Translational Medicine Communications

RESEARCH Open Access

Predictive model and feature importance for early detection of type II diabetes mellitus

Eric Adua[1,2*], Emmanuel Awuni Kolog[3*], Ebenezer Afrifa-Yamoah[2], Bright Amankwah[4], Christian Obirikorang[5], Enoch Odame Anto[1,6], Emmanuel Acheampong[2], Wei Wang[1,7] and Antonia Yarney Tetteh[1]

Abstract

Background: Accurate prediction and early recognition of type II diabetes (T2DM) will lead to timely and meaningful interventions, while preventing T2DM associated complications. In this context, machine learning (ML) is promising, as it can transform vast amount of T2DM data into clinically relevant information. This study compares multiple ML techniques for predictive modelling based on different T2DM associated variables in an African population, Ghana.

Methods: The study involved 219 T2DM patients and 219 healthy individuals who were recruited from the hospital and the local community, respectively. Anthropometric and biochemical information including glycated haemoglobin (HbA1c), body mass index (BMI), blood pressure, fasting blood sugar (FBS), serum lipids [(total cholesterol (TC), triglycerides (TG), high and low-density lipoprotein cholesterol (HDL-c and LDL-c)] were collected. From this data, four ML classification algorithms including Naïve-Bayes (NB), K-Nearest Neighbor (KNN), Support Vector Machines (SVM) and Decision Tree (DT) were used to predict T2DM. Precision, Recall, F1-Scores, Receiver Operating Characteristics (ROC) scores and the confusion matrix were computed to determine the performance of the various algorithms while the importance of the feature attributes was determined by recursive feature elimination technique.

Results: All the classifiers performed beyond the acceptable threshold of 70% for Precision, Recall, F-score and Accuracy. After building the predictive model, 82% of diabetic test data was detected by the NB classifier, of which 93% were accurately predicted. The SVM classifier was the second-best performing classifier which yielded an overall accuracy of 84%. The non-T2DM test data yielded an accurate prediction score of 75% from the 98% of the proportion of the non-T2DM test data. KNN and DT yielded accuracies of 83% and 81%, respectively. NB had the best performance (AUC = 0.87) followed by SVM (AUC = 0.84), KNN (AUC = 0.85) and DT (AUC = 0.81). The best three feature attributes, in order of importance, were HbA1c, TC and BMI whereas the least three importance of the features were Age, HDL-c and LDL-c.

Conclusion: Based on the predictive performance and high accuracy, the study has shown the potential of ML as a robust forecasting tool for T2DM. Our results can be a benchmark for guiding policy decisions in T2DM surveillance in resource and medical expertise limited countries such as Ghana.

Keywords: Type II diabetes, Machine learning, Feature extraction, Prediction, Risk factors

Background

Advances in research and technology have revolutionalised medicine, resulting in improved health outcomes of complex diseases and enhancing longevity. However, there is still much to be achieved in regard to preventing and controlling diabetes mellitus (DM) and its effects or burden has been far reaching (1). From developing to

*Correspondence: eric.adua@ecc.edu.au; eakolog@ug.edu.gh
[1] Department of Biochemistry and Biotechnology, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
[3] Department of Operations and Management Information Systems, University of Ghana, Legon, Accra, Ghana
Full list of author information is available at the end of the article

developed countries, the disease affects 1 in 11 people, and over 400 million people die from DM every year [1]. It is estimated that the DM prevalence will rise from 415 million in 2015 to 640 million in 2040 while 232 million people are not even aware of their status [1, 2].

A significant proportion of diabetes carriers (>90%) identify as having type II diabetes mellitus (T2DM)—a condition of the inability to control plasma glucose due to insulin insufficiency or resistance [1]. When affected, T2DM makes individuals unproductive [3, 4], disables them [5] and renders patients and their families financially impoverished due to life-long spending on medical and hospital bills [6]. Sadly, its effects are not just experienced by the affected alone but also hugely impact the global economy [2]. The American Diabetes Association has even stated that if the current trends of diabetes persist, the economic cost of diabetes will reach $2.1 trillion by 2030 [2].

Despite its widespread implications, there is still no cure for the disease [1, 7]. Current treatments only provide relieve by modifying disease-associated symptoms. Meanwhile, the long latency period of the disease allows for targeting and tackling disease before conditions become irreversible [8–10]. In this latency period, the timing of detection and the accuracy of diagnosis are crucial for ensuring predictive, preventive and personalized medicine (PPPM) [11].

Defined as the medical practice that systematically predicts the onset of chronic disease long before its clinical manifestation, PPPM has the potential to influence treatment in time and influence optimal therapies [8, 12]. Further, PPPM is beneficial in multiple ways including 1) delaying the onset of a chronic disease, 2) designing of targeted drugs, establishing the efficacy, potency and adverse effects of drugs on patients, patient stratification and prevention of disease-associated complication [8, 12–15]. However, before the concept of PPPM can be operationalised, there is the need to recognise the existence of risk factors that are associated with human lifestyles and how these factors influence cardiometabolic health.

Majority of large-scale studies have shown that factors that are antecedence of T2DM are age [16], obesity [17], physical inactivity [17, 18], unhealthy diet [19–21], high blood pressure [22–24], high plasma glucose and high cholesterol levels [25]. With this knowledge, risk estimation scores have been developed including the Framingham Cardiovascular Disease (CVD) risk score [26] and the Systematic Coronary Risk Evaluation (SCORE) established by the European Society of Cardiology [27]. While these scores could signal if an individual will develop a disease, they are built from simple or few models and fail to account for complex variables [28]. Other studies have explored the use of the Suboptimal Health Status Score as a predictor of cardiometabolic diseases [10, 29, 30]. These studies have largely relied on traditional logistic regression and multivariate regression models to make predictions. Although beneficial, the reliance on conventional regression is short-sighted, given they provide a modest information about the interaction between predictors. In addition, logistic regression might be less computationally demanding but does not provide optimal predictions when there is a nonlinear interactions between factors [31] or when there is an imbalance in the number of cases and controls [32]. To overcome these, there is a need for a more advanced predictive tool such as machine learning (ML). ML does not make any statistical assumptions, such as normality, collinearity, linearity or nonlinearity, when building a predictive model. It has proven to be robust in building a predictive model and diversely used in domains such as education, health and business.

ML relies on algorithms that learn from observations or features and create models [33, 34]. Based on these observations, ML scans for patterns, highlights the complex interactions between the predictors and ultimately, optimizes the performance of predictors. Moreover, ML display a better discriminatory power [35], operate with less focus on data distribution [36], handle multidimensional data and create models from big data or utilized for real-time association analysis [37, 38]. Given its ability to transform data into a meaningful information, its application is now seen in medicine. When employed in clinical data, ML learns patterns of health trajectories of patients, can review or expose patient charts and detects subclinical abnormalities in several chronic diseases including coronary artery disease, cardiovascular, rheumatoid arthritis as well as T2DM [39].

Due to the rapid generation of big clinical data and the quest for accurate predictions, the interest in ML has increased dramatically [32, 40–45]. For example, Lai et al. (2019) [33] used Gradient Boosting Machine (GBM) and logistic regression to predict the onset of diabetes in a Canadian population. This study revealed an area under receiver operating curve (AROC) of 84.7% with 71.6% sensitive and 84.0% with 73.4% for GBM and logistic regression, respectively. While this study has shown novel insights, the outcome cannot be generalized. Other ML techniques have also been used elsewhere. Zou et al. [44] used random forest (RF), neural network, and decision tree (DT) to predict diabetes in a population in Luzhou. However, the study could only identify which algorithm was superior to the other and was not able to adequately predict diabetes due to limited indices and imbalanced data [44].

Using feature selection method on a cohort of diabetes individuals, Sneha and Gangil (2019) revealed that RF and DT algorithms had the highest specificities of 98.20% and 98.00% respectively [46]. Utilising four ML methods including k-nearest neighbors (KNN), multifactor logistic regression, multifactor dimensionality reduction and support vector machines (SVM), Farran et al. (2013) reported classification accuracies of 85% of diabetes and 90% for hypertension. Sneha and Gangil (2019) explored the performance of six ML algorithms (RF, Naïve-Bayes (NB), KNN, DT) and SVM. The researchers developed a predictive model for diabetes dataset for each of the ML algorithms. Out of the fifteen (15) attributes in the dataset, ten (10) were found, through a feature selection technique, to produce an optimal predictive model. The researchers generalized the selection of optimal features from the dataset to improve the classification accuracy. The results of their study found DT algorithm and RF to be the highest at 98.20% and 98.00%, respectively.

T2DM arises from the interplay between genetic and environmentally acquired factors including diet, race or ethnicity. Hence, the present study uses four ML algorithms, 1) NB, 2) SVM, 3) KNN and 4) DT to identify predictors of T2DM in ethnically distinct population, Ghana. Moreover, this study ranks the order of importance of the various attributes in the diabetic dataset.

Methodology

Methods and study design

Recruitment of patients was based on a purposive sampling approach where T2DM patients who visited Komfo Anokye Teaching Hospital (KATH) for their medications were asked to participate. After this, we used a convenient sampling approach to recruit healthy individuals from three popular suburbs within the Kumasi metropolis.

Ethics approval

The study was approved by the Kwame Nkrumah University of Science and Technology (KNUST) in Ghana, the Committee on Human Research, Publication and Ethics (CHRPE), and the Human Research Ethics Committee (HREC), Edith Cowan University (ECU). Each of the participants signed an informed consent prior to participating in the study.

Anthropometric examination

Aided by a standard sphygmomanometer (Omron HEM711DLX, UK), blood pressure measurements (Systolic blood pressure (SBP) and diastolic blood pressure (DBP) were noted and recorded. Estimation of body fat was by Body Mass Index (BMI) which is calculated as BMI = weight (kg)/height (m)^2. Waist to height (WHtR ratio was measured as waist (cm)/height (cm).

Clinical data

Fasting blood samples were taken from antecubital vein of each participant into a gel separator, EDTA and fluoride oxalate coated tubes. Serum lipids, comprising total cholesterol (TC), high-density lipoprotein -cholesterol (HDL-c), low-density lipoprotein -cholesterol (LDL-c) and triglycerides (TG), were measured using an automated chemistry analyser (Roche Diagnostics, COBAS INTEGRA 400 Plus, USA). On the same instrument, glycated haemoglobin (HbA1c) in EDTA tubes and fasting blood sugar (FBS) in fluoride tubes were also measured.

Inclusion and exclusion criteria

Cases

T2DM patients who have been clinically assessed by a medical doctor were invited to participate. Those who were identified as having type I diabetes mellitus or in any form of insulin treatment were excluded. The study excluded 34 T2DM from the 253 T2DM patients because of missing information. Thus, 219 participants were included in the final analysis.

Controls

Participants diagnosed with diabetes and/or hypertension were excluded. Moreover, those with digestive, respiratory, genitourinary disorders were excluded. At the end, 219 healthy individuals were included.

The mean age for the cases was 56.54 ± 9.89 and controls was 55.10 ± 9.27. The number of females outnumbered the males (i.e. 61.4% females in controls and 57.3% females in cases) but the difference did not reach statistical significance (p = 0.80). Most of the participants were educated and employed. T2DM patients were primarily sedentary when compared with controls but there was no statistical difference in BMI between the groups. Generally, T2DM patients had higher FBS, HbA1c, and HDL-c when compared with controls. The controls had higher SBP and DBP but WHtR, TC, TG and LDL-c were not statistically different between the groups (Table 1).

Experiment

Data Pre-processing and Feature selection

Figure 1 shows the process model of the classification in this work. As indicated in the figure, the data for each of the attribute is numeric with different form or scaling. The steps in the process model include *cleaning, scaling, feature selection, test/train validation, classifier model building* and *evaluation*. With this dataset, there was no issues with data imbalance as the number of T2DM patients in the dataset (N = 219) was the same as the number of persons without T2DM (controls) (N = 219).

Table 1 Demographic information of T2DM patients and healthy controls

Variable		**Control (n = 219)**	**Case (n = 219)**	***P*-value**
Age (mean ± SD)		55.10 ± 9.27	56.34 ± 9.76	0.1690
Age groups (years)				
	31–40 years	8 (3.7)	13(5.9)	0.1160
	41–50 years	70(32.0)	49(22.4)	
	51–60 years	83(37.9)	82(37.4)	
	61–70 years	44(20.1)	60(27.4)	
	71–80 years	14(6.4)	15(6.8)	
Gender				
	Female	135 (61.4)	133 (57.30)	
BMI(Kg/m^2)				0.8030
	Underweight	11(5.0)	8(3.6)	
	Normal	91(41.6)	95(43.6)	
	Overweight	74(33.8)	73(33.5)	
	Obese	43(19.6)	43(19.7)	
Education				**0.0400***
	Tertiary	29(13.3)	39(17.8)	
	Senior high	72(33.0)	50(22.8)	
	Junior high	71(32.6)	72(32.9)	
	Lower primary	28(12.8)	25(11.4)	
	No formal education	18(8.3)	33(15.1)	
Occupation				**< 0.0001***
	Employed	147(67.4)	144(66.1)	
	Retired	21(9.6)	26(11.9)	
	Keeping house	14(6.4)	21(9.5)	
	Unemployed	26(16.6)	27(12.5)	
Physical activity				
	Sedentary	30(13.8)	49(22.4)	**0.0250***
	Moderate activity	114(52.3)	89(40.6)	
	Active	74(34.0)	81(37.0)	
Clinical/biochemical data				
WHtR		0.56 ± 0.08	0.56 ± 0.08	0.6060
SBP (mmHg)		145.88 ± 24.33	139.61 ± 24.88	**0.0080***
DBP (mmHg)		84.63 ± 14.42	82.40 ± 13.22	0.0940
FBS (mmol/l)		5.86 ± 0.95	9.23 ± 4.31	**< 0.00001***
HbA1c (mmol)		5.30 ± 0.77	8.35 ± 2.17	**< 0.00001***
TC (mmol/l)		4.69 ± 1.26	4.57 ± 1.18	**0.0272***
TG (mmol/l)		1.35 ± 0.97	1.22 ± 0.53	0.1120
HDL-c(mmol/l)		1.24 ± 0.33	1.365 ± 0.32	**0.0001***
LDL-c(mmol/l)		2.88 ± 1.05	2.65 ± 1.09	**0.0270***

Data presented as Mean ± SD. Tests of significance were two tailed (*$p < 0.05$) and bolded

The dataset contained 438 instances (participants) with eleven (11) different features (attributes). While the attributes *Age, BMI, SBP, DBP, HbA1c, FBS, TC, TG, HDL-c, LDL-c* are the predictor variables (attributes), the *T2DM* class is the target variable. This division is essential especially that the approach is to build a predictive model with ML.

Among other factors, the performance of a classification algorithm is largely dependent on the quality of the data. Data that is fraught with errors, such as outliers, influence the performance of a machine algorithm [47]. Hence, the diabetes dataset used in this study was explored to eliminate outliers and errors. By visualizing the data through the lens of boxplot, no outlier was detected. In all the datapoints in the dataset, only

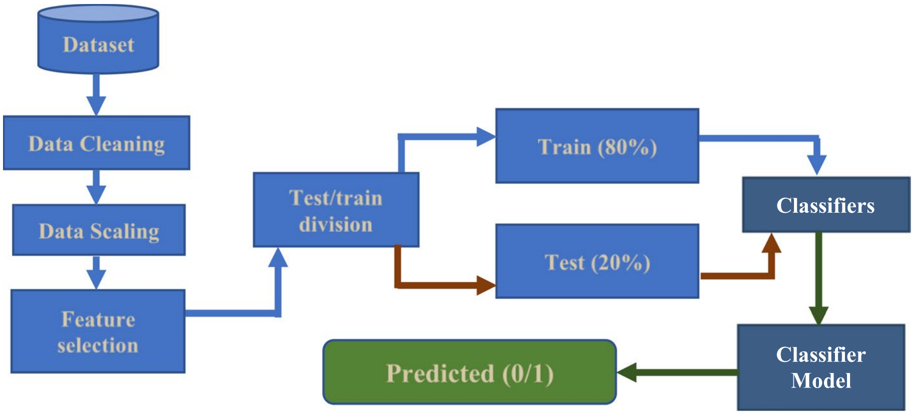


Fig. 1 Model for Classification where 0 is non-T2DM and 1 is T2DM

seventeen (17) of them were found missing. Hence, *Expectation–Maximization (EM)* algorithm was employed to compute for the missing data. EM algorithm incorporates statistical considerations to compute the "most likely, or maximum-likelihood, source distribution that would have created the observed projection data, including the effects of counting statistics " [48].

To improve the performance of the algorithms and eliminate any possible bias, the predictive variables were scaled to a range of (0,1). Data scaling is a method used in ML to normalize the range of predictive variables (features) of data. Furthermore, the importance of each of the attributes (predictor variables) used in this study was explored and ranked according to their respective coefficients. The ranking demonstrates which among the attributes is/are the most important and least important for detecting diabetes. We leveraged on the *Recursive Feature Elimination (*RFE*)* from Scikit-learn using Python to compute and rank the importance of each of the attributes. RFE works by recursively removing attributes and building a model to rank the attributes [49]. It uses the model accuracy (coefficient) to identify which attribute is/are the most important in terms of their predictive influence.

Classification

The predictive models in this study were built on four different ML algorithms: *KNN), SVM*, *NB* and*DT.* While this study sought to predict T2DM and rank the order of the predictive importance of the feature attributes, the goal was to also compare the performance of each of the model classifiers in predicting the unseen data. These classifiers were selected based on their efficacy and the fact that they have been used widely for text classification (Kolog et. al., 2019). Altogether, the total instances of the data used in this study was 438 (219 each for the cases and controls. As shown in Fig. 1, we used *train-test* technique (42) to build the predictive models in this study. With this technique, the data was split into two, where 80% was used to train the algorithms. The remaining 20% of the data was used to test the algorithms. Of the 438 instances of the data (both controls and cases), only 350 (80%) was used for the training while the remaining 68 (20%) was used for the testing. The division of the data into testing and training was random. To avoid imbalance classification, the 350 instances of the training data comprised of 175 each for the case data (diabetic patients) and control data (non-diabetic patients).

NB are probabilistic classifiers that use Bayesian theorem with naïve independent assumptions between the features or attributes [50] (Domingos & Pazzani, 1997). There are three main types of NB algorithms: *Multinomial Naive Bayes, Gaussian Naive Bayes* and *Bernoulli Naive Bayes.* These types are identified according to their classification techniques. Gaussian Naive Bayes was employed in this study because of its versatility to handle both continuous and discrete data. For instance, when the predictors take up a continuous value, Gaussian assumes that these values are sampled from a gaussian distribution. In our study, we sought to predict patients who are with T2DM (cases) or not (controls), C_k (where C_1 = diabetes and C_0 = non-diabetes) given that its predictor variables are x_1, x_2,…,x_p which can be expressed as $P(C_k|x_1,…,xp)$. The Bayesian formula for calculating this probability is Eq. 1. From the equation, $P(C_k)$ is the *prior* probability of the outcome, $P(x)$ is the probability of the predictor variables, $P(x|C_k)$ is the *conditional probability* or *likelihood* and $P(C_k|x)$ is called our *posterior probability.* This is further expressed in Eq. 2.

$$P(C_k|x) = \frac{P(C_k).P(x|C_k)}{P(x)} \tag{1}$$

$$Posterior = \frac{PriorxLikelihood}{Evidence} \tag{2}$$

DT is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility [41]. It consists of nodes and leaves expressed in hierarchical layers as indicated in Fig. 2. Each of the nodes is a divergent point where a particular characteristic of the data is tested, and the data split accordingly [51]. Just like the other ML algorithms, DT is not built based on any statistical assumption of the data, such as normality, collinearity or correlation between explanatory variables. The capability of DT classifiers has prompted its application in diverse domains. It can be used for decision analysis in management sciences and operations research. The algorithms are nevertheless popular ML applications for classification problems.

SVM was originally developed for binary classification, but it was later extended for multiple classifications. It is one of the most popularly used algorithms for both classification and regression due to its efficacy. What the algorithm does is to construct a line (hyperplane (s)) in datapoints expressed in high dimensional vector space [52]. As indicated in Fig. 3, the larger the margin the lower the *generalization error* of the classifier. Therefore, a hyperplane that is farther from the nearest training data point of any class (functional margin) is well separated.

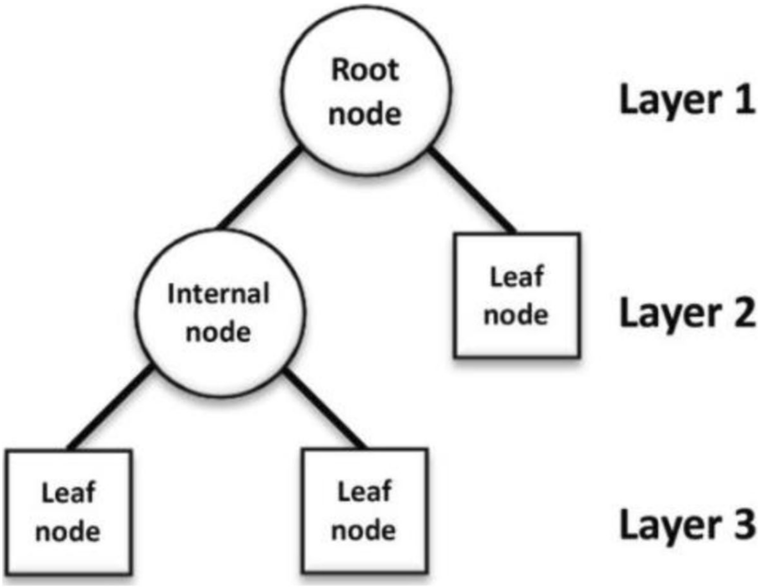


Fig. 2 Basic structure of decision tree showing three different hierarchical layers

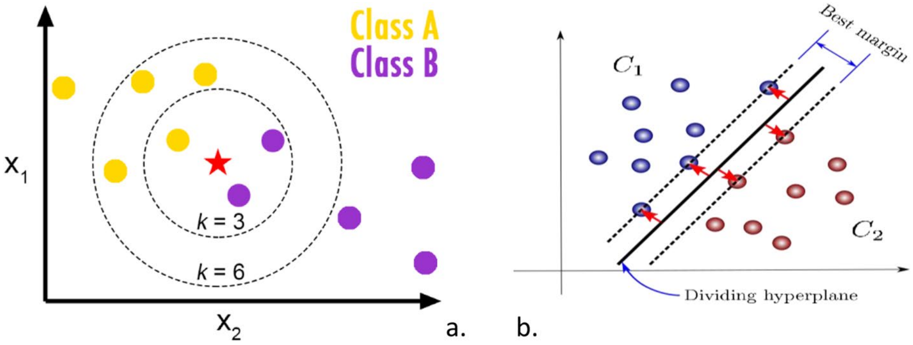


Fig. 3 Pictorial representation of **a**) K-Nearest neighbor classification on datapoints **b**) SVM classification on datapoints in high dimensional vector space

SVM algorithms use a set of mathematical functions that are defined as the kernel. Kernel function contains a mathematical function that takes data as input and transforms it into the required form. Examples of SVM kernel functions are linear, nonlinear, polynomial, radial basis function (RBF) and sigmoid. In this work, we tried all the kernel functions on our data and later arrived at using RBF due to its optimality on our data.

KNN algorithm is a classification algorithm that works by using distance matrix to find *k* most similar instances in the training data for test data [53].The mean outcome of the neighbors is taken as the prediction. Just like k-means in clustering, KNN algorithm commonly uses Euclidean distance. Mathematically, lets represent x_i as input sample with *p* features (x_{i1}, x_{i2},…,x_{ip}), *n* be the total number of input samples (i = 1,2,…,n) and *p* the total number of features (j = 1,2,…,p) (69). The Euclidean distance between datapoints is given by Eq. 3. In this study, we implement KNN from sklearn machine learning library.

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots\ldots + (x_{ip} - x_{jp})^{2\cdot}} \quad (3)$$

Classifiers evaluation

The ML algorithms used in this study were evaluated according to their predictive strengths. Thus, we computed for the *Precision, Recall, F1-score* and *Accuracy* of the algorithms. *Recall* is the proportion of the instances of the test data that were correctly identified by the classifier model based on the trained data, while *Precision* is the proportion of the identified instances of the data that were accurately predicted by the algorithms. The harmonic means of Precision and Recall constitute the F1-score or F-measure. Given the number of real positive (p) cases and the number of real negative (n) cases in the data, the precision, recall and F1-score are indicated in Eqs. 3 – 5, where *tp* is *true positive, fp* is *false positive* and *fn* is *false negative.*

$$Precision = \frac{tp}{tp + fp} \quad (4)$$

$$Precision = \frac{tp}{tp + fn} \quad (5)$$

$$Precision = 2x\frac{PrecisionxRecall}{Precision + Recall} \quad (6)$$

Additionally, we computed the Area under Receiver Operating Characteristics curve (AROC) and the confusion matrix of the algorithms. The figures shown in the second column of Table 3 are ROC curves which depict the abilities of the various classifiers. The discriminatory thresholds of the classifiers are varied. ROC curves are typically used in binary classification to study the output of a predictive model. As indicated in Table 3, ROC curve typically features true positive (sensitivity) rate on the y-axis and false positive rate (1-specificity) on the x-axis.

Confusion matrix, also called *the error matrix,* is a tabular representation of the performance of an algorithm. The computation of the confusion matrix prompted the number of instances that were correctly predicted and falsely predicted. As indicated in Table 3, the first columns contain contingency tables (confusion matrix) for each of the algorithm. From the confusion matrix tables, the *predicted class* is on the row while the *actual class* is at the column.

Results and analysis

Descriptive

The mean of most of the attributes for both the patients with T2DM and without T2DM vary but insignificantly. A notable significant difference is the means of HbA1c and FBS for T2DM and those without T2DM. The mean score of the HbA1c of T2DM patients (mean = 8.1) is higher than that of the non-T2DM patients (mean = 5.3). Generally, the mean score of the parameters in patients with T2DM was higher than the patients without T2DM except for SBP and DBP (Fig. 4).

Relationship structure among features

There exists reasonable overlap in the classification of T2DM and non-T2DM cases based on the two orthogonal linear combinations of the features that explain most of the variability in the data (Fig. 5). In terms of relationships among predictors, there exist a strong positive relationship between SBP and DBP, LDL-c and TC, Age and TG and FBS and HbA1c, based on the angles between the vectors for the features (< 30°). SBP, DBP, BMI, Age and TG seem to be uncorrelated with HbA1c, FBS, HDL-c, as the angle between these vectors is approximately 90°. SBP and DBP contribute highly to classifying the control group, whilst HbA1c, FBS and HDL-c are most influential in classifying subjects with T2DM.

Classification

Table 2 shows the performance of the various ML algorithms in terms of their scores in Precision, Recall, F1-score, weighted average and Accuracy. As indicated in Table 2, all the classifiers performed beyond the acceptable threshold of 70% for Precision, Recall, F1-score and Accuracy. However, the performance of the individual classifiers varied slightly in all the parameters. From the table, after building the predictive model with NB, 82% of diabetic test data was detected by the algorithm of

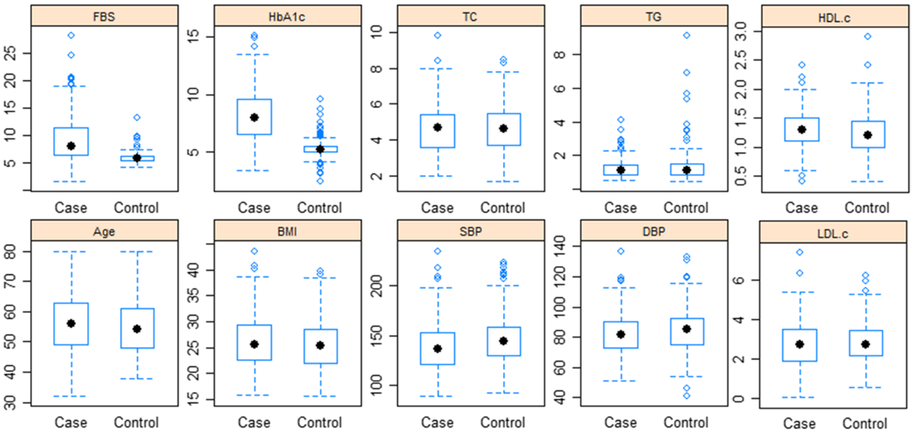


Fig. 4 Distribution of study features across T2DM and non-T2DM subjects. Note that HbA1c is *glycated haemoglobin*, TC is Total *Cholesterol*; BMI is body mass index; FBS is *Fasting blood sugar*; DBP is Diastolic Blood Pressure; TG is *Triglycerides*; SBP is *Systolic Blood Pressure*; HDL-c is high-density lipoprotein cholesterol; *LDL-c is density lipoprotein cholesterol*

which only 93% of them were accurately predicted. The F1-score for diabetic patients was 87%. With regards to the non-T2DM data, 93% of the instances of the test data were detected by the NB algorithm of which only 82% of the detected instances of the non-T2DM test data were accurately predicted. The overall accuracy of the NB algorithm is 87%, which is the highest of the performance of all the algorithms.

The SVM algorithm was the second-best performing algorithm. The algorithm yielded an overall accuracy of 84% for predicting both the cases (T2DM data) and controls (non-T2DM data) as contained in the test data. With SVM performance, only 73% of the T2DM test data was detected but 97% of the detected instances of the data were accurately predicted (precision). In a similar vein, the non-T2DM test data yielded an accurate prediction score of 75% from 98% of the detected proportion of the non-T2DM test data.

KNN and DT yielded overall accuracies of 83% and 81% respectively. Although NB and SVM were better, the performance of DT and KNN signify a good predictive strength. However, the KNN performed better than that of the DT though both classifiers exceed the accepted threshold of 70%. With regards to both algorithms, more than 70% of the instances of the test data were detected by the respective algorithms for both the diabetic and non-diabetic data. Of the detected test data, more than 70% were accurately predicted by the KNN and DT

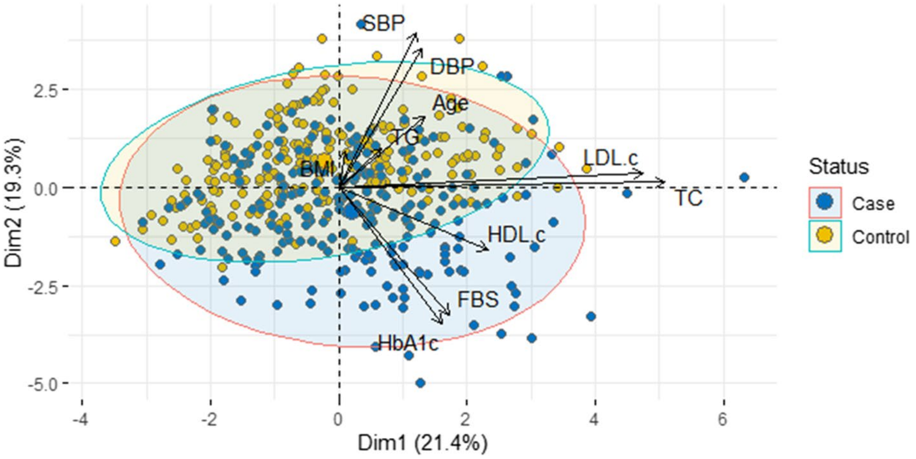


Fig. 5 Principal component analysis (PCA) biplot illustrating the relationships among study features and the clustering patterns in subjects with T2DM and no T2DM based on orthogonal linear combinations of the features. Glycated Haemoglobin (HbA1c), Total cholesterol (TC); Body Mass Index (BMI); Fasting Blood Sugar (FBS); Diastolic blood pressure (DBP); Triglycerides (TG); Systolic Blood Pressure (SBP); High Density Lipoprotein cholesterol (HDL-c); Density Lipoprotein cholesterol (LDL-c)

Table 2 Performance of the classifiers for cases and control

	With all Features				
Classifiers	**Measures**	**T2DM**	**Non-T2DM**	**Weighted avg**	**Accuracy**
NB	Precision	0.93	0.81	0.87	**87%**
	Recall	0.82	0.93	0.87	
	F1- score	0.87	0.86	0.87	
SVM	Precision	0.97	0.75	0.97	**84%**
	Recall	0.73	0.98	0.73	
	F1- score	0.84	0.85	0.84	
KNN	Precision	0.90	0.77	0.84	
	Recall	0.78	0.90	0.84	**83%**
	F1- score	0.84	0.83	0.83	
DT	Precision	0.80	0.78	0.79	
	Recall	0.82	0.76	0.79	**81%**
	F1- score	0.81	0.77	0.79	

Where NB-Naïve-Bayes, SVM -support vector machine, KNN- K-nearest neighbor and DT- Decision tree

classifiers. NB outperformed the other algorithms in terms of ROC, sensitivity, specificity, accuracy and Kappa (Fig. 6).

Table 3 contains both the Confusion matrix and ROC curves of the various ML algorithms for the test set. As earlier described, ROC curve provides the overall assessment of the predictive models. The figures at the right of Table 3 show ROC curves of the four classifiers (NB, KNN, SVM, DT). The top left corner of each of the plot is the "ideal" point—a *false positive rate* of zero (0), and a *true positive rate* of one (1). However, it is highly unrealistic to obtain the extreme Area Under Curve (AUC) score of exactly 0 or 1. Nevertheless, in an ideal situation, AUC of 0.90–1.0 = *excellent*, 0.80–0.90 = *good*, 0.70–0.80 = *fair*, 0.60–0.70 = *poor* and 0.50–0.60 = *fail* (Kleinbaum & Klein, 2010). The AUC measures discrimination and the models classify the cases and controls. Therefore, the larger the area bounded to the reference line, the better in terms of the predictive model. From Table 3, the AUC of all the classifiers was beyond 0.80 (80%) but less than 0.90 (90%) indicating "good" predictive models. While recognising the *good* performance of the ROC curve, some of the classifiers performed better than others. NB has the best performance (AUC = 0.87) followed by SVM (AUC = 0.84), KNN (AUC = 0.85) and DT (AUC = 0.81). This performance shows how well the algorithms discriminate on the dataset.

Feature importance

The predictor attributes of the T2DM dataset used in this study were ranked according to their predictive influence on the target variable (T2DM status). Our feature extraction indicates the relevance of all the features for predicting T2DM. Hence, all the attributes were used for building and testing the predictive models with the various ML algorithms. Although the various attributes were relevant for building the predictive model, the order of importance of each of the feature attributes was computed and ranked according to their strength of influence on predicting T2DM. As indicated in Fig. 7, the best three feature attributes, in order of importance, are *HbA1c, TC* and BMI (rankings were uniform across four ML algorithms). The highly ranked feature attributes are very important when detecting T2DM and these should be prioritized accordingly. While we recognise that all the feature attributes are essential for detecting T2DM,

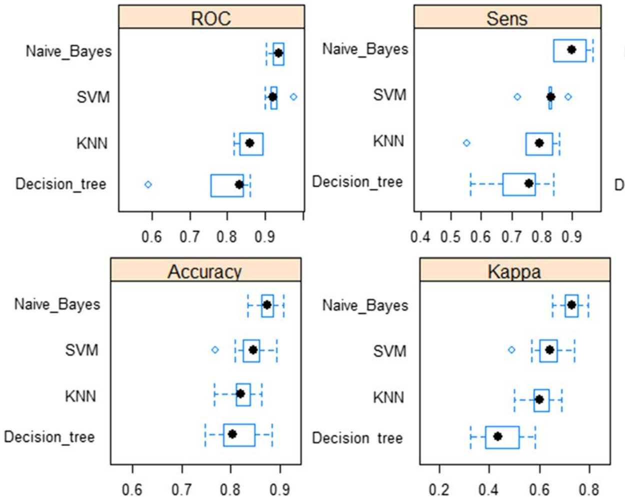


Fig. 6 Comparative analysis of the four machine learning algorithms across sensitivity analysis statistics for the training set. *ROC = Receiver Operating Curve, Sens = Sensitivity score and Spec = Specificity score*

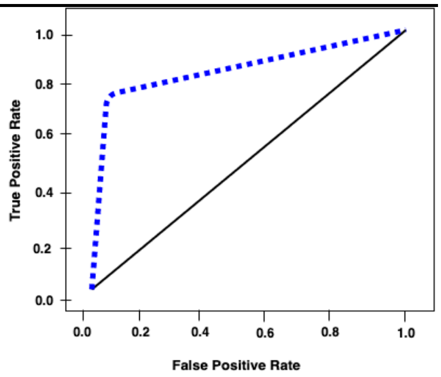
Table 3 Confusion Matrix and ROC curve for each of the classifiers

Confusion Matrix | **ROC Curve**

Naïve-Bayes

		PREDICTED	
		T2DM	**Non-T2DM**
ACTUAL	T2DM	40	1
	Non-T2DM	13	36

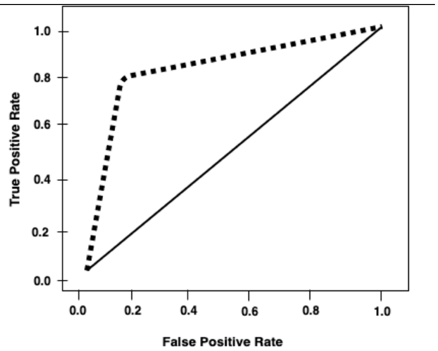
AUC= 87.2 (0.872)



K-Nearest Neighbor

		PREDICTED	
		T2DM	**Non-T2DM**
ACTUAL	**T2DM**	37	4
	Non-T2DM	11	38

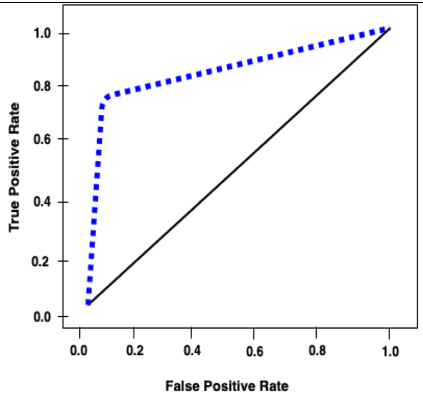
AUC = 83.9 (0.839)



Support Vector Machines

AUC = 85.5% (0.855)

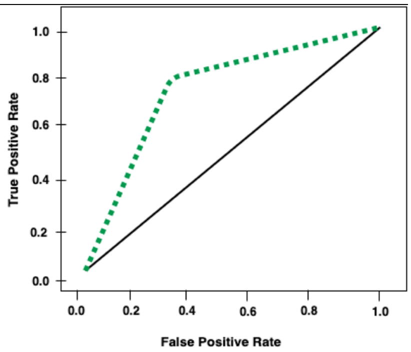
		PREDICTED	
		T2DM	**Non-T2DM**
ACTUAL	T2DM	38	3
	No- T2DM	9	40



Decision Tree

AUC = 80.9 (0.809)

		PREDICTED	
		T2DM	**Non-T2DM**
ACTUAL	T2DM	30	11
	No- T2DM	9	40



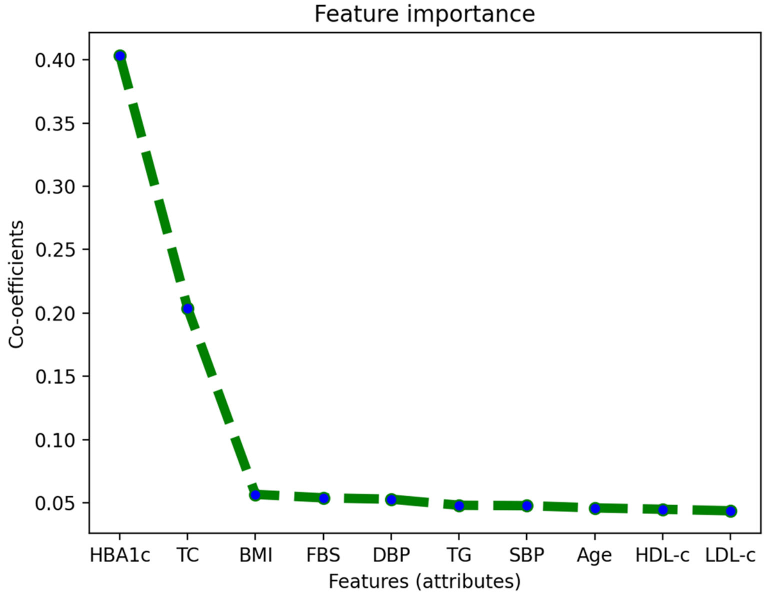


Fig. 7 Feature importance across four machine learning algorithms. Glycated haemoglobin (HbA1c), Total cholesterol (Chol); Body Mass Index (BMI); Fasting Blood Sugar (FBS); Diastolic Blood Pressure (DBP); Triglycerides (TG); Systolic Blood Pressure (SBP); High Density Lipoprotein cholesterol (HDL-c); Density Lipoprotein cholesterol (LDL-c)

this current study has found Age, HDL-c and LDL-c *as* the least of the feature attributes for predicting T2DM.

The ML model was developed to consider Age as one of the attributes. Age is known in literature to affect the other attributes of T2DM. In this study, the ML algorithms inherently aggregated and extracted the weighted the age attribute for the classifier to learn to build a model towards prediction. Therefore, variations in the Age of the patients, in respect of the other attributes, influence the predictive strength of the model towards predicting the test data. For instance, if the Age of patients are highly tilted above 50 years, the indicators of the other attributes are affected, and in effect, influence the predictive model. Likewise, when the Ages are tilted below 40 years the other attributes will vary and the predictive model affected.

Discussion

The emergence of ML techniques has fuelled interest in the predictive modelling of cardiometabolic diseases [43, 44]. In this Ghanaian cohort study, we have demonstrated that ML algorithms can accurately predict T2DM based on laboratory results and anthropometric data. In doing so, four ML classification algorithms NB, KNN, SVM and DT were compared. The predictive performance was generally good for all algorithms. In the analysis, it was found that NB was the best performing classifier with AROC of 87.20%, and also in terms of sensitivity, specificity, accuracy and kappa (Fig. 6). This finding agrees with that of Sisodia and Sisodia (2018) who reported NB (AROC of 76.30%) as the best predictor of diabetes in pregnant women [43]. Sneha and Gangil, (2019) also showed that NB had the best accuracy when compared with DT [46].

The study identified SVM as the second-best performer and having a good discriminatory power. This agrees with previous research. For example, reporting an AUC of 83.47%, Yu et al., [35] highlighted that SVM is efficient, can predict diabetes and outperforms logistic regression in population health surveys. The reason for the good discriminatory power of SVM is suggested to be due to the large margin between hyperplanes that allows for the separation of classes in three dimensional vector space [43]. However, one of the limitations of SVM in terms of its performance on data is the size of the data. SVM has been found in literature to perform extremely well when the dataset is large. Although the performance in this study was above the accepted threshold of the 70%, the size of the dataset could have affected the performance.

With regards to KNN classifier, the accuracy was 81.0% and AROC of 83.9%. While this result is significant (> 70%), it is possible the result could have been affected by bias variance trade off [54]. Despite the fact that KNN is sensitive to the quality of the data, it is also sensitive to the scale of the data and irrelevant features. Hence, features that exhibited weak predictive strength could have influenced the performance of KNN. The DT, which performed significantly but poorly among the classifiers used in the study, is a probabilistic algorithm which works well when the attributes are extremely unique. The poor performance among the other classifiers may have occurred as a result of its sensitivity to small perturbations in the data.

With ML methods exhibiting high Precision, Recall, F1-score, Weighted average and Accuracy (Table 3), the study has demonstrated the ability of the ML techniques to correctly predict T2DM. For example, NB could identify the presence of T2DM in 82 patients out of 100 T2DM patients, SVM could identify 73 out of 100; KNN could identify 78 out of 100 and DT could identify 82 out of 100 T2DM patients. This is especially important as a lower recall rates can lead to misdiagnosis of T2DM.

The phenotypic expression of T2DM is due to a continuum of risk factors. The present study identified 9 variables including blood pressure, FBS, TC, TRG, BMI, HDL-c and LDL-c as predictors of T2DM. Particularly, with regards to the order of importance, we identified HbA1c, TC and BMI as the top three primary predictors of T2DM. These findings are not unexpected but validate those of previous studies [34, 42, 45]. Hitherto, the measurement of FBS was considered the surest way to determining prediabetes and diabetes. However, due to daily fluctuations of glucose levels, there was a need for alternative biomarkers (55). In the course of research, it was known that sugars are pinned to residues of globin chains and forms 1-deoxy-1-N-valyl-fructose after an Amadori rearrangement [55]. Later, this product became known as glycated haemoglobin (HbA1c). While the level of FBS is still the basis for the diagnosis of prediabetes and diabetes in most laboratories, this research has indicated that HbA1c is sensitive and more reliable for diagnosing diabetes than FBS [7, 56]. Further, HbA1c has leverage over FBS by being stable and can detect plasma glucose levels in the previous 3 months. From.

Figure 5, our results confirm that of previous studies that HbA1c is superior to FBS in T2DM diagnosis. Although some researchers prefer other obesity measures to BMI [57–60], BMI is widely used as an indicator of excess body fat and a risk factor for cardiometabolic disease [61, 62]. The use of BMI in the present study instead of the other fat indicators such as waist circumference, abdominal obesity and visceral body fat is justifiable in the light of a previous study (63). Based on 1288 subjects, Bouchard, (2007) revealed the bidirectional relationship between BMI and other fat measures. Specifically, the study showed that BMI strongly correlated with excess fat mass ($r=0.94$), waist or abdominal obesity ($r=0.93$), and abdominal visceral fat ($r=0.72$) [63]. These results are comparable to the findings of several previous studies [31, 35, 64]. For example, using a neural network model, Akella and Kaushik (2020) identified resting blood pressure, serum cholesterol and blood glucose as part of the top 10 variables of importance in cardiovascular disease prediction [65].

LDL-c and HDL-c are important molecules that are dysregulated or modified in T2DM. In T2DM, there is a decline of HDL-c due to the formation of TRG rich HDL-c. TRG-HDL-c is a substrate for hepatic lipases that catalyses the breakdown of HDL-c [66]. Conversely, there is a reduction in the catabolism of LDL-c in T2DM leading to increased levels of LDL-c. This decrease has been attributed to a decline in the expression of apolipoprotein B and apolipoprotein E receptors as well as a decreased affinity of LDL-c [67].

Dinh et al. [39] have revealed that age is a key risk factor for cardiovascular events and diabetes [68] because ageing is linked to physical inactivity and ultimately, T2DM. However, in the present study, our feature selection technique revealed age to be one of the risk factors in T2DM albeit among the least predictors of T2DM. This is to imply that ageing is a determining attribute for T2DM detection. However, other attributes such as HbA1c, TC and BMI ought to be considered before Age when diagnosing for T2DM. It is worth noting that the aged people may physically exhibite symptoms of T2DM but may not necessarily be diabetic. This is may be the reason whyHbA1c, TC and BMI are the most important attributes for diagnosing T2DM. This result also agrees with those reported in the literature. Comparing multiple variables including invasive laboratory data and non-laboratory data (non-invasive), Dinh et al., [39] documented that age was the fifth predictor of diabetes behind LDL-c, TRG, blood urea nitrogen, sodium and blood osmolality. However, in the absence of laboratory variables, their results showed that age was the second most important feature for predicting diabetes.

It should be clear by now that ML can adequately predict T2DM in a Ghanaian population. However, some limitations need to be mentioned. Firstly, the sample size of the participants was small and the prediction may be over/underestimated. However, this does not invalidate our results since Kuhn and Max [36] has stated that large sample sizes, though beneficial, increase computational burden and can impact the results. Secondly, it is important to note that other potential risk factors including family history, physical activity exist, but they were not considered in this study. It is expected that the inclusion of these will further enhance the predictive model. Thirdly, the impact of antidiabetic medications should not be overlooked. Some of the medications being used to control T2DM in this population include glucose-lowering (e.g.

biguanides, thiazolidinediones, sulfonylureas); lipid-lowering (statins) and antihypertensives (e.g. angiotensin II receptor blockers, calcium channel blockers) [29]. Thus, the interpretation of the results should be viewed in light of medication use. Going forward, we will explore the potentiality of ML methods for discovering other biomarkers of T2DM.

The study seeks to influence policy and practice in the various health facilities in Ghana. The present study recommends clinicians to first test HbA1c, TC and BMI for T2DM before any other parameter could be considered. This study underscores the fact that some T2DM risk factors are more important, in terms of their predictive strength, than the other risk factors. Hence, our study attempts to reduce the cost of diagnosing T2DM. Indeed, HbA1c and FBS were the strong biomarkers for predicting diabetes in the Ghanaian population. The data for the current study comprised both undiagnosed (including at-risk) and diagnosed diabetes individuals. Identifying individuals with undiagnosed diabetes has been a challenge but our results reinforce the relevance of HbA1c or FBS for early detection of diabetes or prediabetes. Once detected, such individuals can be targeted for tailored treatments that will delay them from developing the disease. Based on the analysed data, these attributes are enough to show diabetes patients.

Conclusion

Using multiple variables as substrates, the study has shown that ML can generate accurate predictions of T2DM and provide potentially meaningful information. We identified NB as the best algorithm in predicating T2DM when compared with KNN, DT and SVM. When employed, these algorithms can allow the early detection of T2DM, anticipate future events and in turn, stimulate a timely intervention. It is hoped that the findings of this study will guide the selection of appropriate ML algorithm for the prediction of T2DM and help health professionals in Ghana to make well-informed and better decisions.

Acknowledgements
We wish to acknowledge the immense support from the staff at the Diabetic Clinic of KATH, Kumasi.

Authors' contributions
EA and EAK conceived the study; EA collected data, processed blood samples. EA, EAK and EA-Y drafted the manuscript and revised for intellectual content. EAK and EA-Y performed statistical analysis, machine learning and interpreted results. EA, EAK, BA, CO, EOA, EA, WW, AYT and EA-Y edited the ideas and concepts presented. All authors read and approved the final manuscript.

Funding
The study was supported by Edith Cowan University Collaboration Enhancement Scheme 2017.

Availability of data and materials
The data that support the findings of this study are available from Edith Cowan University database, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the University.

Declarations

Ethics approval and consent to participate
This study was reviewed and approved by the Committee on Human Research, Publication and Ethics (CHRPE), Kwame Nkrumah University of Science and Technology (KNUST), Kumasi, and the Human Research Ethics Committee (HREC), Edith Cowan University (ECU), Australia. Written informed consent was obtained from all participants. Further, the study was conducted in accordance with the principles of the Declaration of Helsinki.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Biochemistry and Biotechnology, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. [2]School of Medical and Health Sciences, Edith Cowan University, Perth, Australia. [3]Department of Operations and Management Information Systems, University of Ghana, Legon, Accra, Ghana. [4]Biochemistry & Cellular and Molecular Biology, University of Tennessee, Knoxville, USA. [5]Department of Molecular Medicine, School of Medical Science, Kwame Nkrumah, University of Science and Technology, Kumasi, Ghana. [6]Department of Medical Diagnostics, Faculty of Allied Health Sciences, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. [7]Beijing Key Laboratory of Clinical Epidemiology, Capital Medical University, Beijing 100069, China.

Received: 4 February 2021 Accepted: 19 July 2021
Published online: 23 August 2021

References

1. International Diabetes Federation: IDF diabetes Atlas 9th edition 2019. https://www.diabetesatlas.org/en. Accessed 20 May 2020.
2. Bommer C, Sagalova V, Heesemann E, Manne-Goehler J, Atun R, Bärnighausen T, Davies J, Vollmer S. Global economic burden of diabetes in adults: projections from 2015 to 2030. Diabetes Care. 2018;41(5):963–70.
3. American Diabetes Association. Economic costs of diabetes in the US in 2012. Diabetes Care. 2013;36(4):1033–46.
4. Goettler A, Grosse A, Sonntag D: Productivity loss due to overweight and obesity: a systematic review of indirect costs. BMJ Open 2017;7(10):1–9.
5. Darbà J, Kaskens L, Detournay B, Kern W, Nicolucci A, Orozco-Beltrán D, de Arellano AR. Disability-adjusted life years lost due to diabetes in France, Italy, Germany, Spain, and the United Kingdom: a burden of illness study. Clinicoecon Outcomes Res. 2015;7:1–9.
6. Schofield DJ, Percival R, Passey ME, Shrestha RN, Callander EJ, Kelly SJ. The financial vulnerability of individuals with diabetes. Br J Diabetes Vasc Dis. 2010;10(6):300–4.
7. Association AD. Standards of medical care in diabetes—2010. Diabetes Care. 2010;33(Supplement 1):S11–61.
8. Adua E, Roberts P, Wang W. Incorporation of suboptimal health status as a potential risk assessment for type II diabetes mellitus: a case-control study in a Ghanaian population. EPMA J. 2017;8(4):345–55.
9. Adua E, Frimpong K, Li X, Wang W. Emerging issues in public health: a perspective on Ghana's healthcare expenditure, policies and outcomes. EPMA J. 2017;8(3):197–206.
10. Yan YX, Dong J, Liu YQ, Yang XH, Li M, Shia G, Wang W. Association of suboptimal health status and cardiovascular risk factors in urban Chinese workers. J Urban Health. 2012;89(2):329–38.
11. Lemke HU, Golubnitschaja O. Towards personal health care with model-guided medicine: long-term PPPM-related strategies and realisation opportunities within 'Horizon 2020.' EPMA J. 2014;5(1):8.

12. Suchkov, Sergey, Olga Golubnitschaja, Matt von Herrath, Paolo Pozzilli, Mihail Paltsev, Ashot Mkrtumyan, Martin Frank, Trevor Marshall, and Harry Schroeder. "Predictive, preventive and personalized medicine (PPPM) as a strategic avenue and global tool for advancing T1D-related care: Fundamental, Applied and Affiliated Issues." In EPMA J. BioMed Central. 2014;5(1):1–3.
13. Golubnitschaja O, Costigliola V. European strategies in predictive, preventive and personalised medicine: highlights of the EPMA World Congress 2011. EPMA J. 2011; 2(4):315–32.
14. Golubnitschaja O, Kinkorova J, Costigliola V. Predictive, preventive and personalised medicine as the hardcore of 'Horizon 2020': EPMA position paper. EPMA J. 2014;5(1):6.
15. Anto EO, Roberts P, Coall D, Turpin CA, Adua E, Wang Y, Wang W. Integration of suboptimal health status evaluation as a criterion for prediction of preeclampsia is strongly recommended for healthcare management in pregnancy: a prospective cohort study in a Ghanaian population. EPMA J. 2019;10(3):211–26.
16. Zoungas S, Woodward M, Li Q, Cooper ME, Hamet P, Harrap S, Heller S, Marre M, Patel A, Poulter N. Impact of age, age at diagnosis and duration of diabetes on the risk of macrovascular and microvascular complications and death in type 2 diabetes. Diabetologia. 2014;57(12):2465–74.
17. Venables MC, Jeukendrup AE. Physical inactivity and obesity: links with insulin resistance and type 2 diabetes mellitus. Diabetes Metab Res Rev. 2009;25(S1):S18–23.
18. Slingerland L, Fazilova V, Plantinga E, Kooistra H, Beynen A. Indoor confinement and physical inactivity rather than the proportion of dry food are risk factors in the development of feline type 2 diabetes mellitus. Vet J. 2009;179(2):247–53.
19. DeFronzo RA, Ferrannini E, Groop L, Henry RR, Herman WH, Holst JJ, Hu FB, Kahn CR, Raz I, Shulman GI. Type 2 diabetes mellitus. Nat Rev Dis Prim. 2015;1(1):1–22.
20. Dipnall JF, Pasco JA, Meyer D, Berk M, Williams LJ, Dodd S, Jacka FN. The association between dietary patterns, diabetes and depression. J Affect Disord. 2015;174:215–24.
21. Nilsen V, Bakke PS, Gallefoss F. Effects of lifestyle intervention in persons at risk for type 2 diabetes mellitus-results from a randomised, controlled trial. BMC Public Health. 2011;11(1):893.
22. Lindström J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes Care. 2003;26(3):725–31.
23. Mullican DR, Lorenzo C, Haffner SM. Is prehypertension a risk factor for the development of type 2 diabetes? Diabetes Care. 2009;32(10):1870–2.
24. Ferrannini E, Cushman WC. Diabetes and hypertension: the bad companions. The Lancet. 2012;380(9841):601–10.
25. Klein BE, Klein R, Lee KE. Components of the metabolic syndrome and risk of cardiovascular disease and diabetes in Beaver Dam. Diabetes Care. 2002;25(10):1790–4.
26. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham Study. Am J Cardiol. 1976;38(1):46–51.
27. Conroy RM, Pyörälä K. Fitzgerald Ae, Sans S, Menotti A, De Backer G, De Bacquer D, Ducimetiere P, Jousilahti P, Keil U: Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J. 2003;24(11):987–1003.
28. Dimopoulos AC, Nikolaidou M, Caballero FF, Engchuan W, Sanchez-Niubo A, Arndt H, Ayuso-Mateos JL, Haro JM, Chatterji S, Georgousopoulou EN. Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. BMC Med Res Methodol. 2018;18(1):179.
29. Adua E, Roberts P, Sakyi SA, Yeboah FA, Dompreh A, Frimpong K, Anto EO, Wang W. Profiling of cardio-metabolic risk factors and medication utilisation among type II diabetes patients in Ghana: a prospective cohort study. Clin Transl Med. 2017;6(1):32.
30. Wang Y, Liu X, Qiu J, Wang H, Liu D, Zhao Z, Song M, Song Q, Wang X, Zhou Y. Association between ideal cardiovascular health metrics and suboptimal health status in Chinese population. Sci Rep. 2017;7(1):1–6.
31. 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. Sci Rep. 2020;10(1):1–10.
32. Lee BJ, Ku B, Nam J, Pham DD, Kim JY. Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes. IEEE J Biomedi Health Inform. 2013;18(2):555–61.
33. Lai H, Huang H, Keshavjee K, Guergachi A, Gao X. Predictive models for diabetes mellitus using machine learning techniques. BMC Endocrine Disord. 2019;19(1):1–9.
34. Xie Z, Nikolayeva O, Luo J, Li D: Peer Reviewed: Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. Prev Chronic Dis. 2019;16(1):1–9
35. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Making. 2010;10(1):16.
36. Kuhn M, Johnson K. Applied predictive modeling. 1st Edition. Vol. 26. New York: Springer-Verlag; 2013.
37. Mair C, Kadoda G, Lefley M, Phalp K, Schofield C, Shepperd M, Webster S. An investigation of machine learning based prediction systems. J Syst Software. 2000;53(1):23–9.
38. Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. New Engl J Med. 2017;376(26):2507.
39. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Making. 2019;19(1):211.
40. Harutyunyan H, Khachatrian H, Kale DC, VerSteeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Sci Data. 2019;6(1):1–18.
41. Kamiński B, Jakubczyk M, Szufel P. A framework for sensitivity analysis of decision trees. Central Eur J Operations Res. 2018;26(1):135–59.
42. Mani S, Chen Y, Elasy T, Clayton W, Denny J: Type 2 diabetes risk forecasting from EMR data using machine learning. In: AMIA annual symposium proceedings: AMIA Annu Symp Proc. 2012;606–15.
43. Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia Comput Sci. 2018;132:1578–85.
44. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Front Genetics. 2018;9:515.
45. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data. 2015;3(4):277–87.
46. Sneha N, Gangil T. Analysis of diabetes mellitus for early prediction using optimal features selection. J Big Data. 2019;6(1):13.
47. Kolog EA, Montero CS, Toivonen T. Using Machine Learning for Sentiment and Social Influence Analysis in Text. In: Rocha Á, Guarda T. (eds) Proceedings of the International Conference on Information Technology & Systems (ICITS 2018). ICITS 2018. Advances in Intelligent Systems and Computing, 2018: Vol 721.
48. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc Ser B (Methodological). 1977;39(1):1–22.
49. Brownlee J. Machine learning mastery with python. Machine Learning Mastery. 2nd Edition. Pty Ltd.; 2016. p. 100–20.
50. Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learn. 1997;29(2–3):103–30.
51. Chiu MH, Yu YR, Liaw HL, Chun-Hao L. The use of facial micro-expression state and Tree-Forest Model for predicting conceptual-conflict based conceptual change. Chapter Title & Authors Page 2016, 184.
52. Pisner DA, Schnyer DM: Support vector machine. In: Machine Learning. Elsevier; 2020. p. 101–21.
53. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat. 1992;46(3):175–85.
54. Dybowski R, Gant V, Weller P, Chang R. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. Lancet. 1996;347(9009):1146–50.
55. Gillery P. A history of HbA1c through clinical chemistry and laboratory medicine. Clin Chem Lab Med. 2013;51(1):65–74.
56. Bennett C, Guo M, Dharmage S. HbA1c as a screening tool for detection of type 2 diabetes: a systematic review. Diabet Med. 2007;24(4):333–43.
57. Mahadevan S, Ali I. Is body mass index a good indicator of obesity? Int. J. Diabetes Dev. Ctries. 2016;36(1):140–2.
58. Kok P, Seidell J, Meinders A. The value and limitations of the body mass index (BMI) in the assessment of the health risks of overweight and obesity. Ned Tijdschr Geneeskd. 2004;148(48):2379–82.
59. Tomiyama AJ, Hunger JM, Nguyen-Cuu J, Wells C. Misclassification of cardiometabolic health when using body mass index categories in NHANES 2005–2012. Int J Obesity. 2016;40(5):883–6.

60. Bhurosy T, Jeewon R. Pitfalls of using body mass index (BMI) in assessment of obesity risk. Curr Res Nutr Food Sci J. 2013;1(1):71–6.
61. Freedman DS, Sherry B. The validity of BMI as an indicator of body fatness and risk among children. Pediatrics. 2009;124(Supplement 1):S23–34.
62. Kirk S, Cramm CL, Price SL, Penney TL, Jarvie L, Power H. BMI: a vital sign for patients and health professionals. Can Nurse. 2009;105(1):25–8.
63. Bouchard C. BMI, fat mass, abdominal adiposity and visceral fat: where is the 'beef'? Int J Obesity. 2007;31(10):1552–3.
64. Sarwar A, Ali M, Manhas J, Sharma V. Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. Int J Inf Technol. 2020;12(2):419–28.
65. Akella AB, Kaushik V. Machine Learning Algorithms for Predicting Coronary Artery Disease: Efforts Toward an Open Source Solution. Future science. 2020; 7(6):1–10.
66. Vergès B. Lipid modification in type 2 diabetes: the role of LDL and HDL. Fundamental Clin Pharmacol. 2009;23(6):681–5.
67. Duvillard L, Florentin E, Lizard G, Petit J-M, Galland F, Monier S, Gambert P, Vergès B. Cell surface expression of LDL receptor is decreased in type 2 diabetic patients and is normalized by insulin therapy. Diabetes Care. 2003;26(5):1540–4.
68. Becker J, Nora DB, Gomes I, Stringari FF, Seitensus R, Panosso JS, Ehlers JAC. An evaluation of gender, obesity, age and diabetes mellitus as risk factors for carpal tunnel syndrome. Clin Neurophysiol. 2002;113(9):1429–34.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.