabetes is an increasing chronic disease that affects people health and quality of life. The IDF Diabetes Atlas (2021) reports that 10.5% of the adult population (20-79 years) has diabetes, with almost half unaware that they are living with the condition, while by 2045 1 in 8 adults, approximately 783 million, will be living with diabetes, an increase of 46% [1]. Over 90% of people with diabetes have type 2 diabetes, which is driven by socio-economic, demographic, environmental, and genetic factors. On the other hand, Machine Learning has been a powerfull tool for prediction of events, among them medical related, as well as for the analysis of massive data and extraction of usefull knowledge. To this end, many studies have been conducted for the prediction of diabetes as well as for the identification of factors significantly contributing positive or negative to its management [2,3]. These studies address diabetes case in a broad context, ranging from Diabetic Complications, Genetic Background and Environment to Drugs and Therapies, and Health Care and Management regarding diabetes, and reaching deep learning with image analysis regarding Machine Learning, offering not so detailed insights about the methodologies. In this paper we review a variety of recent research papers towards diabetes diagnosis, diabetes long term prediction and relevant biomarkers regression, using tabular data coming from questionnairies, biometric, laboratory, physical examinations and dietary habits. Our aim is to analyze each paper for its methodology towards data preprocessing, feature selection, model building, evaluation and feature importance extraction, with an

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

Received:
Accepted:
Published:

ultimate goal to unveil patterns, best models and new non invasive predictors, opening new research areas.

## 2. Diabetes

### *2.1. Definition*

Diabetes, according to IDF, is a chronic disease caused either by the disability of pancreas to produce enought insulin, or the body's denial to utilize the produced insulin [1]. Insulin is a hormone created on pancreas, whose job is the transportation of blood glycose into the cells, which in turn offers the necessary energy amount to muscles. This insulin-related disorder, causes increased levels of glycose in blood and consequently damage to a number of human organs and tissues. There are three main types of diabetes, enumerated below:

- **Diabetes Type 1**: This type occurs in 10% of diabetic cases and most of them are children or adolescents. The immunity system attacks to insulin cells, causing disability into pancreas for insulin production. Some usual symptoms indicating this disease, are extensive thirst and urination, sudden weight loss and tiredness. Heredity and viral infections increase the risk of development.
- **Diabetes Type 2**: This is the most common type of diabetes, represanting about 90% of diabetic cases. In this case, the cells, do not recognize the insulin, thus they do not let glycose enter them. Consequently, the blood glycose increases continualy, causing Hyperglycemia, as well as the production of insulin, causing pancreas exhaustion. Long term, pancreas could be unable to maintain the adequate insulin quantity, causing even more increased Hyperglycemia. Many symptoms of type 1 are common, adding slow healing wounds and tingling or numbness in hands and feet. The risk factors are also common to type 1, and also overweight, age, high blood pressure, physical inactivity unhealthy nutrition and ethnicity.
- **Gestational diabetes**: As the name implicize, it affects pregnant women due to elevated levels of blood glycose. The prevalence increases with the age and particularly after 45 years. This disease has many critical complications to both mother and child, such as high blood pressure, overweight babies, obstructed labour and increased probability of developing type 2 diabetes.

### *2.2. Management*

As their common name indicates, the three types of diabetes have some common ways to manage them. First of all, type 1 demands continuous monitoring with the use of appropriate device and multiple insulin injections every day. A healthy well balanced nutrition with a lot of vegetal lipids instead of animal products such as butter and oil, as well as fruits, white meat and whole grain can contribute to a better experience. Furthermore, daily physical exercise such as 30 minutes of mild intensity and avoidance of unhealthy habits such as smoking and drinking are recommended. In type 2 case, it is possible to make use of oral medication such as Metformin and Sulfonylureas or, finally, insulin injection. Finally, regarding gestastional diabetes, women should monitor, as type 1, the levels of blood glycose.

## 3. Machine Learning Background

### *3.1. Models*

Machine Learning as its name denotes, is a domain that studies the creation of intelligent systems, which can learn from humans and especially from current knowledge. In Machine Learning, the tool that utilizes knowledge is called model. A model can be trained on current data, which represent the current knowledge, and therafter make some predictions based on the prior knowledge for new, unknown data. The model is represented by a mathematical function with tunable parameters, which tries to adapt onto the data, minimizing the error function, which evaluates how well the model has done. In Machine

Learning, there are three main subdomains, namely classification, clustering and regression, who, depending on the problem, have different capabilities to make predictions.

In this paper only classification and regression will be presented since the domain of Machine Learning is huge. In classification the model takes input data which contain some observable data (features) and their corresponding target variabe. The concept of classification is to train a model on the observed data and then try to predict the target variable of new unknown observable data. The target variable is called class and it is discrete (0 or 1, Yes or No, etc.). Some classification models, which are utilized in the proposed studies, are referenced briefly below:

- **Logistic Regression**: As its name denotes, this model uses a logistic function to estimate the probability for a vector of observable data $X_i$, to belong to a class $y$.
- **Naive Bayes**: NB assuming independence between features, applies Bayes Rule to predict the class that maximizes the respective likelihood.
- **K Nearest Neighbour**: KNN is an algorithm, which given an observation tries to find the K most similar, known data, summs their classes and assigns the majority class to the unknown obsveration. The similarity can be calculated with various methods, among them most common are Euclidean distance, Manhattan and Mahalanobis.
- **Support Vector Machine**: This model mapps the data onto the space and then tries to find the best hyperplane that separates the data of different classes. The term 'best' is evaluated by the margin, the greater value, the better it is.
- **Decision Tree**: DT is a data-driven model which does not make any assumption about data distribution, but constructs a tree-shaped struct of simple if-else rules based on input features, and based on this rules tries to make predictions.
- **Random Forest**: RF is an ensemble method consisting of multiple DT, as applies also in nature. The RF sums all DTs predictions and by majority voting, elects the class.
- **Multilayer Perceptron**: MLP maybe is the simplest form of Artificial Neural Network. It is a type of feed-forward ANN, simulating the way that human brain takes decisions. It consists of nodes and artificial neurons, which all cooperate to export a simple numerical value.
- **Gradient Boosting**: Gradient Boosting refers to a broad family of ensemble models which combines trees with the mathematical term 'gradient'. In every iteration, it combines the predictions of trees, also called 'weak learners' to produce a better result than previous iteration. After the last iteration it gives the final prediction. Such models are Extreme Gradient Boost (XGBOOST), Gradient Boosting Machine (GBM), Light GBM and CategoricalBoosting (CatBoost).
- **Ensemble Models**: In this category we set Voting classifier and Stacking classifier. Both they utilize the concept of combination between weak classifiers. In Voting classifier the weak classifiers prediction are combined with majority, soft or weighted voting to make the final prediction. In Stacking classifier, the weak classifiers produce a probability output and then all of them are fed to a final classifier, which is trained based on probabilistic outputs, rather than conventional features. Finally, the final classifier makes a prediction.

In the Regression category, most of the aforementioned models can also work efficiently. However, a classical model that belongs to this category is Linear Regression. In this case, a linear function tries to predict continuous values, utilizing least squares method.

### *3.2. Imputation & Normalization*

Empty values or features with very different numerical range can lead to bad models. With imputation, empty values are filled with predefined ones, using methods such as Mean/Mode or MICE. In the former, the values take the mean, median or most common values within a specific feature. In the latter, using a model such as Linear Regression or Random Forest, the empty values predicted by them, setting as input features all other non-empty ones, which exist in the dataset. The Normalization of data is achieved using MinMax or Z-Score. MinMax sets all values in the range $[0, 1]$, while Z-Score tries to give

gaussian like distribution to each features. Z-Score is beneficial for Machine Learning models that assume a-priori normal distribution of data.

### 3.3. Balancing

Balancing refers to the ratio between classes size. When one class has significant higher number of data, than the others, then the trained model could be biased towards this majority class. This means, that the model is more likely predicting the majority class. To this end, resampling techniques such as majority undersampling, minority oversampling and Synthetic Minority Oversampling have been developed. The majority undersampling constructs a new dataset by taking a specified percent of majority class data, so the classes ratio become 1. The second technique resamples many times the minority class data so their respective number in the new dataset is equal to majority class number of data. Finally, the last one creates new synthetic minority class data by applying linear interpolation between known data, thus the number of minority data is equal to the minority ones.

### 3.4. Feature Selection

Very often, some features existing in a dataset are not always useful. That means, they do not contribute to model performance comparing with their complexity and memory size, lead to overfitted models or have similar behaviour as other features of the dataset (linear correlated). So it is critical to choose the best subset of these features. The feature selection techniques, in general, can be categorized to the following three categories:

- **Filter**: Pearson correlation coefficient, chi-squared test and ANOVA.
- **Wrapper**: Sequential feature selection or backward elimination.
- **Embedded**: L1 and Ridge regression.

In addition, PCA finds the best features according to variance and SHAP is a relatively new method which evaluates feature contribution on predicted class.

### 3.5. Evaluation

The evaluation of a model predictive capability is the most important aspect of Machine Learning, since it depicts clearly, with numbers how well the model has done and allows comparisons between different models. The main idea of model evaluation is the comparison of models predicted value against the real value of an instance. Assuming, a binary classification problem, where there are a positive and a negative class, the confusion matrix illustriates principal evaluation metrics, who are used in Machine Learning.

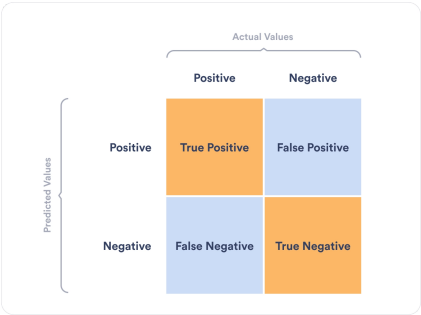



**Figure 1.** Confusion Matrix

The basic produced metrics are: Accuracy, Sensitivity, Specificity, F1-Score and AUC, listed below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Specificity = \frac{TN}{FP + TN} \tag{2}$$

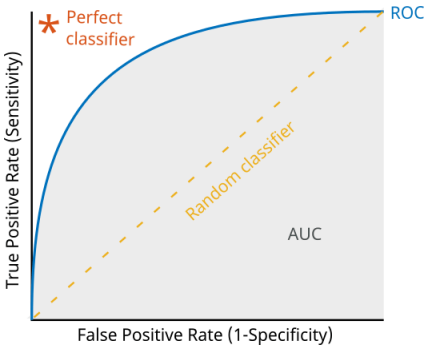



**Figure 2.** AUC and ROC

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

AUC is the area between the curve Sensitivity - (1-Specificity) Figure 2.

## 4. Relevant Sections

### *4.1. Related Work*

As mentioned before, Diabetes is a popular and well studied topic in the science world. Consequently, a vast majority of research paper has been dealt with the identification of diabetic cases, the long term prediction of diabetes development and the regression of critical biomarkers such as Fasting Plasma Glycose or HbA1c. Of course there are another use cases, such as classification/prediction of diabetes complications, as well as different input data e.g images [2,3]. Regarding bibliography quest, the research frame limited to T2D diabetes mellitus case using tabular data. To this end, high quality review publications were found in [2] and [3]. The first, handles the following sections: a) Prediction and Diagnosis, b) Diabetic Complications, c) Genetic Background and Environment, d) Drugs and Therapies, and e) Health Care and Management. Here, researchers note the majority (85%) of supervised algorithms found during their study, the superiority of SVMs, as well as the frequent utilization of clinical data such as Electronic Health Records (EHR), summing up optimistic, that diagnosis, etiopathophysiology and treatment of T2D through employment of machine learning and datamining techniques in enriched datasets that include clinical and biological information, have many benefits to offer during the continuous systematic exploration. The latter, as newer publication (published 2021), studies about 90 research papers, covering a broad space of the intersection between machine learning, deep learning and diabetes mellitus and demonstrates a variety of use cases, ranging from diabetes identification to complications prediction. Their useful findings are oriented to the following three axes:

- Datasets structure
- Top performing models
- Most frequent models
- Complementary techniques
- Asceding models
- Performance evaluation

Thus, clean and well structured datasets are preferred against big and complex ones. Random Forest and Decision Tree are the top performing models, achieving an AUC value of nearly 0.99. The top used models are Deep Neural Networks, tree-type and SVMs, however noting the advanced capabilities of the first one on dirty and very big datasets. The complementary techniques can improve the performance of every model, so resampling

techniques are employed and for the crucial job of feature selection Linear Regression and Principal Component Analysis seems to have a broader acceptance over scientists. Neural Networks are found by this research that perform better on large datasets and particularly on over 70,000 records. Finally, the difference on evaluation metrics from study to study caused some difficulties during their screening, as a consequence it's better to adopt common metrics such as AUC, Accuracy, Sensitivity and Specificity.

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

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 [4–6]. Regarding diabetes research field, the literature deals with the identification of diabetic people, early or long term (1-10 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 on 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[7] developed by Leicester University and FINDRISC [8] developed by University of Helsinki. The former utilizing a Logistic Regression model, took 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 Leicester 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 [2,3]. Another diabetes complications studies utilizing Machine Learning and Deep Learning include neuropathy and nephropathy [2,3]. 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 [2,3,9].

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.

#### 4.2.1. Current state classification

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 [10–19].

In [10], 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 into 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 achieving AUC 84.7%, misclassification rate 18.9%, Sensitivity 71.6% and Specificity 83.7% at threshold equal to 0.24. The information gain feature importance evaluation ranked the features with the following order: FBS, HDL, BMI. They summarized suggesting the incorporation of such models to online programms for further assistance of physicians during patient assessments.

In [11] a 138,000-records dataset containing 14 features such as age, pulse rate, breathe, diastolic and systolic pressure metrics, biometric data, physique index, fasting glucose, 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 an 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 was best with a value at 0.8084. Using only fasting glycose as input feature trains the models still better than PCA and mRMR techniques, yielding accuracy at 0.7597. Through feature screening procedure, FPG, weight and age were the most usefull. 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.

In [12] authors used also NHANES dataset to classify cases of diabetic and undiagnosed diabetic versus prediabetic and healthy, and undiagnosed diabetic and prediabetic versus healthy, using different feature subsets like survey data or laboratory results. First the data were standardized and then downsampled in order to produce a balanced dataset, which then splitted to 80/20 train/test set respectively. The models which were trained are Logistic Regression, SVM, Random Forest, XGBoost and an ensemble of all individual models. The ensemble model was a weighted soft voting classifier whose weights were the percentage of each individual model AUC in comparison with the sum of all AUCs. Formalistically, the calculation was performed according to the following equation

$$w_i = \frac{AUC_i^2}{\sum_{i=1}^{4} AUC_i^2} \tag{5}$$

. Each tunable model underwent hyperaparemeter tuning and the final performance was calcualted using 10-fold cross valdiation. For the first aforementioned classification case using only survey data (123 features, 1999-2014 period) XGBoost performed the best achieving AUC 0.862, Precision 0.78, Recall 0.78 and F1-Score 0.78. When laboratory results included, for period 1999-2014 XGBoost performed the best with AUC 0.957, Precision 0.89, Recall 0.89 and F1-Score 0.89. For the second classification case and for period 1999-2014 using survey data the ensemble model showed its superiority achieving AUC 0.737, Precision 0.68, Recall 0.68 and F1-Score 0.68. Utilizing laboratory data for period 1999-2014 XGBoost classified the cases with best scores, namely AUC 0.802, Precision 0.74, Recall 0.74 and F1-Score 0.74. The feature performance experiment which based on error rate, showed that waist circumference, age, leg length, sodium intake, blood osmolality,

blood urea nitrogen, Triglycerides, LDL, Choride, carbohydrates, Diastolic and Systolic pressure, BMI and fiber intake were the top ranked features. They concluded machine learned models based on survey questionnaire can provide an automated identification mechanism for patients at risk of diabetes and cardiovascular diseases. They also identified key contributors to the prediction, which can be further explored for their implications on electronic health records.

In [13] researchers utilized a 36,652 record-dataset from Henan rural cohort study, to test the ability of machine learning algorithms for predicting risk of type 2 diabetes mellitus (T2DM) in a rural Chinese population. Among features included were sociodemographic (e.g. Age, Sex, Income, Education), anthropometric (e.g. Waist circumference and waist to hip ratio), biometric (e.g. pulse pressure and heart rate), laboratory results (LDL-c, HDL-c, TG, Insulin, Creatinine, Uric acid, Urine glycose, etc.) and some history of diseases (Hypertension, Coronary Heart Disease and Family history of T2D). The positive class was defined as a positive diagnosis of T2D by a physician or an FPG value greater or equal than 7.0 mmol/L according to American Diabetes Association. At the preprocessing step, apart from data cleaning, the SMOTE method was employed in order to overcome the bias obstacle due to class imbalance. A number of linear, non-linear and ensemble models was employed on two datasets, the one containing laboratory results and the other excluding them. After hyperapameter tuning through 10-fold cv grid search, the Gradient Boosting Machine model gave the best evaluation metrics on both datasets, achieving AUC 87.2% and 81.7%, Accuracy 81.2% and 70.28%, Sensitivity 76.04% and 78.96%, and Specificity 81.71% and 69.43%, respecitively. Finally, using SHAP attribute evaluation this study reveals as most relevant predictors of T2D the urinary parameters, sweet flavor, age, heart rate and creatinine.

In [14] authors used NHANES dataset of 16429 records with nutritional, behavioural, socio-economic and non-modifiable demographic features (114 nutritional/dietary/food-intake associated; 13 other modifiable/health behaviour associated; 12 socio-economic/ demographic) spanning years 2007-2016. Missing values were imputed using MICE package. The dataset was divided into training, validation and testing set. Due to class imbalance, three new dataset variations were created using three sampling techniques, namely minority class Oversampling, Random Oversampling Examples and SMOTE. Logistic Regression, Artificial Neural and Random Forest were trained in the four datasets. For ANN parameter tuning was conducted whereas for the other two models the default parameters and 10-fold cross validation was utilized. The experiments showed that Logistic Regression, trained on minority oversampled dataset, was the best model achieving an AUC value of (75.35%). The odds ratio analysis of best performing logistic regression indicates folic acid, food folate, self reported health of diet and calcium as factors minimizing risk of diabetes, while total number of people in household, total fat, cigarette smoking, weight, BMI and waist circumference increase the risk. They concluded that their findings are a step towards personalised clinical nutrition, such as risk-stratified nutritional recommendations and early preventive strategies aimed at high risk individuals as well as in the nutritional management of people with type 2 diabetes.

In [15] four ordinary Machine Learning models were evaluated for their efficiency at classifiyng diabetic patients on 20,227 records taken from Department of Medical Services in Bangkok. The dataset, containing 10 typical demographic, biometric, heart pressure and rate results, as well as family history of diabetes, was normalized through MinMax algorithm and then Gain ratio utilized for the identification of most important features during class prediction. Thus, BMI and family history of diabetes were the most influencing attributes, which next played a primary role in the creation of the 'interaction variables'. Especially, as a first step, every continuous attribute was binarized (e.g age $\gtreqless$ 60, diastolic blood pressure $\gtreqless$ 90, etc. ). Then, BMI and family history of diabetes, due to their superiority were taken two by two with all the rest attributes with 'AND' clauses to create those interaction variables (e.g if BMI < 23 AND diastolic blood pressure < 90, then $Y = 0$, if family history of diabetes = True and age $\geq$ 60, then $Y = 1$, 0 otherwise, etc.). The

produced interaction variables where added to the existing 10-features datateset. Finally, after 80/20 train/test split and hyperapameters tunning, all the models were evaluated on both datasets. Among them, Random Forest trained on the dataset, included interaction terms, yeld the best metrics, namely, 97.5% Accuracy, 97.4% Precision and 96.6% Recall.

In [16], also NHANES dataset was utilized by Boosting, Random Forest, Logistic Regression and SVM models, to examine whether a diabetes case exists. After dataset balancing using SMOTE and attribute evaluation using backward feature selection as method and AIC (Akaike Information Criterion) as objective, 18 attributes came of after this preprocessing. The attributes can be categorized to four relevant sectors. Demographic, Dietary, Examination and Questionnaire features. Regarding models evaluation, CatBoost resulted to be the best, with metrics as follows: AUC 0.83, Accuracy 82.1%, Sensitivity 82% and Specificity 51.9%. SHAP evaluation also took part here, revealing that sleep time, energy and age are high influencing the diabetes outcome.

In [17], the LightGBM model was compared onto ZMHDD ethiopian Hospital dataset against typical pattern recognition models. Basic demographic, anthropometric as well as blood pressure, cholesterol, pulse rate and FBS data were utilized to predict the diabetes status of a patient. After median imputation of missing values and MinMax normalization the LightGBM model was trained using 10-fold cross validation for hyperapameters tuning. The optimized model achieved AUC 0.98, Accuracy 0.98, Sensitivity 0.99 and Specificity 0.96. Pearson correlation coefficient showed that FPG, total cholesterol and BMI have the strongest linear relationship with the diabetes prevalence among other attributes, even if their value did not reach more than 0.37. This high computational-efficient model outperformed the well established models in all metrics, providing a vital help to poor countries, which haven't the appropriate hardware to support more complex operations.

In [18], scientists examined the efficiency of a hard voting ensemble model against simple models on a 1,787 record- dataset (898 positive cases and 889 negative) from Centro Medico Nacional Siglo XXI at Mexico City. The feature selection using LASSO method yeld 12 features among them, sociodemographic, anthropometric, laboratory (urea, HDL under treatment, TG, diastolic pressure under treatment, systolic pressure without treatment),as well as the existence of hypertension and lipid medication. It's notable, that there are heart pressure data with and without treatment, offering a more objective knowledge for building the models. Z-score standardization was performed to have comparable features. An SVM, Linear Regression and Artificial Neural Network were trained and tuned on 75% of the dataset with 10-fold cross-validation, as well as the aforementioned hard voting ensemble incorporating all of them. The testing stage onto the remaining percentage of data showed, that single SVM has a slightly better performance on AUC compared to hard voting. Namely 92.8% against 90.5%, also Accuracy 89.82% Sensitivity 87.85% and Specificity 92.35%. Furthermore, the most usefull features were Lipids level in treatment and Hypertension treatment.

In [19] scientists probed a somehow reverse strategy. They did not used the typical features as the previous studies (only age and gender), but they examined only a variety of symptoms who come along with T2D. Those features are polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, itching, obesity, etc. The dataset consists of 520 records, 16 features and one target class for diabetes diagnosis, underwent balancing through SMOTE looking for the 5 nearest neighbours. The features relevance -importance evaluated using Pearson coefficient, Gain ratio, Naive Bayes AUC and Random Forest AUC. As result, polyuria, polydipsia, sudden weight loss and gender were the most influencing features. At the evaluation step a plethora of typical models was trained and tested as well as the impact of SMOTE and cross validation. Thus, the best models were Random Forest and KNN trained on a balanced dataset using 10-fold cross-validation. Namely, AUC 99.9% and 98.9% respecitively, Accuracy 98.59%, Recall 98.6% and Precision 98.6%.

#### 4.2.2. Biomarkers regression

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, [9] utilized 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 purpose 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. Outlier detection took part using the formula $\overline{X} \pm (3 \times \text{SD})$ and each outlier value was marked as missing. Records and features with more than 50% of missing values were dropped. The remaining missing values were imputed with MICE method. 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). The data partitioned into 6-months intervals (T6, T12, T18, T24, T30) according to the submission date of each record, thus 5 subdatasets were created and each Machine Learning model was trained in each subdataset and validated using 100 times random sampling with replacement (bootstrap). 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. When the whole dataset was available for training and testing, RF performed the lowest RMSE at 0.745 (95% CI 0.733-0.757). To measure how well the regressor fits the actual FPG value given the input features, $R^2$ coefficient was utilized. 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. Finally, the feature importance was assessed for every model through the five time-frame datasets using different metrics (because each model has different structure) like $\beta$-coefficient and permutation importance on MSE. 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.

#### 4.2.3. Long term prediction

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 [20–24]. In [20] the study examined a cohort of 7949 people with known and unknown family history of diabetes. The involved features 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. 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:

$$S_l = AUC_l^{val} \cdot (1 - \Theta_l) \tag{6}$$

Where

$$\Theta_l = \mu(\sigma(X'_{ijkl})) \tag{7}$$

and

$$X'_{ijkl} = \frac{X_{ijkl} - \mu(X_{ijkl})}{\sigma(X_{ijkl})} \tag{8}$$

. $X_{ijkl}$ 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 [21], scientist deployed regression, tree and gradient boosting models as well to predict the development of diabetes within 9 years. The dataset, consisted of 38,379 records and a large number of demographic, laboratory, pulmonary test, personal history and family history data, was imputed with mean/mode values and splited to 80/20 train/test. The tunable models were optimized with stratified 10-fold cross-validation. The comparison of models metrics showed that XGBoost achieved the best. That is, 0.623 AUC, 0.966 Accuracy, 0.970 Sensitivity and 0.690 Specificity. In addition, a survival analysis was performed utilizing Cox Regression and XGBoost Survival Embeddings to give a clear picture about the mean time to develop someone diabetes. Finally, the SHAP attribute evaluation showed that FPG, HbA1c and family history of diabetes contributed the most to diabetes, while, in contrary to [13], the uric acid contributed at the very least.

In [22], researchers evaluated the efficiency of Machine Learning models such as, Random Forest, Gradient Boosting, MLP and Naive Bayes against the classical Linear Regression. The cohort study included a total of 3,687 participants and by using the baseline-follow-up method, the prediction of the 3-year diabetes development is intended, taking into account demographic, smoking, drinking, history of health conditions and laboratory data. According to the LR feature analysis results, eight factors, including age, family history, impaired fasting glucose (IFG), impaired glucose tolerance (IGT), hypertension, triacylglycerol, alanine aminotransferase (ALT) and gamma glutamyl transpeptidase (GGT) were finally selected as modeling variables. The training phase included 75/25 train/test split, while the evaluation 10-fold cross validation and showed that Random Forest is the best proposed model. Then hyperapameter optimization using again 10-fold cross validation was performed and the final model was analysed against SHAP evaluation and this, in turn, compared with LR feature analysis. Finally, AUC of RF was 0.835.

In [23], 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 and Specificity were 0.856, 0.798 respectively. Finally, they concluded that due to the superiority of ensemble models, these can be embedded into recommendation systems to prevent patients from development of diabetes.

In [24] a dataset of 500,000 records from Hanaro medical foundation containing diagnostic results and questionnairies through 5 years, in the form of baseline-follow up, was utilized to conduct a multi classification experiment for prediction of prediabetic, diabetic and normal people in the following 1 year. Participants who suffered from a relative condition such as diabetes, hypertension and hyperlipidimia or took a respective medication, were excluded. Due to class imbalance, majority undersampling as well as SMOTE techniques were conducted, to avoid majority class bias. The extraction of important features consisted of two stages. In the first stage, continuous and nominal features were elected through ANOVA and $x^2$ test, respectively and in the second stage, 12 features were selected using Recursive Feature Elimination having as criterion an impurity index of a decision tree. Among the 12 features, were FPG, HbA1c, demographic data, BMI, Gamma glutamyl transpeptidase (gamma-GTP), uric acid, lifestyle habits (smoking, drinking) and family history of diabetes. This method showed, that FPG, HbA1c and gamma-GTP were the most informative features. Regarding model creation, logistic regression, SVM, Random Forest, XGBoost were compared to more sophisticated ensemble classifiers such as confusion matrix-based classifier integration approach (CIM), soft voting and stacking. 10 fold cross validation was used for both hyperparameters tuning with grid search and testing. CIM model slightly yeld the best results, with Accuracy, Precision and Recall values of 0.77. As a final experiment, researchers used follow up data from 1, 2, 3, 4, 5 consecutive years, proving that the more data exist, the better results receive.

In addition, in Figure 1, a summary of the reviewed literature is provided, whose purpose is to offer an olistic grasp about datasets used along with their unique features, complementary techniques added such as data normalization, resampling tecniques and most significant features extraction, and best models selection along with their respective evaluation metrics.

**Table 1.** Summary of reviewed studies.

| Study & Purpose | Dataset | Complementary techniques | Important Features | Best model |
|---|---|---|---|---|
| [9] Kopitar et al. FPG regression | 3,758 records, physical activity, lipid profile results, alcohol, smoking, diet, stress, cardiovascular results and health history. | Outlier detection, MICE imputation, Bootstrap random sampling with replacement, $R^2$ model calibration | Hyperglycemia, Age, Triglycerides, Cholesterol and blood pressure results | LightGBM, RMSE 0.8 mmol/L |
| [10] Lai et al. Diabetes classification | 13,309 records from CPCSSN[1]. Personal data and recent laboratory results | Misclassification cost matrix, grid search, adjusted threshold and 10 fold cross validation, information gain | FPG, HDL, BMI, Triglycerides | GBM AUC 0.847, Misclassification rate 0.189, Sensitivity 0.716 and Specificity 0.837 |
| [11] Zout et al. Diabetes identification | 138,000 records with glycose, physical examination, biometric, demographic and laboratory results | Random sampling, mRMR, PCA and 5 fold cv | FPG, Weight and Age | Random Forest using all 14 available features with Accuracy 0.8084, Sensitivity 0.8495 and Specificity 0.7673 |
| [12] Dinh et al. Diabetes identification | NHANES, survey data and laboratory results | Standardization, majority downsampling, ensemble model weighting optimization, hyperapameters tunin and 10-fold cv | Waist circumference, age, blood osmolality, Sodium, blood urea nitrogen and Triglycerides | Case 1:a) With survey data; XGBoost, AUC 0.862, Precision, Recall and F1-Score all 0.78 b) With laboratory results; AUC 0.957, Precision, Recall and F1-Score all 0.89. Case 2: a) With survey data; Ensemble, AUC 0.737, Precision, Recall and F1-Score all 0.68 b) With laboratory results; XGBoost, AUC 0.802, Precision, Recall and F1-Score all 0.74 |
| [13] Zhang et al. Diabetes identification | 36,652 from Henan rural cohort, including sociodemographic, anthropometric, biometric, laboratory results and history of diseases | SMOTE, hyperapameter tuning and 10-fold cv | Urinary parameters, sweet flavor, age, heart rate and creatinine | Experiments with and without laboratory results: Both XGBoost. AUC 0.872 and 0.817, Accuracy 0.812 and 0.702, Sensitivity 0.76 and 0.789, and Specificity 0.871 and 0.694 |
| [14] DeSilva et al. Diabetes identification | 16,429 records from NHANES with nutritional, behavioural, 146 socio-economic and non-modifiable demographic features | MICE Imputation, minority class oversampling, ROSE and SMOTE. Hyperparameter tunning, cv and Odds ratio | Folate, self reported diet health, number of people in household, total fat and cigarette consumption | Logistic Regression trained on minority oversampling dataset. AUC 0.746 |
| [15] Phongying et al. Diabetes identification | 20,227 records Department of Medical Services in Bangkok, including demographic, biometric, heart pressure and rate results and family history of diabetes data | MinMax normalization, Gain ratio, interaction variables and hyperapameter tuning | BMI and family history of diabetes | Random Forest trained on interaction variables dataset, achieving 0.975 Accuracy, 0.974 Precision and 0.966 Recall |
| [16] Qin et al. Diabetes identification | 17,833 records NHANES, including Demographic, Dietary, Examination and Questionnaire features | SMOTE, backward feature selection with objective AIC and SHAP | Sleep time, energy and age | CatBoost with AUC 0.83, Accuracy 0.821, Sensitivity 0.82 and Specificity 0.519 |

[1] Canadian Primary Care Sentinel Surveillance Network www.cpcssn.ca

**Table 2.** Continue

| Study & Purpose | Dataset | Complementary techniques | Important Features | Best model |
|---|---|---|---|---|
| [17] Rufo et al. Diabetes identification | 2,109 records from ZMHDD hospital, including demographic, anthropometric, blood pressure, cholesterol, pulse rate and FBS data | Median imputation, MinMax normalization, Pearson correlation coefficient hyperapameters tuning and 10-fold cv | FPG, total cholesterol and BMI | LightGBM. AUC 0.98, Accuracy 0.98, Sensitivity 0.99 and Specificity 0.96 |
| [18] Benita et al. Diabetes identification | 1,787 records from Centro Medico Nacional Siglo XXI at Mexico City, including sociodemographic, anthropometric, laboratory data such as HDL and diastolic pressure under treatment and systolic without treatment | Standardization, LASSO feature selection and hyperapameter tuning with 10-fold cv | Lipids level in treatment and Hypertension treatment | SVM achievning AUC 0.928, Accuracy 0.898, Sensitivity 0.878 and Specificity 0.923 |
| [19] Dritsas et al. Diabetes identification | 520 records dataset from Kaggle, including symptoms such as polyuria, polydipsia, sudden 226 weight loss, weakness, polyphagia, genital thrush, itching, obesity, etc | SMOTE using 5-NN, Pearson coefficient, Gain ratio, AUC of NB and RF, and 10-fold cv | Polyuria, polydipsia, sudden weight loss and gender | Random Forest and KNN, achieving AUC 0.99 and 0.98 respecitively, Accuracy 0.985, Recall 0.986 and Precision 0.986 |
| [20] Lama et al. Long term diabetes prediction | 7,949 records, socioeconomic and psychosocial factors, physical and laboratory results, physical activity, diet information and tobacco use | Median imputation, SHAP, 5-fold cv grid search with objective as 6. Risk profiles | BMI, waist-hip ratio, age, systolic and diastolic BP, and diabetes heredity | Random Forest, AUC 0.7795 |
| [21] Shin et al. Long term diabetes prediction | 38,379 records including demographic, laboratory, pulmonary test, personal history and family history data | Mean/mode imputation, hyperapameter tuning with stratified 10-fold cv, SHAP and survival analysis | FPG, HbA1c and family history of diabetes | XGBoost with 0.623 AUC, 0.966 Accuracy, 0.970 Sensitivity and 0.690 Specificity |
| [22] Mao et al. Long term diabetes prediction | 3,687 records, including demographic, smoking, drinking, history of health conditions and laboratory data | LR feature analysis, hyperparameter tuning, 10-fold cv and SHAP | Age, Impaired Fasting Glycose and Glycose Tolerance. | Random Forest with AUC 0.835 |
| [23] Fazakis et al. Long term diabetes prediction | 2,331 records from ELSA, including biometric, anthropometric, hematological, lifestyle, sociodemographic and performance index variables | Feature selection techniques: LASSO, Correlation, Greedy stepwise. Random undersampling. Adjusted threshold with objective $J^3$, multiobjective optimization | Not applicable | Weighted Soft Voting ensemble with base classifiers LR and RF. AUC 0.884, Sensitivity 0.856 and Specificity 0.798 |
| [24] Deberneh et al. Multiclass long term diabetes prediction | 500,000 records containing diagnostic results and questionnairies | Majority undersampling and SMOTE, ANOVA, $x^2$ and RFE, grid search, 10fold cross validation. | FPG, HbA1c, gamma-GTP | CIM Accuracy, Precision, Recall 0.77 |

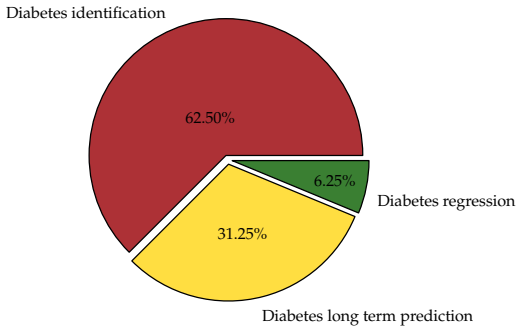



**Figure 3.** Percent of reviewed studies by purpose

## 5. Discussion

Interpret the findings based on evidences(tables)

The findings are oriented towards the following thematic axes.

1. Types of hypotheses addressing diabetes through Machine Learning using tabular data.
2. Data preprocessing.
3. Features involved.
4. Selection and identification of most important Features.
5. Methodology structure towards model building.
6. Evaluation metrics.
7. Best models.

In the end, relevant literature referring review studies are compared with ours.

### *5.1. Types of hypotheses addressing diabetes through Machine Learning using tabular data*

To start with, recent applications on diabetes using tabular datacan be categorized to three discrete hypotheses. Current state diabetes identification, long term diabetes prediction and biomarkers regression. As Figure 3 depicts, the majority of hypotheses are in the context of current state diabetes identification with 62.50%, then long term diabetes prediction with 31.25% and lastly diabetes biomarkers regression with 6.25% in a sum of 16 articles. This distribution is quite reasonable, because each hypothesis has different data collection requirements. First of all, the current state diabetes identification requires only present health condition, while in long term case, the class variable refering to diabetes state is filled many years after measurements completion, maintaining a baseline-follow up method. Thus, the current state study has simpler data collection process than the long term one. Then, biomarkers regression such as FPG or HbA1c has even more complex dataset creation, because the target variable should be measured continually and systematically by the individuals with an invasive way. Moreover, the features values should be filled every time along with the measurements by indivudials, something that could lead to false or uncompleted data, due to lack of professional intervetion. Among them, the most challening and interesting use case is the long term forecasting, because it would provide an early assessment of diabetes development.

### *5.2. Data preprocessing*

Most research articles studied, give high importance in data preprocessing techniques such as empty values imputation, data balancing and transformation, probing a variety of them. The imputation is done with Mean/Mode or the more complex one MICE, while most of reviewed researches do not handle this issue, rather drop the empty values due to high data quantity. The evidences show that there is not a go-to method, but it depends on every specific problem setting, however MICE utilization in datasets with relatively low percent of missing data could simulate possible hidden patterns between features and therefore produce more realistic datasets.

Data balancing is undoubtly one of the most important stages, even thought the models have the highlights, because balancing adjusts the bias of the whole experiment. As

from literature deducted, general approaches include oversampling minority instances, undersampling majority instances, synthesizing artificial data (SMOTE) in order to product equal instances for all classes. Interestingly, Fazakis et al. [23] successfully used undersampling technique to match real life positive diabetic cases age wise, because with only 2,000 instances an equal size of classess could lead to significant bias towards positive diabetic cases. Also, Fazakis et al [23] and Lai et al [10] utilized adjusted decision threshold of their classifier, yielding very good results.

Data transformation techniques included only in [12,18] (Standardization) and in [15,17](MinMax). Standardization contributes in general in all models, while MinMax normalization does not affect tree based models. This in contrary to [15,17], which trained LightGBM and Random Forest, respectively, mainly for fair comparisons against other MinMax dependent models (SVM and KNN).

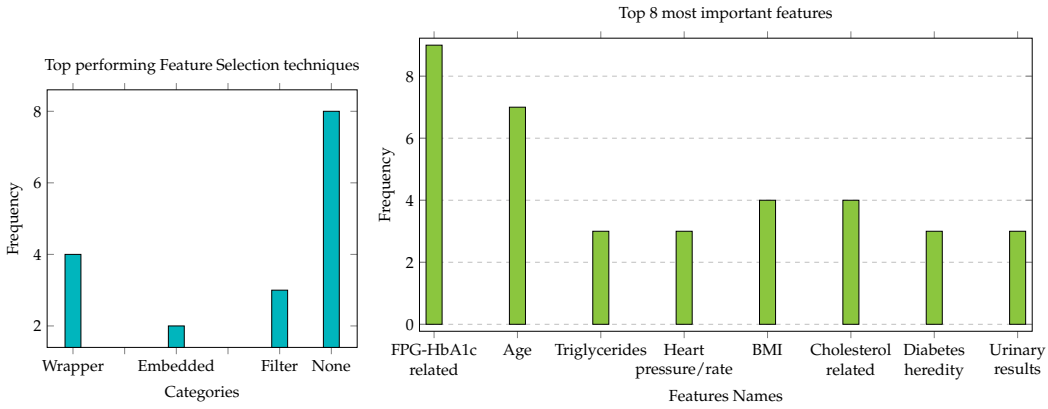
### *5.3. Features involved*

Most datasets have a size of some thousands records (excluding [19] with only 520) and include a variety of data representing many aspects of humans. Datasets are mainly provided by clinics, hospitals and institutes such as NHANES [11,14,16]. The features involved are sociodemographic like age, sex, education level, salary and marital status, anthropometric/biometric like height, weight, BMI, waist circumference, systolic and diastolic pressure, and pulse rate, laboratory results like FPG, HbA1c, HDL, LDL, total cholesterol, triglycerides, urea measurements, creatinine and gamma-GTP, lifestyle behaviour such as sleep time, smoking, drinking and physical activity, family history of diabetes and dietary consumption data such as folate, carbohydrates and sugar [14,16]. Most of those features are verified by medical literature that associate with diabetes ([1]), while others such as urea measurements, gamma-GTP and dietary habits are under investigation. An advantage of tabular data is the low storage overhead comparing to images which would demand thousands times more memory to be stored. However, such a plethora of features demands a systematic and time consuming registration effort, while also the participants could not be available to all measurements.

### *5.4. Selection and identification of most important features*

Regarding the aforementioned problem, the application of feature selection techniques could contribute remove features which seem to be unrelated to diabetes and thus help the data collection process, as well as the computational efficiency of model building. As can be seen from Figure 4a, the top performed models utilized a variety of methods which belong to Wrapper, Filter and Embedded. The most notable finding is that the majority of researchers chose to promote models trained on the whole feature set, rather than to a reduced version. A reason for that is the scope of each study. For example De Silva et al [14] examined dietary features contribution, while [19] presented the influence of common diabetic symptoms. However, the most useful result of such a study is the interpretability of associated factors, in order to moderate them towards risk minimization. Luckily, with the new advancements of Machine Learning, the prediction models are not any more black boxes, but with feature important methods, every feature contribution can be quantified. For instance, [14] using odds ratio metric, found folate consumption, self reported diet health and total fat consumption, while [19] highlights the significance of polyuria, polydipsia sudden weight loss and gender as best indicators of diabetes, using Pearson correlation, Gain ratio, Random Forest and Naive Bayes AUC. Another finding, is the usual deployment of SHAP method, (like in[16,20–22]), which provides a common interface for all types of models to quantify feature influence, as well as the influence of features on a particular participant [20]. Summing up all of them, in Figure 4b an enumeration of best predictor categories during this review can be seen. Glycose related biomarkers FPG and HbA1c, as expected, found to be the most important predictors as they are known to be the principal indicators of diabetes [1]. Age, BMI, heart operation metrics, Lipid profile and diabetes heredity seem to agree both on such data driven techniques as

medical practice, in contrary to ethnicity which is not highlighted on the reviewed studies. Finally, urinary parameters seem to be promising about their predictive capabilities ([13,19]), while dietary factors should be more investigated [13,14,20].




**(a)** **(b)**

**Figure 4**

### *5.5. Methodology structure towards model building*

In general this stage offers wide freedom of management. However, our findings present with clarity a common pattern that exists in model building. Firstly, a number of researches use many different feature subsets, that arose either from a variety of feature selection methods or from medical bibliography or from the hypothesis, which is examined [9,11–13,15,23]. Datasets with laboratory data are compared against non invasive data, to evaluate if the latter can provide relatively reliable results [12]. According to table 1, the vast majority studies prefer train/test split usually with 80/20 ratio, hyperparameter tuning using mostly grid search and 10-fold cross validation. Due to the data quantity, the split percent as well as the number of folds are considered as good choice. The figure 5 presents a typical workflow towards model building, as extracted by the reviewed literature, (e.g balancing at preprocessing, wrapper or none feature selection technique, the training procedure, the evaluation metrics and the popular feature importance methods).

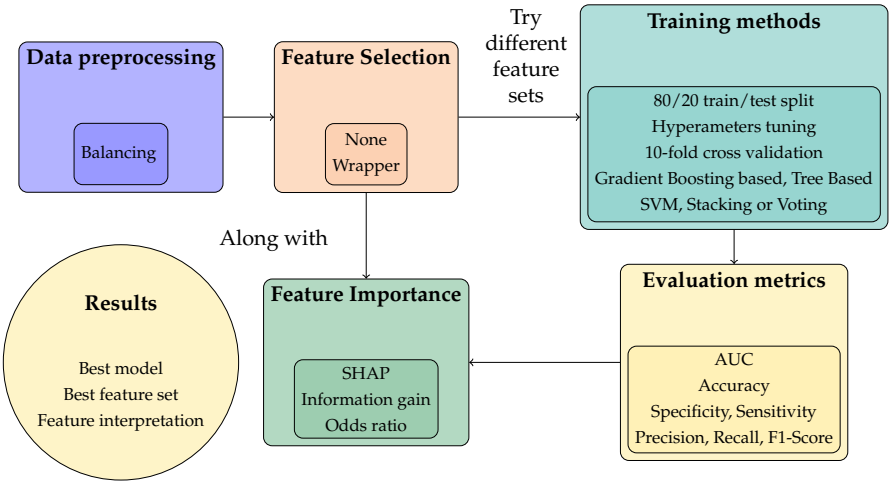



**Figure 5.** Typical methodology workflow

### 5.6. Evaluation metrics

In most of examined studies, the evaluation metrics are the typical ones such as AUC, Accuracy, Sensitivity or Recall, Specificity, Precision and F1-Score. These provide a complete understanding of performance. However, in such use cases as disease detection, Sensitivity must have the priority because the higher it is, the less possible is for an individual with diabetes to diagnosed as healthy. Other metrics used are misclassification rate in [10], AIC in [16], robustness in [20] and Youden index in [23]. Generally, researchers should be urged to include different existing or new, custom metrics.

### 5.7. Best models

Figure 6 depicts the general categories of top models as elected on reviewed studies, clustered by the hypothesis. For the identification purpose, Gradient Boosting models have a clear advantage, with Random Forest coming second. For long term forecasting, Stacking and Voting share the top with Random Forest, while for biomarkers regression the one and only model preferred belongs to Gradient Boosting family. SVM, kNN as well as logistic regression are not yet preferred, first of all due to their worse performance and secondly because these models demand data transformation such as MinMax normalization due to data heterogenity, which would be time and resource consuming in datasets of some decades of thousands. On the other hand Gradient Boosting and tree like models stay unaffected by the different numerical ranges of features.

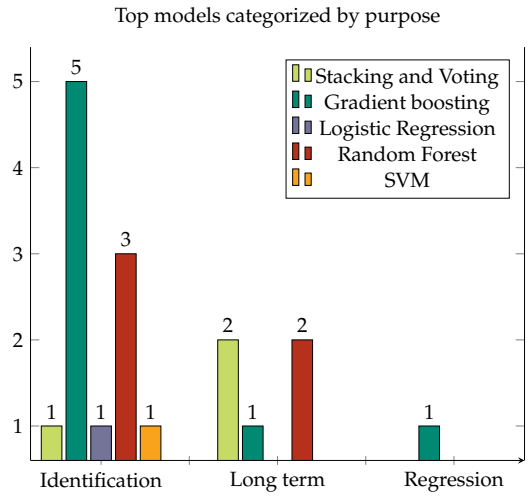



**Figure 6.** Top performing models categorized by purpose

### 5.8. Comparing to previous reviews

Both reviews [2,3] examine a broader context of Artificial Intelligence and T2D mellitus, including deep learning, unsupervised learning and association rules. Kavakiotis et al [2], although it is older publication, deals with a more general overview of applied methods, which includes hypotheses about diabetic complications prediction, data driven investigation of both drugs and therapies, genetic background and enviroment, and health care management. In contrast to our research, they found SVM to be the best model towards diabetes hypotheses. Fregoso et al [3] as a publication of 2021, agrees with us in the finding that tree based, ensemble models have the top performance. In contrary to us, they found Linear Regression coefficients and PCA as the most popular feature selection techniques and heterogenity in model assessment metrics. SHAPley values are not of great concern in both studies, while as proved to our study, it is a wide acceptable interpretation method. Our limitations, include the examination of quite new articles along with few high quality older ones, that use tabular data and machine learning models. An advancement of our research is that it is trying to provide an in-depth, detailed methodology description of all

overviewed articles, unveil common successfull patterns, highlight new asceding features and subsequently provide new research areas.

## 6. Conclusions

By their nature, the identification, or even better a very early forecast of T2D mellitus, are very interesting cases of study, because people could be able to stay informed about their health situation and try to prevent further negative development. In this study we overviewed the recent Machine Learning applications towards T2D mellitus prediction using tabular data such as demographic, biometric, laboratory, lifestyle and dietary data, with a goal to investigate common patterns towards ML models implementation and discover both asceding methods and features. Interestingly we found Gradient Boosting and Tree based models as the most successfull ones, which are trained and optimized usually with grid search, as well as that SHAPley and Wrapper algorithms are the quite popular feature interpretation and evaluation methods. Apart from classical laboratory biomarkers, urinary results and non-invasive dietary informations are promising features opening new research areas, which possibly could lead to prevention or easier management of T2D mellitus.

**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 |

## References

1. About diabetes. https://idf.org/aboutdiabetes. Accessed 09/06/2023.
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. 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.
4. 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.
5. 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.
6. Seabold, S.; Perktold, J. statsmodels: Econometric and statistical modeling with python. In Proceedings of the 9th Python in Science Conference, 2010.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. Phongying, M.; Hiriote, S. Diabetes Classification Using Machine Learning Techniques. *Computation* **2023**, *11*. https://doi.org/10.3390/computation11050096.

16. Qin, Y.; Wu, J.; Xiao, W.; Wang, K.; Huang, A.; Liu, B.; Yu, J.; Li, C.; Yu, F.; Ren, Z. Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type. *International Journal of Environmental Research and Public Health* **2022**, *19*. https://doi.org/10.3390/ijerph192215027.
17. 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.
18. 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.
19. Dritsas, E.; Trigka, M. Data-Driven Machine-Learning Methods for Diabetes Risk Prediction. *Sensors* **2022**, *22*. https://doi.org/10.3390/s22145304.
20. 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.
21. Shin, J.; Lee, J.; Ko, T.; Lee, K.; Choi, Y.; Kim, H.S. Improving Machine Learning Diabetes Prediction Models for the Utmost Clinical Effectiveness. *Journal of Personalized Medicine* **2022**, *12*. https://doi.org/10.3390/jpm12111899.
22. Mao, Y.; Zhu, Z.; Pan, S.; Lin, W.; Liang, J.; Huang, H.; Li, L.; Wen, J.; Chen, G. Value of machine learning algorithms for predicting diabetes risk: A subset analysis from a real-world retrospective cohort study. *Journal of Diabetes Investigation* **2023**, *14*, 309–320, [https://onlinelibrary.wiley.com/doi/pdf/10.1111/jdi.13937]. https://doi.org/https://doi.org/10.1111/jdi.13937.
23. 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.
24. 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.