Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-16407-5

Diabetes detection based on machine learning and deep learning approaches

Boon Feng Wee[1] · Saaveethya Sivakumar[1] · King Hann Lim[1] · W. K. Wong[1] · Filbert H. Juwono[2]

Received: 20 November 2022 / Revised: 19 May 2023 / Accepted: 21 July 2023

Abstract
The increasing number of diabetes individuals in the globe has alarmed the medical sector to seek alternatives to improve their medical technologies. Machine learning and deep learning approaches are active research in developing intelligent and efficient diabetes detection systems. This study profoundly investigates and discusses the impacts of the latest machine learning and deep learning approaches in diabetes identification/classifications. It is observed that diabetes data are limited in availability. Available databases comprise lab-based and invasive test measurements. Investigating anthropometric measurements and non-invasive tests must be performed to create a cost-effective yet high-performance solution. Several findings showed the possibility of reconstructing the detection models based on anthropometric measurements and non-invasive medical indicators. This study investigated the consequences of oversampling techniques and data dimensionality reduction through feature selection approaches. The future direction is highlighted in the research of feature selection approaches to improve the accuracy and reliability of diabetes identifications.

Saaveethya Sivakumar, King Hann Lim, W.K Wong and Filbert H. Juwono contributed equally to this work.

✉ Boon Feng Wee
19966010@student.curtin.edu.au

Saaveethya Sivakumar
saaveethya.s@curtin.edu.my

King Hann Lim
glkhann@curtin.edu.my

W. K. Wong
weikitt.w@curtin.edu.my

Filbert H. Juwono
filbert@ieee.org

[1] Department of Electrical and Computer Engineering, Curtin University, Kent St, Bentley, West Australia 6102, Australia

[2] Department of Electrical and Electronic Engineering, Xi'an Jiaotong, Liverpool University, 111 Ren'ai Road, Suzhou, Jiangsu Province 215123, P. R. China

Published online: 10 August 2023

Keywords Diabetes detection · Machine learning · Deep learning · Feature selection · Anthropometric measurement

1 Introduction

Diabetes mellitus is a chronic metabolic disease affecting the human body by converting blood sugar into energy. People who diagnosed with diabetes cannot regulate their blood sugar levels in the bodies, resulting in high blood sugar levels and blood pressure. Suppose diabetes is not detected, diagnosed, and treated adequately at early stage, it could lead to other life-threatening diseases such as diabetic retinopathy and neuropathy, kidney failure, and other cardiovascular diseases [33]. Even with significant advances in the medical field over the past century, diabetes is still becoming more common in most societies. Its prevalence is rising in all countries, regardless of the people's income level. According to research on diabetes prevalence projection for 2030 and 2045 conducted [50], it is estimated that people diagnosed with diabetes are expected to rise to 10.2% of the worldwide adult population (578 million) by 2030 and continue to rise to 10.9% (700 million) by 2045. Hence, it is vital to develop intelligent systems that aid medical personnel in diabetes diagnosis and decision support. Traditional lab-tests based diabetic detection methods are mostly time-consuming and expensive. In general, clinicians take approximate prediction and diagnosis of diabetes mellitus of patiets by taking oral glucose tolerance, fasting blood sugar, or random blood sugar tests [25]. In order to perform further confirmation on diagnosis of diabetes in the patients, Glycated hemoglobin (A1C) test was introduced and implemented to the public in 1980 [25], where this test analyzes percentage of blood sugar attached to the haemoglobin for three months [25]. The procedures of this test are complicated, time-consuming and require medical professionals and specific equipment to be at the scene to perform the lab tests, needless to mention that if these are not available at the scene, transportation and storage of blood samples etc. require another amount of expenses [35].

Type 2 diabetes mellitus (T2DM) is the most common diabetes category and it is characterized as hyperglycemia due to insulin resistance or insufficient insulin production in the body [8]. Various factors such as lack of physical activity, imbalance diet, tobacco use, excessive alcohol intake, and obesity are suggested to be contributed to the manifestation of T2DM [28, 36]. T2DM also contributes to the risk for cardiovascular diseases such as coronary heart disease, stroke, peripheral arterial disease, and aortic diseases; all lead by high blood pressure as the consequences of infecting diabetes mellitus [28]. In addition, T2DM can also be diagnosed genetically as insulin resistance and insulin secretion are genetically linked to hyperglycemia [18].

Many existing data-driven diabetes detection models are currently constructed and studied by other researchers [9, 64, 71]. However, most researchers took datasets recorded with lab-test-based medical indicators for the models' training and validation process [20, 44, 51]. However, these prediction models are considered excessive as lab-test-based measurements can already be used to tell and diagnose if a person has diabetes or not with promising accuracy [49]. A more preliminary diagnosis solution without any lab test measurement is more demanded. Therefore, research on anthropometric measurements' features and their impact on diabetes detection models should be performed. In all data-driven diabetes detection classification solutions, machine learning is one of the most popular options due to its classification function using statistical approaches, without requirement of high computation power [10]. Although the relationship between possible risk factors in causing diabetes is

certainly nonlinear [68] and machine learning algorithms are more suitable to classify linear functions, this issue can be solved by augmenting various kernel functions to the machine learning models in predicting nonlinear functions using statistical approaches [10].

Another popular option for this task is by utilizing deep learning approaches. With high computing power, Deep Neural Network (DNN) can perform classification with minimal manual engineering optimization required while achieving satisfying results. Moreover, DNN can solve any function with nonlinear variables compared with machine learning models that required to be paired with certain algorithms and functions to achieve this [10]. Hence, its robust and less complicated learning functions made it one of the best options in solving this classification problem. Hence, to construct a model that is simple to be implemented and able to analyze attributes of all extracted features from a dataset, a hybrid model combining both machine learning and deep learning algorithms is suggested.

The reason that this particular diabetes detection task is selected because diabetes mellitus is the prime factor that leads to other life-threatening cardiovascular diseases. Due to the advancements made in data science domain in the past decade, classification of diabetes using machine learning and deep learning approaches have become more and more possible. Therefore, a pre-diabetes classification tool, trained with non-invasive lab measurements data is demanded. This is because development of such tool can greatly decrease the burden of current healthcare system from every aspect. On top of that, possibility in analyzing causes of diabetes mellitus using machine learning and deep learning methods should also be investigated.

This review paper is essential in finding possibilities of constructing a data-driven diabetes classification model, trained with datasets recorded with anthropometric or non-lab-invasive medical measurements. Multiple machine learning and deep learning approaches are investigated in order to observe their respective advantages and disadvantages in solving this diabetes classification task. The approaches conclude every aspect including selection of datasets, methodology in imputing missing data, feature selection, sampling, and most importantly, classification algorithms used in performing the task.

This comprehensive review is separated into seven sections: Section I: Introduction for this review paper; Section II: Data Collection concluded what data are collected and extracted for the models' training and validation; Section III: Data Pre-processing summarized techniques implemented by researchers in improving overall quality of extracted dataset; Section IV: Feature Selection summarized feature selection algorithms implemented by researchers in selecting relevant variables in the extracted dataset; Section V: Machine Learning and Deep Learning Models summarized models constructed by researchers and findings in their works; Section VI: Comparisons concluded the performances of machine learning and deep learning-based models side by side; Section VII: Future Works and Conclusion summarized findings in this review paper.

2 Data collection

Since classification of diabetes by machine learning and deep learning approaches are highly relied on the datasets implemented, selecting an appropriate dataset has then become one of the most critical processes in training the model [12]. In recent studies, most existing data-driven diabetes detection models are trained using a publicly available diabetes dataset for machine learning and deep learning purposes, named Pima Indians Diabetes Database (PIDD) [14]. Released in 1988, this dataset records nine features of 768 female instances

aged at least 21 years old. Recorded features are as follows: Age, Body-Mass-Index (BMI), Diabetes Pedigree Function (DPF), number of pregnancies, plasma glucose in an oral glucose tolerance test, blood pressure, triceps skin fold thickness, 2-hour serum insulin, and presence of diabetes in sample's body.

In this PIDD dataset: age, BMI, number of pregnancies, diabetes pedigree function, and triceps skin fold thickness do not require complex lab equipment to perform the measurement. However, for the other two features in the dataset (Blood Glucose Levels and Insulin Dose Taken), they require certain machines or equipment to obtain the measurement. 2-hour serum insulin test and oral glucose tolerance test (OGTT) are performed by first taking blood samples of the subjects, the subjects are then requested to drink a certain amount of glucose solution. The subjects' blood samples are then taken again for every 30 to 60 minutes.

Other than Blood Glucose Level and Insulin Dose taken, it is worth to mention difficulties in performing diabetes pedigree function measurement. Based on original description of the PIDD dataset, this feature is obtained by calculating the score of likelihood of diabetes based on family history [56]. First, to perform this calculation, data of the subjects' family members are required, and this procedure is time-consuming. Secondly, based on different diabetes pedigree algorithms, outputs of diabetes pedigree function measurement would not be the same, i.e. do not have a common universal standard.

PIDD consists of adequate number of recorded instances and can be readily inserted into any learning model, without excessive data pre-processing. Therefore, PIDD is widely applied in machine learning and deep learning methods for Diabetes Detection. Along with this popular PIDD database, other databases are also used in training constructed detection models. Quan et al. [71] introduced prediction model using Luzhou, China database. This database recorded 14 medical examination features of 68994 healthy and diabetic samples respectively (total no. of samples = 137,998) [71]. These trained models are then taken to test and validated using another dataset recorded with 13,700 samples and the same features. Quan el al. also used the PIDD dataset to validate the constructed detection models. Recorded 14 features in the Luzhou, China database are as follows: Age, pulse rate, breathing, left systolic pressure (LSP), right systolic pressure (RSP), left diastolic pressure (LDP), right diastolic pressure (RDP), height, weight, physique index, fasting glucose, waistline, low-density lipoprotein (LDL), and high-density lipoprotein (HDL). For this dataset, measurements of systolic and diastolic pressure require sphygmomanometer [59], and both high and low-density lipoprotein require blood test called as lipoprotein panel before the analysis can be made.

It is worth investigating the performance of classification models when trained on datasets that do not include invasive lab measurements, as both the PID and Luzhou, China datasets rely on these measurements. This study direction is crucial due to advancements in the medical field, where accurate classification of diabetes outcomes can already be achieved using conventional methods and datasets stated earlier. In contrast to these datasets, there is a current demand for a diabetes classification tool that is more cost-effective and time-efficient, and it is undeniable that datasets play a vital role in addressing this need. By implementing non-invasive datasets with machine learning and deep learning approaches, there is potential for developing the desired classification tool. Therefore, several studies have focused their investigations on the utilization of non-invasive datasets rather than implementing conventional invasive datasets.

In Swapna et al. [58] research, they used electrocardiograms of 40 people collected for 10 minutes while lying down in a relaxed position, where 20 of them had diabetes, and the remaining 20 people were healthy [58], while Sandhiya et al. [51] used dataset obtained

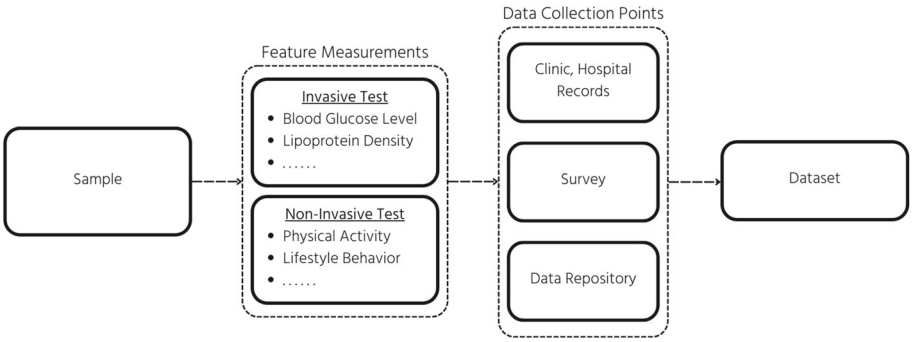

Fig. 1 General representation of data collection

from University of California, School of Information and Computer Science (UCI) Machine Learning Repository. This diabetes dataset recorded both lab and non-lab-tests-based features such as insulin doses taken, blood glucose levels measurements before and after meal intake, amount of meal ingestions, and amount of exercise activity taken by samples [51].

Anna et al. [9] trained their diabetes prediction models using dataset that records non-lab-test-based medical indicators of 451 children ages 6 and 18. Recorded features are as follows: age, sex, weight, height, presence of Type 1 Diabetes, and physical exercise records such as step counts, sedentary activity duration, and also light, moderate, and vigorous activity duration per week [9]. The physical activity duration and intensity measurements are performed by equipping pedometers and accelerometers on the samples. Intention of using this dataset in the research is to solve several problems faced by the children when taking lab-test-based diabetes detection tests, such as being afraid of needles when taking medical tests. Hence, their goal is to study the reliability of diabetes detection models when they are trained with non-lab-test-based datasets.

Vidhya et al. [62] extracted training data from UCI Diabetic Repository. This dataset records related lifestyle behaviors and medical indicators of samples including: Presence of

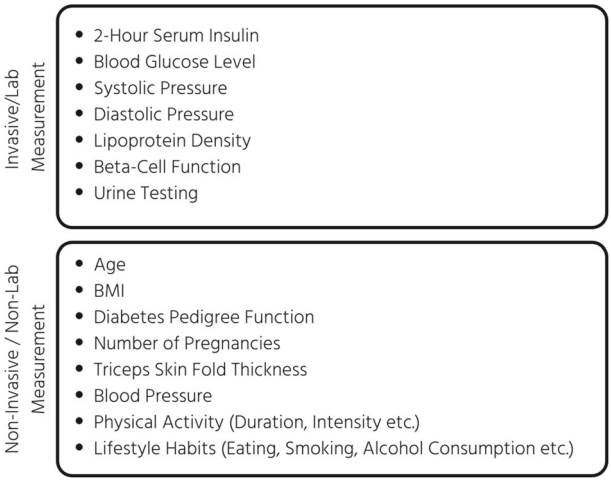

Fig. 2 Popular features extracted for machine learning / deep learning-based detection model

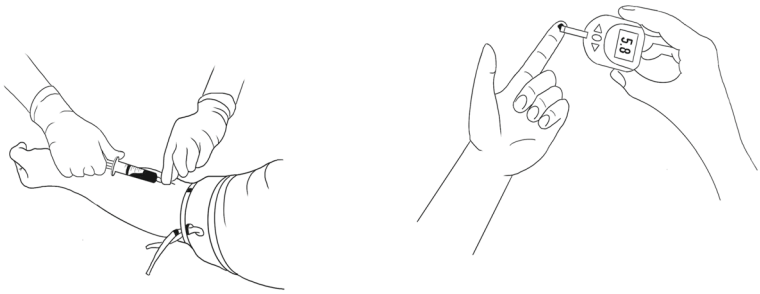

Fig. 3 Examples of invasive measurements in diabetes classification task: blood extraction (left) and fingerstick glucose monitoring (right)

diabetes in family history, BMI, food habit, conduction of regular exercises, age, consumption of alcohol, late-night binging, eating habit, average duration in seated position per day, blood glucose level, presence of smoking habit, job nature, and gender.

Rani et al. [45] used a dataset that recorded multiple medical indicators as follows: age, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity and presence of diabetes in samples [45]. These data are recorded as an analog value, representing YES or NO, instead of taking exact numerical medical test values. It is worth to mention that these indicators can be observed by investigating daily life routines or physical appearances without taking any lab tests, similar to Anna et al. [9] research.

As a brief summary, in most existing diabetes detection models, features fed to the models are invasive and require expensive machines, plus the data collection procedures are tedious and complex. At the same time, limited research has proven simple physical features or lifestyle investigations such as physical activity duration and existences of smoking behaviour can also be used in constructing diabetes detection model [9, 62], but their reliabilities are yet to be determined. Figure 1 and Fig. 2 categorized several common and popular features used in existing research in this field. Figures 3 and 4 showed several common measurements in diabetes classification task. It can be observed that for most invasive measurement, it involves in extracting the patient's blood, where the patients are required to perform fasting before the measurement.

Figure 5 shows a pie chart displaying the proportion of datasets used in all reviewed papers. It can be observed that over 101 classification models constructed in all reviewed papers, only 12 of them are trained with non-invasive datasets. This figure further verifies that most people

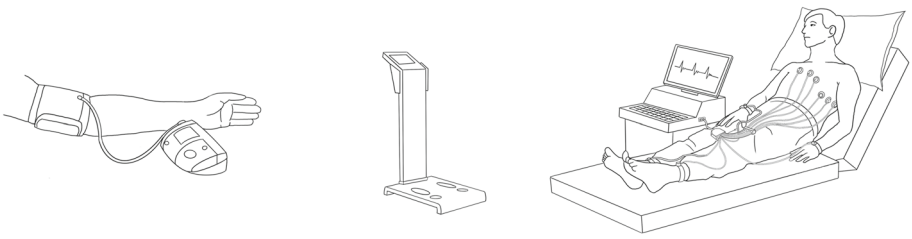

Fig. 4 Examples of non-invasive measurements in diabetes classification task: sphygmomanometer for blood pressure measurement (left), kern platform scales for body mass measurement (middle), multi-channels ECG machine for electrocardiogram (right)

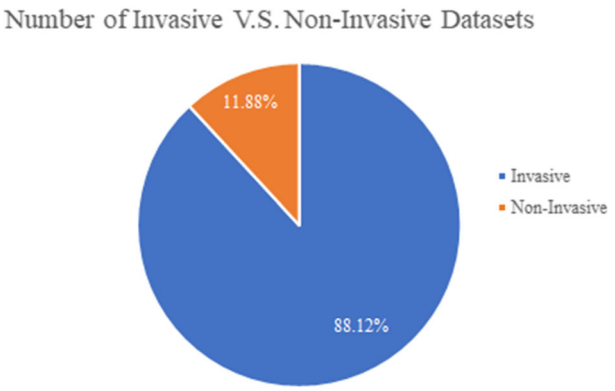

Fig. 5 Proportion of datasets in terms of invasive or non-invasive

has ignored the possibility in constructing a classification model using non-invasive datasets, and this should not be the case as it can bring more advantages than conventional methods such as cost, accessibility, patient comfort, safety, and easier application in real-world, if it is successfully invented and verified.

Machine Learning and Deep Learning models heavily rely on datasets for their training and performance. However, obtaining large datasets can be challenging due to the significant time and cost involved. While using non-invasive datasets can help to some extent, it is not the perfect solution. The reason is that before collecting data to build the dataset, it is essential to first conduct an analysis of the correlation between the target feature and diabetic outcome. Blindly collecting unrelated data would only increase the cost and time required to develop an effective data-driven diabetes classification model. Therefore, careful consideration and analysis are necessary to ensure the relevance and usefulness of the collected data. Nonetheless, performance of classification models based on such datasets are necessary to be studied.

Despite of that, many existing datasets available online has certain flaws and disadvantages such as having missing data or very few recorded samples. To solve these issues, researchers applied various pre-processing techniques. Figure 6 displays the generalized flow of construction of a classification model, it can be seen that model's performance is primarily affected by the quality of dataset itself. Nature of the dataset has to be determined carefully before data collection to avoid unnecessary cost and performance dragging issue.

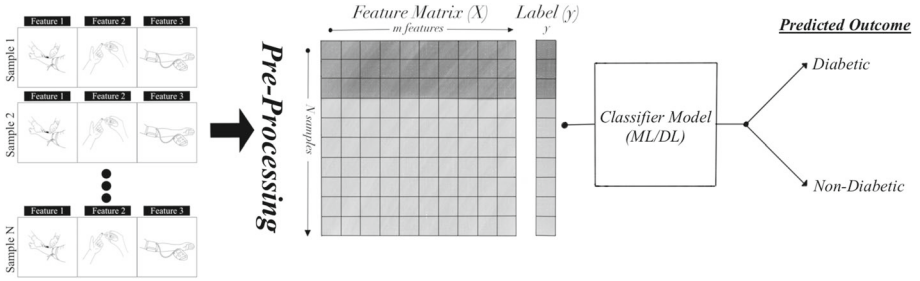

Fig. 6 Generalized flow of classification model construction

3 Data pre-processing

Despite there are numerous kinds of datasets implemented in solving this task, it is still undeniable that most of them does not meet the quality constraint in training the machine learning and deep learning models, due to reasons as follows: First, it is unavoidable that many datasets contain missing or wrongly data when being constructed. Existence of such data points may affect the models' performance to some extent, and this must be avoided especially when dealing with medical or healthcare related tasks. Second, features recorded in the dataset may not be correlated to the target diabetic outcome, involving these features to train the classification models will not only drag the models' performance, but also increase the computational cost and time required. Other than these, various scales, units and distributions may be different in datasets, and this may cause domination of certain feature in the learning process, leading to incorrect and unfair comparisons between different features. Class balancing issue is also a concerning task in constructing the perfect dataset, as it can prevent biasing of model towards majority class. Therefore, before feeding the dataset to train the model, data pre-processing procedures such as data imputation, feature selection, data normalization and class balancing has to be performed accordingly to solve the stated issues. On top of that, depending on nature of the dataset, encoding of data has to be performed in order to allow the model in processing and understanding the categorical information effectively.

3.1 Data denoise and imputation

Despite the fact that PIDD [14] has many advantages, there are few concerning issues especially using them for machine learning or deep learning applications. PIDD dataset records only female samples, together with the number of pregnancies of that individual. As a result, the models trained with PIDD dataset might not be applicable on males. PIDD dataset has a large number of missing/abnormal recorded data that requires data filtering to avoid biased and inaccurate prediction. Nevertheless, if these abnormal data are eliminated, the dataset may result in a smaller number of samples and it may inadequately represent the actual distribution of data in big population. Hence, model will not generate a good result when released to the public.

As stated in [70], they pointed out that a major limitation of PIDD is its low number of samples and features recorded. Due to this issue, the generalization of machine learning and deep learning models is questionable. Moreover, the authors also pointed out that data variability, data quality, feature processing, and result interpretability are issues that need to be solved in constructing diabetes detection models. As such, both data pre-processing and feature selection processes are vital in constructing a prediction model that able to achieve accurate performance.

As mentioned earlier, missing or abnormal data exists in PIDD dataset. There are total of 30% missing data in the Triceps Thickness Fold class, and Insulin Dose Class has 49%. These data are recorded in the form of zero. Occupying such big portion of the total dataset, this matter should not be ignored, as this would affect the classification accuracy.

In eliminating the missing data [1, 17, 29, 47, 52], filled the missing value with mean value of respective class. Such approach is popular as this could ensure continuity of the dataset. However, these filled data could not represent the actual distribution, hence classification will be inaccurate when new samples outside the dataset are taken into the model for classification.

Another popular approach in solving this issue is by simply eliminating samples recorded with missing data [29, 44, 71]. As explained earlier, this action will further decrease number of available samples in the dataset. Boundary between diabetic and non-diabetic output will be inaccurate if the models are trained with such dataset.

Santosh et al. [29] performed their research on studying how different approach in treating PIDD dataset would affect the model's performance. In this research, all classes in the dataset are taken to train the models. After that, authors used three different approaches in handling missing data: Removing the samples, replace the missing data with mean values, and replace it with zero. As result, in the testing process, removing samples with missing data achieved the best accuracy among all three methods, followed by replace as mean value and finally replace as zero. Result implies that replacing missing data with mean value will slightly overcome the bias issue brought by missing data. Although removing samples with missing data achieved the best accuracy in this research, it is still unclear how the models will perform when new datasets are introduced to the models.

As a brief conclusion, to eliminate missing data issue in datasets, researchers have performed both data imputation and missing data removal. While it seems that both methods possess certain advantages over the other, at the end the selected solution is still dependent on the nature of the dataset: If a small dataset such as PIDD is selected, data imputation is recommended as further data elimination will only decrease number of available data points in it, especially when dealing with deep learning-based classification models where number of training data is crucial for effective training; On the other hand, if a large dataset is selected and when number of data is no longer a concern, it is best to remove small noises that exist in it in order to maintain the original data distribution. However, despite stated that removing noises from small dataset will lead to ineffective training due to limited number of data, oversampling the dataset will come in handy in solving this issue.

To increase number of available classes and samples, Maria et al. [17] used a technique known as Data Augmentation in improving the PIDD dataset. This technique's main purpose is to perform oversampling, i.e., increasing number of elements in datasets, enabling the detection models to have more samples for training. By implementing Sparse Autoen-

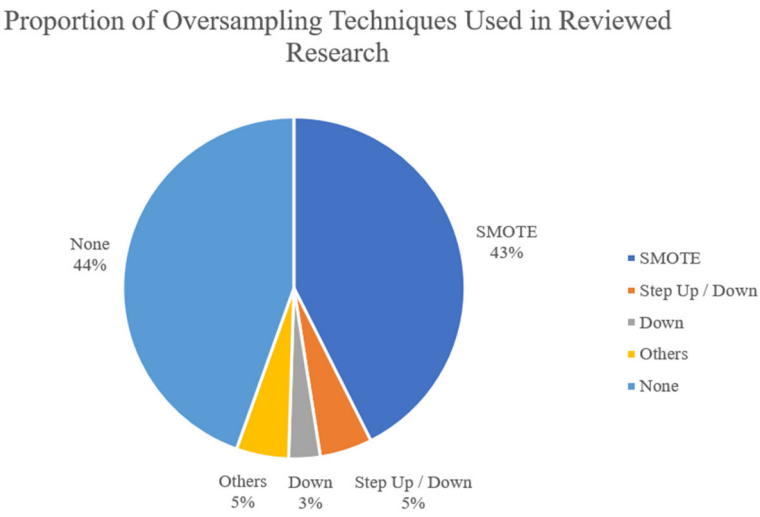

Fig. 7 Proportion of oversampling techniques implemented in reviewed papers

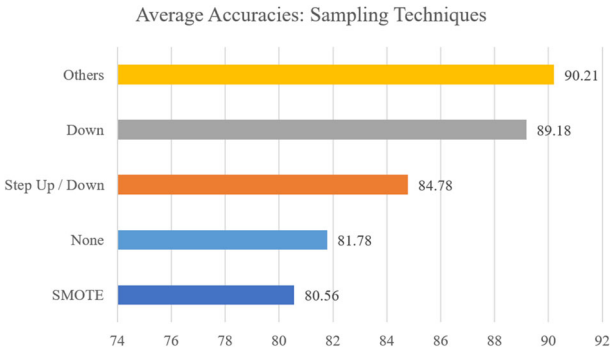

Fig. 8 Average accuracies oversampling techniques implemented in reviewed papers

coder (SAE), original PIDD with 8 features (excluding presence of diabetes in samples) are increased to 400, and transformed into a 20 by 20 matrix. Variational Autoencoder (VAR) is then implemented to increase number of samples to 449 class 0 (non-diabetic) and 484 class 1 (diabetic) respectively, solving the issues that the dataset has too less samples and also the samples are unbalanced simultaneously. As a whole, issues such as missing data or outliers in the dataset should be handled before dataset is fed to the detection models as it is essential in pursuing a flawless prediction model.

As displayed in Figs. 7 and 8, blindly oversampling the datasets will only cause negative impact on the models's performance. Comparing proportion and average accuracies achieved by models implemented with no oversampling technique and SMOTE, they both have similar proportion, where 44% of the models did not implement any oversampling technique and 43% are implemented with SMOTE. Models implemented with SMOTE achieved a lower average accuracy in solving diabetes classification task despite having larger number of training samples. There are many factors that lead to this result, whereas the most critical point is the quality of the dataset. It is stated so as most researcher did not perform any data

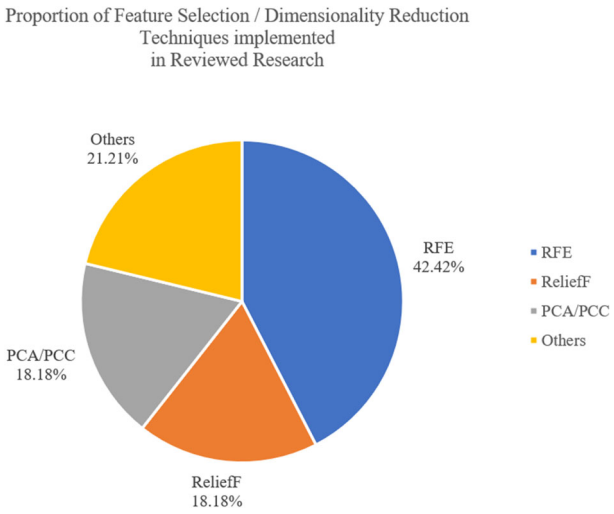

Fig. 9 Proportion of feature selection techniques implemented in reviewed papers

pre-processing before implementing SMOTE oversampling, causing the SMOTE algorithm continues to generate incorrect datapoints due to noises-filled dataset, hence leading the model to incorrect training. As a whole, issues such as missing data or outliers in the dataset should be handled before dataset is fed to the detection models as it is essential in pursuing a flawless prediction model (Fig. 9).

On top of that, several researchers prefer using machine learning-based algorithm in oversampling their selected datasets. Also shown in Fig. 10, categorized as “Others", these models achieved the best average accuracy of 90.21% as this type of oversampling algorithm can generate a more generalized and accurate datapoints. While using such algorithm can definitely increase the required computational time and cost, feature selection can aid in mitigating this issue.

3.2 Data normalization

In every medical database, values of different recorded classes fluctuate greatly. Suppose two classes with huge differences in data range are fed, the machine will consider the feature with higher values has more significant contribution to the output. Not only will it affect the model’s accuracy, but also make the correlation between recorded features in the database difficult to interpret and analyse. For convenience, many researchers performed min-max normalization method in scaling every class in the dataset. Min-max normalisation is one of the most common method in normalising data, where its function is to transform all features’ to have minimum value of 0 and maximum of 1. Min-max normalization method has a major problem, such that it will take rare extreme cases, or statistically known as outliers into consideration. Equation (1) shows mathematical expression of min-max normalization.

$$x_{scaled} = \frac{x - x_{max}}{x_{max} - x_{min}} \tag{1}$$

To eliminate any possible outlier in the dataset, study by Nadeem et al. [39] suggested Z-score normalization method is suitable for this task. Although Z-score normalization is able to handle outliers in dataset, it cannot produce a output where all features have the same scale, unlike min-max normalization that ensures all features scale between 0 and 1. Equation (2) shows mathematical expression of Z-score normalization.

$$Z_{score} = \frac{x - \mu}{\sigma} \tag{2}$$

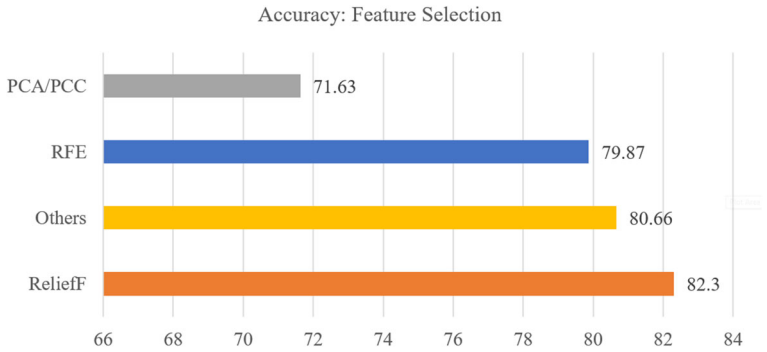

Fig. 10 Average accuracies of feature selection techniques implemented in reviewed papers

The issue of high dimensional features with redundant component is still existed in the dataset although the data is normalized. In order to decrease both computational time and power, it is essential to select important features according to function, which will be discussed in next section.

4 Feature selection and dimension reduction approaches

Although diabetes is found related to destruction of Beta-cells and elevation of TNF-alpha expression in human body[37], the leading medical indicators to these defects are still uncertain [23]. Diabetes is often associated with the following features: Hyperglycemia, obesity, abnormal systolic blood pressure, presence of diabetes in family history, defect in immune system and genes, dyslipidemia, liver and kidney dysfunctions, and various other factors [66]. These may include other physical and clinical data in which when fully incorporated may yield high dimensionality in feature implementation.

As with all machine learning application domains, there is a need to prune features such that model description is as compact as possible. This is in line with "Ocham's Razor" principle in computation which states that in the case presented with 2 models of varying complexity, the simpler model and description should be selected. This may refer to machine learning model complexity but primarily refer to the features inclusion in model. In another words, selection of features such that features are minimal (hence optimal model description). This is even more evident in the case of diabetic detection where features included are often in higher order. Another related approach is feature reduction. This method, as the name suggest, applies principles to select projected features/transform feature prior to application to machine learning models. Generally, 2 approaches are common in machine learning application: feature selection (which seeks to select useful and discriminative features) and dimensionality reduction (augmentation of features by projection methods). However, years of implementation shows that these yields varying degree of improvements when implemented.

In both cases, selecting highly relevant features/feature projection can minimize the computational power required and decrease computation time for the detection process. Therefore, it is vital to compute each feature's correlation and relative importance using specific algorithms. In this context, the question arises: what are the optimal feature selection algorithm/optimal feature reduction approaches applied in this domain?

In this section, two common feature selection are discussed namely PCA (Principal Component Analysis) and PCC (Pearson Correlation Coefficient). Both PCA and PCC are commonly used in dimension reduction and feature selection respectively. PCA is a feature selection algorithm that is widely used in computer science domain, especially datasets with high dimensionality . In general, function of PCA is to relocate samples from its initial space, and denoting the samples with a lower-dimensional representation [57], while the denoted data still able to represent statistical characteristic such as covariance of the original data.

Common feature dimension reduction includes PCA and Linear discriminant analysis (LDA). Both utilises features projection on new axis and subsequently selecting the projection that captures highest variances. PCA is widely used in image pre-processing and feature dimension reduction due to its robustness in finding discriminant projections. In Song et al's [57] research, applied PCA to reduce dimension of images from the dataset before implementing them to their proposed models. After feature selection is performed, dimensions of

the images are reduced to 40%-70% of the original image, while the results show that the processed images are able to achieve higher accuracy than unprocessed images.

Never the less, it is noteworthy that feature reduction/feature selection could have adverse effects and thus would require case by case considerations when applied. In such cases, train /test data need to be randomly separated. this is highlighted in the case study presented in Sivaranjani et al. [55], authors shows that PCA feature selection method yielded slight improvements to the SVM-based model. In the RF-based model, its performance decreased after dimensionality reduction. Authors pointed out that this could be caused by reduction in dimensionality further decreased the number of available data in the PIDD dataset, where it is considered that this dataset thus limiting generalisation. Similar to their research result, Francesco et al. [38] works showed the same trend in performances of their constructed machine learning models, where PCA algorithm only brings minimal improvements to most of the models, and even decreased the performance of RF-based model after PCA is introduced.

As a brief summary for this section, the average accuracy achieved by all models implemented with feature selection algorithms is 81.16%, with the average accuracies of 80.31% and 84.51% achieved by machine learning and deep learning models respectively. As it can be seen of two figures above, although both PCA / PCC and RFE feature selection algorithms take up to 60% of the whole proportion, these two algorithms achieved the lowest average accuracies among all others feature selection algorithms, where they achieved average accuracies of 71.63% and 79.87% respectively. With average accuracies of lower than 80%, it can be concluded that not all feature selection algorithms are suitable to be implemented in solving diabetes classification task, and may bring adverse effects on the models' performance if such algorithms are introduced. From Fig. 11, it perfectly displayed the positive impacts brought to machine learning models when appropriate feature selection algorithms are implemented. Note that three machine learning algorithms displayed are the top three popular algorithms in all reviewed research papers.

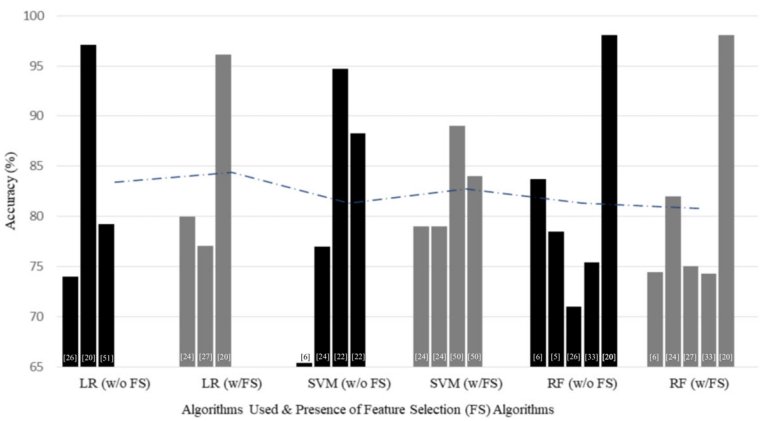

Fig. 11 Impact of feature selection on ml models

4.1 Feature selection

Most feature selection approaches utilise certain form of feature evaluation and scoring system. More rudimentary approaches include sequentially configuring subset of features and evaluation. The limitation to this is that features sets combination fitness are mostly non-linear. Therefore selecting sequentially using "greedy" approach may not yield much optimallity. More constructive approaches include evaluating feature similarities. One example is PCC [53], which is a popular algorithm used to compute correlation between two variables in one dataset. In many research of this domain, people often use this algorithm to choose the appropriate features that have strong correlation with the diabetes outcome in the samples. In Nour et al. [1] research, they computed all the correlations between every variable in the dataset to study their respective statistical relationship with each other.

Zhou et al. [69] applied a dataset that contains 141 indicators of 9765 samples, including 5 classes of diabetic complications. Their study focused on detecting samples with different diabetic complications, using the 141 medical indicators recorded in the dataset. It is impractical to extract all variables in the dataset and study their respective impact in detecting every diabetic complication. Hence, before feeding the data to the models, they implemented PCC and computed the correlation between each indicator with the diabetic complications. With different correlation values obtained by every indicator in every diabetic complication, for every complication, they chose top ten indicators that obtained the highest marks in the PCC algorithm . As result, all four models used in this research achieved good result in classifying every kind of diabetic complications, indicating that feature selection using PCC is feasible in this domain.

Lukmanto et al. [34] applied Fuzzy Support Vector Machine on the PIMA indian dataset and investigated the feature selection approach to gauge the improvements achieved. Authors highlighted that The results show a promising accuracy of 89.02% in predicting patients with DM by implementing feature selection strategies using F-Score Feature Selection. this method is somewhat related to PCC in which correlation scores between features were evaluated for elimination and feature reduction.

Hou et al. [21] implemented fisher score to evaluate dataset from 19 features. The values of accuracy, precision, sensitivity, F1 score, MCC and AUC were computed respectively. Results show that their method can be successfully used to select features for diabetes classifier and improve its performance, which will provide support for clinicians to quickly identify diabetes. In short, purpose of feature selection in data driven solution is not only to improve the model's performance, but is also essential in analysing impacts of the feature to the outcome class, i.e., in diabetes classification task, feature selection can provide insights to medical personnel in analyzing correlation between target feature and diabetes outcome in data science perspective, hence

As a comparison, Fig.11 shows effect of feature selection on accuracy of diabetes detection machine learning models (LR, SVM, RF), as compared with models with no feature selection performed. The data were collected from several research papers that incorporate feature selection in its approaches. As shown, Blue line indicates mean accuracies achieved by all the models before and after feature selections are performed, while the difference indicate the impact of feature selection. As it can be observed in most studies, most models with feature selection yielded improved results as compared to the models without features selection.

However, even it shows that positive feedback can be brought to the machine learning-based models when appropriate feature selection algorithms are implemented on the dataset, statistical data summarized in this literature review shows that even without feature selection, deep learning-based models can still perform better than feature selection implemented

machine learning-based models, as shown in Fig. 12. As a matter of fact, more evidences showed that after feature selection is deployed on the dataset, the models performed worse than it did before feature selection. While many cases showed that feature selection is crucial in optimizing the models' performance, most researchers did not realize that consequences of feature selection are always correlated to the number of available samples. When the original size of a dataset is already considered small at the beginning stage, further dimensionality reduction will only decrease number of available data, hence causing ineffective training of models. This can be directly verified from Fig. 12 that when feature selection is implemented with deep learning approaches, average accuracies achieved drastically becomes lower as a result of decreased available training data. Although studies showed that feature selection is important in decreasing computational burden, it is also important to mitigate the side effects by incorporating a larger dataset with better quality. While availability of good quality large datasets is always scarce, oversampling techniques as stated in last section are always suitable in mitigating side effects brought by feature selection itself.

From the literature review, it is noteworthy that feature selection / dimension reduction need to be performed with proper analysis despite the facts that many application domains claiming these procedures as standard procedures in machine learning. Are the feature dimension large enough to require feature selection? Are the data sufficient to provide generalisation? these are important considerations and thus the decision ultimately depends on the data acquired.

5 Machine learning and deep learning models

In most data driven diabetes classification task, machine learning and deep learning approaches are popular solutions. Before discussing their respective pros and cons, it is important to first understand how the models are evaluated in this task.

For binary classification purposes, a confusion matrix can be implemented in evaluating the performance of the model. The confusion matrix summarizes the predicted and

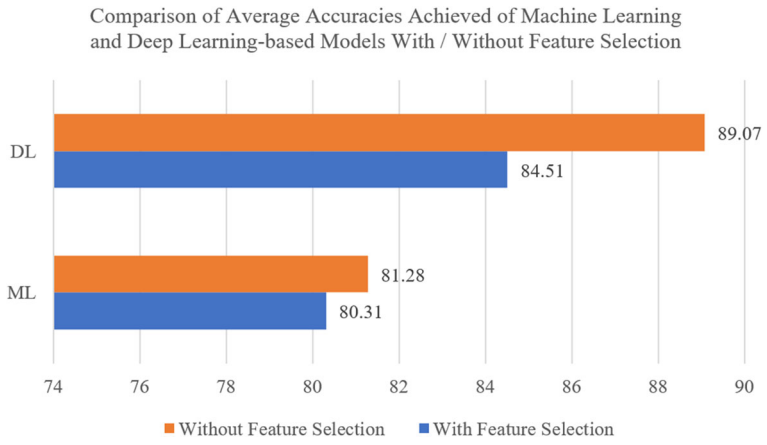

Fig. 12 Comparison of models' accuracies before and after feature selection

actual classes, allowing researchers to calculate the various evaluation metrics that provide insights into the model's performance. These results are essential to be analyzed as it helps in determining the model's threshold and tuning of hyperparameters to achieve optimized performance.

A confusion matrix displays the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for a set of predictions compared to their actual ground truth labels, as shown in Fig. 13

And hence, the evaluation metrics below can be computed using the outcomes:

Accuracy: The proportion of correct predictions to the total number of predictions made. It is calculated as
$Accuracy(Acc) = \frac{TP+TN}{TP+TN+FP+FN}$.

Precision: The proportion of true positives to all positive predictions made by the model. It is calculated as
$Precision(Pr) = \frac{TP}{TP+FP}$.

Recall: The proportion of true positives to all positive samples in the data. It is calculated as
$Recall(Re) = \frac{TP}{TP+FN}$.

F1 Score: The harmonic mean of precision and recall, which provides a balance between the two. It is calculated as
$F1 = \frac{2 \times Pr \times Re}{Pr+Re}$.

Specificity: The proportion of true negatives to all negative samples in the data. It is calculated as
$Specificity = \frac{TN}{TN+FP}$.

In addition to the aforementioned evaluation measures, numerous academics have conducted additional research on analyzing the performance of the models using statistical curves such the Receiver Operating Characteristic (ROC) curve and area Under the Curve (AUC).

Plotting TPR (also known as sensitivity) versus FPR (1-specificity) at various classification thresholds results in the ROC curve. The area under the ROC curve, or AUC, summarizes the model's overall performance across all potential classification thresholds. Each point on the ROC curve represents a specific threshold value, and the curve itself demonstrates how well the model is able to distinguish between positive and negative classes at various threshold

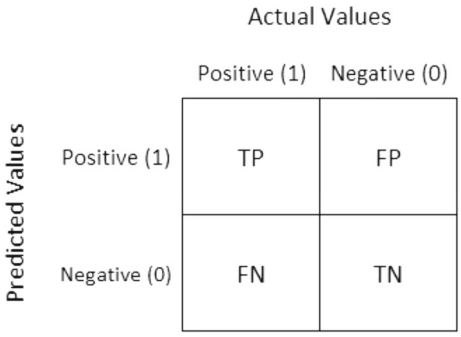

Fig. 13 Example of a confusion matrix

values. Higher values correspond to better performance; the scale runs from 0 to 1. An AUC score of 0.5 indicates that the constructed model is randomly guessing the outcomes. At the same time, an AUC value of 1 suggests that the model can perfectly distinguish between the positive and negative classes.

In other words, while the confusion matrix offers specific details about the model's performance at a specific threshold, the ROC curve and AUC assess the model's ability to distinguish between the positive and negative classes across all potential threshold values. Additionally, the ROC curve and AUC may be more useful when working with unbalanced datasets where one class may be underrepresented because they focus on the overall tradeoff between TPR and FPR rather than particular classification criteria.

5.1 Machine learning-based diabetes prediction models

The scope of machine learning enables the pattern learning and identification through statistical models and deep learning models [11]. These are all achieved by implementing linearly complex statistical algorithms in the models [10]. In many existing diabetes detection solutions, this approach has proven effective and reliable. Possibilities in merging machine learning models with other assistive algorithms such as PCA and PCC made it one of the popular approaches. Researchers are able to probe hyperparameters of the constructed models, hence applying them in solving real-life complex nonlinear applications. When comes to hyperparameters of machine learning models, it refers to the nature of the machine learning algorithms itself, which includes: model architecture, learning rates, number of epochs (iterations), number of branches (specialized for Decision Tree and Random Forest), number of feature clusters, and so on. These parameters are set in advance before the training and validation process begins [65].

Amani et al. [64] implemented two machine learning and one deep learning models and used the PIDD dataset to train the models. Machine learning models they've used in this research work are Support Vector Machine (SVM) and Random Forest (RF). SVM is a popular algorithm where its main function is to determine decision boundaries of linear functions after data analysis is performed [19]; RF is another algorithm where it randomly builds Decision Trees (DT) on different samples [3]: Each DT is structure similar to a flowchart diagram with multiple nodes. Each node in DT represents a classification rule set by the algorithm. Down along the branches attached to the node, they finally lead to the leaf nodes, which each leaf node represents decision output of this iteration [3]. Since native SVM models are not able to solve such nonlinear function, the authors applied kernels function known as Radial Basis Function (RBF) to map the nonlinear to a linear space where it is convinced that the data could be separated in an easier way. As result, the SVM and RF models achieved an accuracy of 73.94% and 79.26% respectively.

In Quan et al. [71] work, they compared machine learning and deep learning diabetes prediction models using J48 and RF. J48 is a statistical-based classifier that is able to produce DT based on data fed to it [6]. Their study used PIDD and a dataset that records hospital physical examination data in Luzhou, China. Both algorithms indicated that blood glucose level is the most impactful feature in diabetes detection function. Hence, they have also conducted another test for both datasets on how removing blood glucose levels from them and using only blood glucose levels for training would affect the models' accuracy. As results, both RF and J48 achieved average accuracy of 72.59% and 75.19% respectively in the PIDD and Luzhou Diabetes Dataset. One significant finding from this research paper is that it was noticed when performing PCA feature selection on the PID dataset, it will cause accuracy

drops in all models implemented in the research, whereas mRMR does not have this issue on the same dataset; However, when incorporating the Luzhou, China dataset, all sorts of feature selection algorithms will only cause drawbacks in models' accuracies. This implies the fact that not all feature selection algorithms are beneficial in improving the models' performances, and this may be the reason that RF model trained with PID dataset in this research performed worse than the one stated in earlier, as in Amani et al. [64] research, they chose to feed the model with all features instead of incorporating any feature selection algorithm. It is also noteworthy to point out that in Quan et al. [71] work, they did not perform any oversampling in mitigating the consequences brought by feature selection.

Nour Abdulhadi et al. [1] compared six machine learning models in their study: RF, Logistic Regression (LR), Voting Classifier, LDA, SVM, and polynomial SVM. In short, LR performs classifications by computing probability of the output, where it assumes every variable is independent to each other [13]; Voting classifier is an algorithm that consists of various base classification models. Its final decision output is based on the prediction basis and findings of the classifiers [30]; Similar to LR, LDA is suitable in solving a more complex function when there are more than two output classes available in the dataset [61]. As a result, RF achieved the highest accuracy of 82%; Logistic Regression and Voting Classifier achieved 80%; Linear Discriminant Analysis (LDA), Polynomial SVM, and SVM all achieved the same accuracy of 79%. Unlike Amani et al. [64] approach, Nour Abdulhadi et al. did impute their implemented PID dataset by filling the missing data with mean values of respective class. This is to avoid further decrement of available training data in the dataset. On top of that, they also chose to feed all features in training the constructed models. Therefore, compared to Amani et al. [64] and Quan et al. [71] models, Nour Abdulhadi et al. models are trained with PID dataset with more available samples and features, and hence performed better in terms of accuracy.

Muhammed et al. [52] performed their research using six machine learning algorithms to construct the diabetes detection models: SVM, K-Nearest Neighbour (KNN), Logistic Regression (LR), Naive Bayes (NB), DT, and RF. KNN performs classifications by determining class percentage of "K", where K denotes number of samples from different classes around the test subject in the data distribution [60]. NB is an algorithm where it performed the classification based on probability of attributes with Bayesian's Theorem. It assumes feature taken in prediction process is statistically independent from the others [46]; They pointed out that to achieve higher accuracy, a database with more samples and zero missing value is required. For data pre-processing, the authors replaced all missing data with mean values and performed min-max normalization on the cleaned dataset. Result shows that both KNN and SVM models performed the best among all six machine learning models, with 77% accuracy in the experiment; Followed by LR and NB models which both achieved accuracy of 74%; Lastly, the DT and RF-based models achieved the lowest accuracy of 71% in this research. In treating the PID dataset, Muhammed et al. and Nour Abdulhadi et al. implemented a similar approach, and the accuracies achieved by their constructed SVM-based models are similar, with 77% and 79% achieved respectively; However, when comes to the RF-based models, both models' performances have a vast difference, with a difference of 11% in achieved accuracy. Similarly, the tree-based DT model also achieved a low accuracy of 71% in this research. The reasons that these decision tree-based models can perform differently will be discussed in the later paragraph.

In Amin et al. [20] research work, they implemented various feature selection algorithms for their constructed DT models such as RF and DT paired Iterative Dichotomiser 3 (DT-ID3) algorithms. As result, the proposed DT+(DT-ID3) model achieved the best accuracy among all three methods, with accuracy of 99% achieved. When full features in the dataset

are extracted for training, the DT model achieved a satisfying accuracy of 98.2%. Now, also as a decision tree-based model, unlike the previous two researches, Amin et al. model can even perform better than a neural network-based model. It is less convincing to state that this result is acceptable and readily to be implemented in solving real-life situation tasks, as the model is trained from the low quality PID dataset, without being oversampled and imputed. It was suspected in this way as it is a common issue found in many tree-based models that they can be easily overfitted if not tuned properly. Excessive numbers of hyperparameters such as number of branches and maximum depth can easily make the models to possess unnecessary complex structure and capture every minute detail in the dataset, including any irrelevant noise and patterns, thus causing overfitting. It is found that authors rarely discuss and investigate overfitting issues that possibly exist in models, and this should not be ignored in any relevant research.

In Sajratul et al. [47] research work, they performed k-mean with Greedy Stepwise Search to compute better feature selection in PIDD dataset. Implementation of Greedy Stepwise Search found out that feature subset with number of pregnancies, blood glucose level, BMI, age, diabetes pedigree function and k-mean cluster gives minimum error to the outcome. These features are fed into LR and RF-based models, achieving 77.08% and 75% of accuracy respectively. Other than analysing the features using Greedy Stepwise Search approach, they also analyse it by incorporating conventional statistical approach, which is via histogram of features versus diabetic outcome. Their findings showed that when both glucose and B.M.I. levels in the PID dataset increased, significant increment in diabetes risk was also spotted, and it was found the other way round when comes to the Skinfold Thickness. After verifying it in parallel with the outcome of Greedy Stepwise Search, unwanted features were removed. Hence the computational cost required were decreased significantly, while maintaining an adequate accuracy. This has shown that rather than simply implementing certain algorithm in feature selection, researchers should also analyse the algorithms' outcome using any sort of approach, and this step is crucial in building a convincing decision support data driven tool.

Huma et al. [42] proposed sampling methods such as Linear Sampling, Shuffled Sampling, Stratified Sampling, and Automatic Sampling to extract features. These sampled features are fed into NB and DT models achieving 76.33% and 86.62% respectively. Different from the previous stated works, Huma et al. did perform oversampling on the PID dataset before using it to train the models. It was also stated in their paper that pre-pruning is essential when constructing a tree-based predictive tool. Based on investigation carried on this paper, even their tree-based DT model is able to achieve a high accuracy of 86.62%, this result still can be improved by implementing data imputation in solving the missing data issue, which is not stated in their work; And also, feature selection can be performed in eliminating any feature that drags the model's performance.

Francesco et al. [38] analyzed features in the PIDD dataset to study if all the features recorded in it are potential risk factors in developing diabetes. After feature selection, minimal improvements are visible in all models. Best accuracies achieved by each of the models are as follows: J48: 74.2%, Hoeffding Tree: 77%, Jrip: 75.5%, BayesNet: 74.9%, RF: 75.4%. Out of these five machine learning models, it is worth to mention that the RF-based model is the only model that achieved lower accuracy after feature selection is performed on the training dataset.

In Fayroza et al. [26] research, they used three different machine learning models: KNN, NB, and LR. For the PIDD dataset, they only performed min-max normalization on the data and did not clean the missing data. As result, LR achieved the best among all three

models, evaluated with three different performance measuring tools. Table 1 shows detailed performance achieved by each model in the paper.

Rani et al. [45] constructed their proposed model using an MLP-based approach while comparing it with LR and RF-based models. Their feature selection processes are as follows: Calculating feature scores in the dataset using an inbuilt class with Tree Based Classifier; Top 10 features with the highest score are then selected from the original dataset for the model training and testing purpose. As result, after performing feature selection, their LR-based model achieved a lower accuracy of 96.153% (compared with 97.115% before feature selection), while the RF-based model achieved the same accuracy of 98.076% throughout the whole experiment. Different from the PID dataset, all features recorded in dataset implemented are related to sequela of infecting diabetes, unlike certain features from PID dataset such as Triceps Skinfold Thickness and Number of Pregnancies, where they are not confirmed to have direct relationship with diabetic outcome. In short, before constructing a dataset, preliminary investigation should be carried out in finding correlation between target features and outcome. This step is essential as construction of a dataset is expensive, thus any possible waste of resources in collecting data should be avoided carefully.

Nadeem et al. [39] implemented a fusion model consists of SVM and ANN algorithms to detect diabetes. This model is then compared with other two SVM and ANN-based models. NHANES and PIDD datasets were fused and used by them for the purpose of this research. After data pre-processing are performed on the two datasets, where the missing data are filled with mean values and normalized using Z-score normalisation method, the data are then fed to the machine learning models. As result, for the two datasets, the Fusion of SVM-ANN model outperformed the other two models, with average of 94.67% achieved; Whereas the SVM and ANN models achieved accuracy of 88.3% and 93.63% respectively. As a direct result from the fusion of NHANES and PID datasets, where 10,627 records with 8 features are available for training and validation purposes, significant improvements can be observed at the performance of the SVM-based model, compared with SVM models in previously reviewed paper. Instead of performing oversampling, Nadeem et al. chose to directly increase the dataset size by fusing two datasets. This is also a considerable solution in mitigating low number of available training and validation data. Despite of that, improvements still can be made in this research, as in the dataset fusion, only 3,556 (33.46%) samples were recorded as “diabetic", while the rest are non-diabetic. Class imbalance problems are common in real life situation, if this can be solved, a more effective training can be performed on the models. Other than that, authors did not perform any feature selection or feature analysation in this research. Since number of data is now sufficient, feature selection may bring positive feedbacks on the models’ performances, and this should be further investigated.

Harleen et al. [24] performed their research using Linear Kernel SVM, RBF kernel applied SVM, KNN, and Multifactor dimensionality reduction (MDR). They’ve also used the PIDD dataset in this research. To improve the dataset, they removed the outlier using statistical methods, and also filled in the missing values by predicting them using k-NN imputation algorithm. They also used the Boruta wrapper algorithm to select relevant and important features from the dataset. As result, linear kernel SVM performed the best among all models, with accuracy of 89% achieved in the testing process. Accuracy achieved by all five models are tabulated in Table 1.

In existing studies, Gradient Boosting Machine (GBM) is also one of the popular approaches in solving this function. In short, GBMs are machines that capable in fitting new algorithm models (base classifiers) in response to the variables [41]. Example algorithms that are similar to GBM are Extreme Gradient Boosting (XGBoost) and LightGBM, where all of these three algorithms are widely use in classification problems.

Leon et al. [27] performed their research using five machine learning models: Language Model (LM), RF, XGBoost, Regularized Generalised Linear Models (Glmnet), Light Gradient Boosting Machine (LightGBM). Glmnet algorithm is an algorithm package that fits multiple functions and capabilities of regression models such as Cox, Poisson, Multinomial, Logistic, and Linear Regression Models [67]. Different from these models. With elastic net, Glmnet linearly combines both L1 and L2 penalties that exist in lasso and ridge methods [67]. These models are also used to importance of every available features in the dataset. Dataset they used contains 111 variables of 27050 samples, where after feature selection and data pre-processing are performed, 59 features of 3723 samples are finally implemented to the machine learning models. The models are used to train train and test using five subsets from the dataset, marked as T6, T12, T18, T24, and T30. Using AUC as validation metric, performance of the five machine learning models used in this research are summarized in Table 1.

In Lai et al's [31] research work, they implemented four machine learning models: GBM, LR, RF, and Rpart to perform diabetes detection. Using the PIDD dataset to train the models, RF performed the best with an AROC value of 85.5%, followed by GBM with AROC of 85.1%. Both LR and Rpart diabetes detection models achieved AROC of 84.6% and 80.5% respectively.

Birjais et al. [7] used GBM, LR, and NB for their research. PIDD datatset is also implemented by them in this research. Similar to the last research paper, the authors also used KNN imputation to predict and fill in the missing values in the PIDD dataset to reduce bias. Although no feature selection is performed, their constructed machine learning-based diabetes detection models still achieved high accuracy of 86%, 79.2%, and 77% by GBM, LR, and NB respectively.

As a brief conclusion for this section, all Machine Learning-based models reviewed are able to perform well. With high average accuracy of 80.6% achieved by all Machine Learning-based models reviewed in this paper, it is considered that machine learning algorithms are capable in solving diabetes detection and classification function. Among all 75 machine learning-based models reviewed, LR, SVM and RF-based models are popular in this domain, taking up to 53% (40 models) among all models. Refer to Fig. 14, LR, SVM, and RF-based models achieved average accuracies of 79.2%, 81.4%, and 81.9% respectively. In order

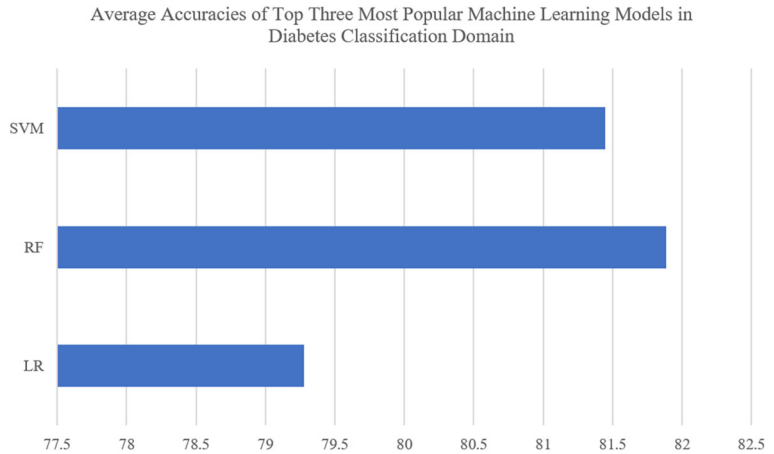

Fig. 14 Average performances (accuracy) of three popular ML-based models (LR, SVM, RF)

to achieve such high accuracy, excessive hyperparameters tuning and feature selection is required to be performed. On top of that, these high results are highly dependent on the quality of the datasets. In LR, RF, SVM implemented models, 27 models are trained with PID dataset, and achieved average accuracy of 79.2%; Whereas the remaining 13 reviewed models are trained with comparatively large datasets with better quality, and achieved average accuracy of 83.3%. In comparison, deep learning models do not require complex maneuver in achieving adequate result, where this will be discussed in the later section.

5.2 Deep learning-based diabetes prediction models

In general, deep learning algorithms attempt to simulate thinking and learning patterns of human brains. One of the benefits is that deep learning algorithms possess various built in functions such as feature selection and feature extraction [54]. Therefore, lesser maneuvers are required in operating such models. In returns, deep learning models need to be fed with larger amount of data for a more precise training. Other than that, due to its powerful and robust operation, they also require equipment with higher computational power than machine learning models do [10].

Among all research reviewed, Convolutional Neural Network (CNN), Deep Belief Network (DBN), and Deep Neural Network (DNN) are the top three popular algorithms used in diabetes detection problem. In general, CNN is extremely popular in data-driven classification problem, where it has already proven capable in image-processing and classification [4, 32]. When data is fed to the algorithm, it will be passed down to the Convolutional Layer where in this layer, convolution will be performed in generating a feature map that clusters and summarizes information of the data [2]. Feature map will then passed down to the Pooling Layer, where in this phase the data size is decreased using specified pooling method [2]. This layer is essential in decreasing required computation power and time. Between Fully Connected Layer and the Output Layer, activation functions exist in solving classification problems. For binary-classes classification problem like diabetes detection, Sigmoid or Softmax functions are commonly implemented [2].

DBN, on the other hand, is a deep learning algorithm that consists of multiple restricted Boltzmann machines (RBMs), stacking on top of each other [63]; where RBM is capable in learning probability distribution of given dataset. In this stacks of RBMs, data are inputted to the first layer of DBN, where their outputs serve as input of the second layer in the stack [63].

Deep neural networks (DNN), is the resulting product of combining multiple hidden layers [22]. In one hidden layer of the ANN, it consists complex combination of mathematical and statistical functions that evaluates the data fed to the algorithm. As stated in the name of DNN - Deep: DNN consists of multiple hidden layers, where output of one layer is served as input of another, similar to the operation of DBN. Complex structure of the DNN enables it to have a robust and precise operation and accuracy in dealing with classification problem [22], in exchange of high computational time and power required.

Maria et al. [17] used CNN algorithm to construct their predictive model. The authors also used the PID dataset for the model's training and validation process. However, as deep learning model requires a large amount of data to achieve a more accurate performance, they've pointed out that most medical datasets available online for machine learning and deep learning purposes record too few samples. Hence, the authors increased the number of samples by using VAE and increasing the number of features by using SAE. For the missing or abnormal data in the dataset, their approach is inputting that particular feature's mean

value. In this research, they've also implemented Multilayer Perceptron (MLP) algorithm, combined with data augmentation methods mentioned above. As a result, the CNN model with the SAE data augmentation method achieved the best result among all state of art, with an accuracy of 92.31%. As mentioned in machine learning section, it is always stated by researcher that number of available training data in PID dataset is a critical point in models' performances. Despite having such dataset, author suggests that this issue can be addressed by implementing various data imputation algorithms when increasing the size of the dataset in real life is not a feasible option.

When implementing a much larger dataset, without much need of data pre-processing, deep learning models can perform well when tuned properly. The research by Sandhiya et al. [51] used two feature selection methods: Conditional Random Field and Linear Correlation Coefficient based Feature Selection (CRF-LCFS). Dataset they chose is obtained from the UCI repository. As a result, before feature selection, their proposed CNN model achieved an accuracy of 82.5%, and it is improved to 84% after feature selection was performed.

P.Prabhu et al. [43] constructed the model based on DBN. Their proposed model then achieved a Recall score of 1, Precision score of 0.6791, and F1 score of 0.808, which performed the best compared to other machine learning-based predictive models.

In research conducted by Safial I. A. et al. [5], they compared accuracies achieved by their constructed DNN-based prediction model with five-fold k and ten-fold cross-validation techniques implemented in the PID dataset. At the end of their research, they concluded that the five-fold cross-validation method performed better than the ten-fold cross-validation method. The DNN-based model achieved an accuracy of 98.04% with the five-fold cross-validation method and 97.27% with the ten-fold cross-validation method.

Nadesh et al. [40] conducted their research using a DNN-based approach. They used an algorithm called Feature Importance (FI), also named Extremely Randomized Trees, for feature selection. This algorithm selects four features from the database based on scores obtained in the FI entropy calculation. They've also used 10-fold cross-validation method in evaluating their DNN-based diabetes prediction model. The constructed DNN-based model is then trained with a train/test split of 60/40, 70/30, and 80/20. As a result, in terms of accuracy, when the training portion goes higher, the performance of DNN model will improve, with an accuracy of 98.16% obtained when a train/test split of 80/20 is used at last. When the 10-fold cross-validation method is used on this model, it achieved the lowest accuracy of 96.10% among all methods, compared with accuracy of 96.77% achieved by the 60/40 split.

Rakshit et al. [44] used Two-Class Neural Network model to classify individuals with and without diabetes from the dataset. After studying various research works on PIDD, the authors used all features recorded in the dataset, except Triceps skinfold thickness, then adjusted the scaling of every feature for the model's learning process. 80% of the data were used to train the model, and the remaining 20% were used in testing the model. As a result, the Two-Class Neural Network-based diabetes predictive model achieved an accuracy of 83.3%.

In [15], Nesreen et al. implemented the model in Just Neural Network (JNN) environment. JNN-based predictive model is able to compute the relative importance of each feature extracted from the dataset, enabling researchers to perform analytic works more precisely. The computed result is then used to aid in feature selection to improve the outcome. This model's accuracy in predicting diabetes is 87.3%, with 76% of the recorded data taken for training and 24% for testing.

Other than conventional invasive dataset such as PID dataset, Vidhya et al. [62] conducted research work on diabetes complication models using DBN, SVM, and ANN using non-invasive datasets. This research work focused on finding the most impactful risk factors of diabetes in diabetic complications using classification algorithms stated above. They

took a dataset from the UCI diabetic repository, recorded with 13 features based on the life behaviours of samples. They used Restricted Boltzmann Machine (RBM) in pre-processing the dataset. Results are as follows: DBN achieved training/testing stage classification accuracy of 81.19/80.99, SVM achieved 72.72/62.81, and ANN achieved 76.52/57.61. In selecting the features, authors implemented unsupervised learning and pre-training of Restricted Boltzmann Machine before extracting demanded features. Since their result showed that diabetes classification based on non-invasive dataset is possible and able to achieve high accuracy, this research direction should not be ignore and further investigated.

Another study by Ryu et al. [48], they constructed their diabetes prediction model using DNN approach. In their study, they focused on screening samples with undiagnosed type 2 diabetes, using the NHANES dataset. It is worth to mention that they also focus on taking non-invasive features that do not require lab tests to be performed or any sort of blood samples. As result, the proposed DNN-based model scored an AUC of 80.11. Both studies by Vidhya [62] and Ryu [48] et al. have proven that non-invasive datasets can be implemented in deep learning approach to solve diabetes classification task. However, in order to achieve good results, presence of relatively large dataset is vital.

Deep learning approaches have also been proven that its strong, robust and precise ability in handling image processing will come in handy when dealing with detection of diabetes retinopathy - a complication caused by diabetes, which one of its consequences is damaging retina of human eye. Lam et al. [32] and Arcadu [4] used CNN-based algorithms in diagnosing diabetic retinopathy, while Gadekallu et al. [16] implemented a DNN-based approach, paired with PCA-Firefly Feature Selection algorithm. All three researches stated have proved that diabetic retinopathy diagnosis using deep learning approaches is feasible, with highest accuracy of 97% achieved in Gadekallu et al's model [16].

In [9, 29, 38, 42, 45, 47, 64, 71], other than machine learning approaches, they've also implemented deep learning models in detecting diabetes. The researchers then studied which approach can perform better for this function. Detailed information on these research are tabulated in Table 2.

As a brief summary for this section, compared with Machine Learning-based Diabetes Detection Models, the Deep Learning-based models are able to perform better, with an average accuracy of 86.7% achieved in this diabetes classification function. Based on Fig. 15

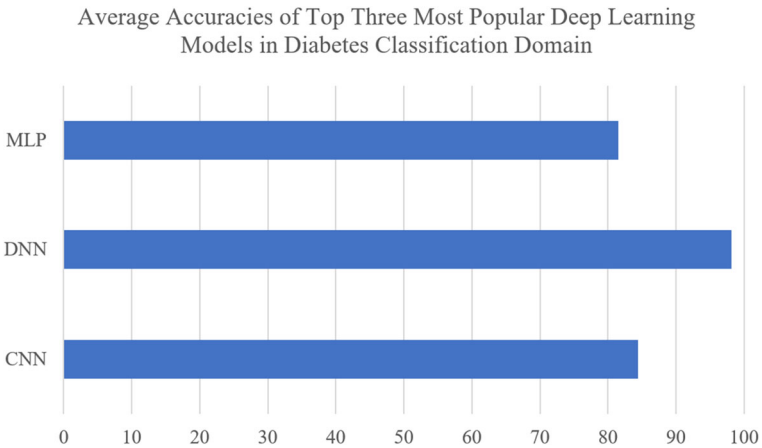

Fig. 15 Performance of three popular DL-based models (CNN, DNN, MLP)

above, it shows the performances of three popular deep learning-based models: CNN, DNN, and MLP. In general, they achieved an average accuracy of 84.37%, 98.1%, and 81.49% respectively.

6 Challenges, research gaps and comparisons

Although progression has been made in recent research, diabetes classification by machine learning or deep learning approaches still present several issues to be addressed. Furthermore, more potentials and improvements to be explored are found in this comprehensive literature review. These stated points will be discussed in this section.

6.1 Challenges

In general, future challenges in diabetes classification using machine learning and deep learning techniques can be concluded into three aspects: Availability of large and good quality datasets, medical analyzation of features using data driven solution, and ethical issues in implementing this solution to the public. A huge dataset must be used to improve training and attain greater accuracy and performance. Building an accurate diabetes classification model requires a sizable database for model training. However, creating such a database involves a significant investment of time and money, causing that obtaining the perfect dataset comes with great number of issues to overcome. Researchers suggest exploring non-lab-invasive lab measurement databases, but it is crucial to analyze target features and diabetic outcomes beforehand to prevent excessive costs associated with building the database. For instance, the availability of current diabetes datasets presents another difficulty because they are frequently not globally representative and may create racial or regional biases during model training. This problem becomes important since datasets are crucial to machine and deep learning models. Additionally, because classification models frequently operate as "black boxes", it is challenging for medical professionals to understand the results adequately. For diabetes classification to advance and patient treatment to improve, these obstacles must be overcome.

Medical personnel must interpret the outcome produced by a diabetes classification machine learning or deep learning-based model for several reasons. First, machine learning or deep learning classification models are a supportive tool aiding treatment decision. While the model's outcome can provide insights into patients developing diabetes, the people need to interpret the information generated by the model when making treatment decisions, such as prescribing medication or suggesting lifestyle modifications and further diagnostic tests. This solution is less persuasive to the patients without considering the outcome. This is crucial since it is the patients' constitutional right to comprehend the care being given. Therefore, knowing the model's results enables medical providers to inform patients about their risk of developing diabetes or their diagnosis effectively. It gives patients the power to decide what is best for their health, including following treatment programs, forming healthy behaviours, and actively managing their disease.

6.2 Research gaps

Despite the fact that several studies by Swapna et al. [58] and Vidhya et al. [62] have proven that non-invasive measurements can be implemented in diabetes classification tasks, a standard procedure is not proposed for collecting the required data for further investigation and

improving the classification models. On top of that, features such as ECG studied by Swapna et al. are not proven to have a direct relationship to diabetes mellitus. Hence, this result needs to be further investigated from medical perspective until a common consensus is on par with machine learning or deep learning perspective.

Furthermore, it is common that diabetes complications-related measurements such as diabetes retinopathy or the presence of slow wound healing in samples are taken in training the classification models. For instance, research by Rani et al. [45] implemented datasets containing measurements related to diabetes complications, achieving average high accuracy of 97.3% among all models constructed. While these measurements are non-invasive in particular perspective, they should act as supportive features instead of main features that help to train the classification models. This is stated in this way because with the diabetes complications measurements, certain conventional anthropometric diagnosis tests can already be utilized to diagnose the presence of diabetes. Study such as Vidhya et al. approach, where they implemented datasets containing health behaviors related measurement should be investigated with more focus, as their constructed models not only aids in diabetes classification but also provide a new perspective in analyzing possible health behaviors that cause diabetes. Vidhya et al. models' performances are not satisfying (Accuracy: 80.99% with DBN and 62.81% with ANN); hence it is advised that the dataset can be imputed with diabetes complications-related features in strengthening the models' performance.

Although it is stated so, non-invasive diabetes complications such as diabetic retinopathy still require to perform fluorescein angiogram test, and a modern mydriatic fundus camera is the tool most commonly used for this purpose, where it averagely costs more than 1000 USD. This undoubtedly will increase the cost burden in constructing the datasets. Therefore, the selection of data collection and standardization of datasets for machine learning or deep learning-based diabetes classification tools are also topics that should be investigated in inventing a modern data-driven diabetes classification tool.

6.3 Comparison and findings

Tables 1 and 2 displayed two generalized information on the performances achieved by Comparing performances achieved by both machine learning and deep learning-based models: Among all 75 reviewed machine learning-based diabetes detection-based models, they achieved an average accuracy of 80.6%; Among all reviewed deep learning-based models, they achieved an average accuracy of 86.7%. In common situations, deep learning-based models are able to perform better than machine learning-based models, and this is caused by complex yet powerful structure design of the deep learning algorithm itself. In exchange for that, the "black-box" nature in operation of deep learning models is more complicated and opaque when compared with machine learning-based models. This is because the deep learning-based models tend to perform the whole function in one line, i.e., its operation concluded both feature extraction, selection, and classification in one iteration; While this has greatly decreased the complexity in handling the model, it also decreases the flexibility in maneuvering the model at the same time, such that it could lead to a problem that researchers will have difficulties in analyzing every small component during the classification process.

Although it is stated in this way, it was also spotted that when comparing certain studies, a large dataset does not always benefit to the model's performance even the algorithms implemented are the same. This phenomenon can be found from studies by Amani et al. [64], Sandhiya et al. [51] and Maria et al. [17], where they all implemented CNN algorithms, but obtained vastly different results in terms of accuracy. This can be observed directly from

Table 1 Summarized performance of machine learning models reviewed

Algorithm	Database	Studies	Results	Metrics
BayesNet	PID	[38]	74.9	Acc
DT	PID	[52]	71	Acc
		[42]	96.62	Acc
			98.2	Acc
	Self-Prepared	[20]	99	Acc
			98.3	Acc
GBM	PID	[31]	85.1	Acc
		[7]	86	Acc
Glmnet	EHR, Slovenia	[27]	N/A	N/A
Hoeffding Tree	PID	[38]	77	Acc
J48	Luzhou	[71]	78.53	Acc
	PID	[71]	75.34	Acc
		[38]	74.2	Acc
Jrip	PID	[38]	75.5	Acc
KNN	PID	[52]	77	Acc
		[24]	88	Acc
			69	Pr
		[26]	68	Re
			68	F1
LDA	PID	[1]	79	Acc
LightGBM	EHR, Slovenia	[27]	N/A	N/A
LM	EHR, Slovenia	[27]	N/A	N/A
LR	PID	[1]	80	Acc
		[52]	74	Acc
		[47]	77.08	Acc
			94	Pr
		[26]	70	Re
			79	F1
		[31]	84.6	Acc
		[7]	79.2	Acc
	Self-Prepared	[45]	97.115	Acc
			96.153	Acc
MDR	PID	[24]	83	Acc
NB	PID	[52]	74	Acc
		[42]	76.33	Acc
			79	Pr
		[26]	68	Re
			72	F1
		[7]	77	Acc

Table 1 continued

Algorithm	Database	Studies	Results	Metrics
RF	EHR, Slovenia	[27]	N/A	N/A
	Luzhou	[71]	80.84	Acc
	PID	[64]	83.67	Acc
		[71]	76.04	Acc
		[1]	82	Acc
		[52]	71	Acc
		[47]	75	Acc
		[38]	74.3	Acc
		[31]	85.5	Acc
	Self-Prepared	[45]	98.076	Acc
			98.076	Acc
Rpart	PID	[31]	80.5	Acc
Voting Classifier	PID	[1]	80	Acc
XGBoost	EHR, Slovenia	[27]	N/A	N/A
SVM	NHANES	[39]	88.3	Acc
			94.67	Acc
	PID	[24]	89	Acc
			84	Acc
		[64]	65.38	Acc
		[1]	79	Acc
		[52]	77	Acc
		[1]	79	Acc

Table 2. While it not stated in Amani et al. paper, Sandhiya et. al. and Maria et al. did mentioned the number of epochs in training their models: 400 and 600 respectively. Moreover, Sandhiya et al. model may have a more complex structure compared with Maria et al. model: With one more convolutional layer and multiple dense layers placed in their model, this may caused that their model is overfitted to the training samples. In short, in diabetes classification task, similar to RF model, researchers often prevent their deep learning-based models from being too complex, as it might easily cause overfitting case to occur.

Compared with machine learning-based approach, the whole diabetes detection process is separated into smaller sections, where researchers have greater freedom in analyzing and tuning the models. Similar to deep learning-based models, this freedom comes with a cost: The classification algorithms will have to be paired with complex activation functions, feature selection algorithms, and various data sampling and pre-processing steps in order to generate an adequate result, due to its weakness in solving any nonlinear function as diabetes detection and classification.

Table 2 Summarized performance of deep learning models reviewed

Algorithm	Database	Studies	Results	Metrics
ANN	UCI Diabetic Repository	[62]	62.81	Acc
CNN	Diabetec Retinopathy	[4]	N/A	N/A
		[31]	N/A	N/A
	PID	[64]	76.81	Acc
		[17]	92.31	Acc
			80.52	Acc
	UCI Diabetic Repository	[51]	84	Acc
			82.5	Acc
DBN	PID	[43]	1	Re
			67.91	Pr
			80.8	F1
	UCI Diabetic Repository	[62]	80.99	Acc
DNN	NHANES	[48]	80.11	AUC
	PID	[5]	98.04	Acc
			97.27	Acc
		[40]	96.1	Acc
			96.77	Acc
			97.54	Acc
			98.16	Acc
JNN	Urmia	[15]	87.3	Acc
MLFNN	PID + Removing Samples		84.17	Acc
	PID + Replace as mean	[29]	81.73	Acc
	PID + Replace as 0		80.38	Acc
MLP	PID	[17]	79.22	Acc
			85.71	Acc
			80.52	Acc
		Yakin	85	Acc
	Self-Prepared	[9]	81.3	Acc
		[45]	77.9221	Acc
			91.3853	Acc
NN	Luzhou	Quian	78.41	Acc
	PID	[71]	77.67	Acc
		[44]	83.3	Acc
PNN	Self-Prepared	[9]	84.35	Acc

7 Conclusion and future work

In conclusion, constructing a data-driven diabetes detection model is imminent as prevalence of diabetes around the globe has been rising day by day. Reliabilities of diabetes detection models based on both non-lab test and non-invasive measurement should be further investigated in finding possibilities of reducing medical expenses and labor required in diabetes detection and treatment. In order to achieve this, a dataset with better quality, as in: more

recorded features and samples, no missing or abnormal values, is required. After reviewing more than 50 machine learning and deep learning models in this paper, it is concluded that both types of algorithms (machine learning and deep learning) have their own benefits in different fields.

With findings showing that feature selection is beneficial to most of the machine learning-based diabetes detection models, it is important to implement feature selection algorithms to the dataset, then perform a cross-test with dataset that has not performed with feature selection, in case feature selection has a negative impact on the tested models. Furthermore, although in most deep learning algorithms, features extraction and selection functions have already built in in them, studies still suggest that a pre-feature selection process can be performed to study their impact on the classification, for that this region is rarely studied in existing research. Last but not least, researchers should also investigate the most cost-effective features in constructing datasets for this purpose, where it is one of the top priorities in addressing the cost issue mentioned aforehand. On top of all, cost, ethical, and medical analyzation issues are the most important factor to be considered when inventing a data driven solution for diabetes classification task.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data Availability Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Abdulhadi N, Al-Mousa A (2021) Diabetes detection using machine learning classification methods. In: 2021 International Conference on Information Technology (ICIT). IEEE, p 350–354
2. Albawi S, Mohammed TA, Al-Zawi S (2017) Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET). Ieee, p 1–6
3. Ali J, Khan R, Ahmad N, Maqsood I (2012) Random forests and decision trees. Int J Comp Sci Iss 9(5):272
4. Arcadu F, Benmansour F, Maunz A, Willis J, Haskova Z, Prunotto M (2019) Deep learning algorithm predicts diabetic retinopathy progression in individual patients. NPJ Dig Med 2(1):1–9
5. Ayon SI, Islam MM (2019) Diabetes prediction: a deep learning approach. Int. J. Inf. Eng. Electron. Bus 12(2):21
6. Bhargava N, Sharma G, Bhargava R, Mathuria M (2013) Decision tree analysis on j48 algorithm for data mining. Proc Int J Adv Res Comp Sci Softw Eng 3(6)
7. Birjais R, Mourya AK, Chauhan R, Kaur H (2019) Prediction and diagnosis of future diabetes risk: a machine learning approach. SN Appl Sci 1(9):1–8

8. Chawla R, Madhu S, Makkar B, Ghosh S, Saboo B, Kalra S et al (2020) Rssdi-esi clinical practice recommendations for the management of type 2 diabetes mellitus 2020. Indian J. Endocrinol. Metab 24(1):1
9. Czmil A, Czmil S, Mazur D (2019) A method to detect type 1 diabetes based on physical activity measurements using a mobile device. Appl Sci 9(12):2555
10. Dargan S, Kumar M, Ayyagari MR, Kumar G (2020) A survey of deep learning and its applications: a new paradigm to machine learning. Arch Comput Methods Eng 27(4):1071–1092
11. Das K, Behera RN (2017) A survey on machine learning: concept, algorithms and applications. Int J Innov Res Comp Commun Eng 5(2):1301–1309
12. Dinh A, Miertschin S, Young A, Mohanty SD (2019) A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med. Inform. Decis. Mak. 19(1):1–15
13. Dreiseitl S, Ohno-Machado L (2002) Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform 35(5–6):352–359
14. Dua D, Graff C (2017) UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
15. El Jerjawi NS, Abu-Naser SS (2018) Diabetes prediction using artificial neural network. Int J Adv Sci Technol 121
16. Gadekallu TR, Khare N, Bhattacharya S, Singh S, Maddikunta PKR, Ra I-H, Alazab M (2020) Early detection of diabetic retinopathy using pca-firefly based deep learning model. Electronics 9(2):274
17. García-Ordás MT, Benavides C, Benítez-Andrades JA, Alaiz-Moretón H, García-Rodríguez I (2021) Diabetes detection using deep learning techniques with oversampling and feature augmentation. Comput Methods Programs Biomed 202:105968
18. Ge Q, Xie XX, Xiao X, Li X (2019) Exosome-like vesicles as new mediators and therapeutic targets for treating insulin resistance and ß-cell mass failure in type 2 diabetes mellitus. J Diabet Res 2019
19. Ghosh S, Dasgupta A, Swetapadma A (2019) A study on support vector machine based linear and non-linear pattern classification. In: 2019 International Conference on Intelligent Sustainable Systems (ICISS). IEEE, p 24–28
20. Haq AU, Li JP, Khan J, Memon MH, Nazir S, Ahmad S, Khan GA, Ali A (2020) Intelligent machine learning approach for effective recognition of diabetes in e-healthcare using clinical data. Sensors 20(9):2649
21. Hou J, Sang Y, Liu Y, Lu L (2020) Feature selection and prediction model for type 2 diabetes in the chinese population with machine learning. In: 4th International Conference on Computer Science and Application Engineering (CSAE 2020). p 1–7
22. Hu J, Zhang J, Zhang C, Wang J (2016) A new deep neural network based on a stack of single-hidden-layer feedforward neural networks with randomly fixed hidden neurons. Neurocomputing 171:63–72
23. Kaul K, Tarr JM, Ahmad SI, Kohner EM, Chibber R (2013) Introduction to diabetes mellitus. Diabetes 1–11
24. Kaur H, Kumari V (2020) Predictive modelling and analytics for diabetes using a machine learning approach. Appl Comput Inform 1330
25. Kazmi NHS, Gillani S, Afzal S, Hussain S (2013) Correlation between glycated haemoglobin levels and random blood glucose. J Ayub Med Coll 25(1–2):86–88
26. Khaleel FA, Al-Bakry AM (2021) Diagnosis of diabetes using machine learning algorithms. Mater Today Proc
27. Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G (2020) Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Sci Rep 10(1):1–12
28. Krishnappa M, Patil K, Parmar K, Trivedi P, Mody N, Shah C, Faldu K, Maroo S, Parmar D (2020) Effect of saroglitazar 2 mg and 4 mg on glycemic control, lipid profile and cardiovascular disease risk in patients with type 2 diabetes mellitus: a 56-week, randomized, double blind, phase 3 study (press xii study). Cardiovasc Diabetol 19(1):1–13
29. Kumar S, Bhusan B, Singh D, Kumar Choubey D (2020) Classification of diabetes using deep learning. In: 2020 International Conference on Communication and Signal Processing (ICCSP). IEEE, p 0651–0655
30. Kumari S, Kumar D, Mittal M (2021) An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. Int J Cognit Comput Eng 2:40–46
31. Lai H, Huang H, Keshavjee K, Guergachi A, Gao X (2019) Predictive models for diabetes mellitus using machine learning techniques. BMC Endoc Disord 19(1):1–9
32. Lam C, Yi D, Guo M, Lindsey T (2018) Automated detection of diabetic retinopathy using deep learning. AMIA Summits Transl. Sci. Proc. 2018:147
33. Liu G, Li Y, Hu Y, Zong G, Li S, Rimm EB, Hu FB, Manson JE, Rexrode KM, Shin HJ et al (2018) Influence of lifestyle on incident cardiovascular disease and mortality in patients with diabetes mellitus. J Am Coll Cardiol 71(25):2867–2876
34. Lukmanto RB, Suharjito Nugroho, A., Akbar, H, (2019) Early detection of diabetes mellitus using feature selection and fuzzy support vector machine. Procedia Computer Science 157:46–54. https://doi.org/10.

1016/j.procs.2019.08.140. (**The 4th International Conference on Computer Science and Computational Intelligence (ICCSCI 2019)?: Enabling Collaboration to Escalate Impact of Research Results for Society**)
35. Mallone R, Mannering S, Brooks-Worrell B, Durinovic-Bello I, Cilio C, Wong FS, Schloot N (2011) Isolation and preservation of peripheral blood mononuclear cells for analysis of islet antigen-reactive t cell responses: position statement of the t-cell workshop committee of the immunology of diabetes society. Clin Exper Immunol 163(1):33–49
36. Maratni NPT, Saraswati MR, Dewi NNA, Yasa I, Eka Widyadharma IP, Putra IBK, Suastika K (2021) Association of apolipoprotein e gene polymorphism with lipid profile and ischemic stroke risk in type 2 diabetes mellitus patients. J Nutr Metabol 2021
37. Mathis D, Vence L, Benoist C (2001) ß-cell death during progression to diabetes. Nature 414(6865):792–798
38. Mercaldo F, Nardone V, Santone A (2017) Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. Procedia Comput. Sci. 112:2519–2528
39. Nadeem MW, Goh HG, Ponnusamy V, Andonovic I, Khan MA, Hussain M (2021) A fusion-based machine learning approach for the prediction of the onset of diabetes. In: Healthcare, vol. 9. MDPI, p 1393
40. Nadesh RK, Arivuselvan K et al (2020) Type 2: diabetes mellitus prediction using deep neural networks classifier. Int J Cogn Comput Eng 1:55–61
41. Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Front Neurorobot 7:21
42. Naz H, Ahuja S (2020) Deep learning approach for diabetes prediction using pima indian dataset. J Diabet Metabol Disord 19(1):391–403
43. Prabhu P, Selvabharathi S (2019) Deep belief neural network model for prediction of diabetes mellitus. In: 2019 3rd International Conference on Imaging, Signal Processing and Communication (ICISPC). IEEE, p 138–142
44. Rakshit S, Manna S, Biswas S, Kundu R, Gupta P, Maitra S, Barman S (2017) Prediction of diabetes type-ii using a two-class neural network. In: International Conference on Computational Intelligence, Communications, and Business Analytics. Springer, p 65–71
45. Rani DVV, Vasavi D, Kumar K et al (2021) Significance of multilayer perceptron model for early detection of diabetes over ml methods. J. Univ. Shanghai Sci Technol 23(8):148–160
46. Rish I, et al (2001) An empirical study of the naive bayes classifier. In: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, vol. 3. p 41–46
47. Rubaiat SY, Rahman MM, Hasan MK (2018) Important feature selection & accuracy comparisons of different machine learning models for early diabetes detection. In: 2018 International Conference on Innovation in Engineering and Technology (ICIET). IEEE, p 1-6
48. Ryu KS, Lee SW, Batbaatar E, Lee JW, Choi KS, Cha HS (2020) A deep learning model for estimation of patients with undiagnosed diabetes. Appl Sci 10(1):421
49. Sacks DB, Arnold M, Bakris GL, Bruns DE, Horvath AR, Kirkman MS, Lernmark A, Metzger BE, Nathan DM (2011) Guidelines and Recommendations for Laboratory Analysis in the Diagnosis and Management of Diabetes Mellitus. Diabetes Care 34(6):61–99. https://doi.org/10.2337/dc11-9998 https://diabetesjournals.org/care/article-pdf/34/6/e61/609322/e61.pdf
50. Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, Colagiuri S, Guariguata L, Motala AA, Ogurtsova K, Shaw JE, Bright D, Williams R (2019) Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas, 9th edition. Diabetes Res Clin Pract 157:107843. https://doi.org/10.1016/j.diabres.2019.107843
51. Sandhiya S, Palani U (2020) An effective disease prediction system using incremental feature selection and temporal convolutional neural network. J Ambient Intell Humaniz Comput 11(11):5547–5560
52. Sarwar MA, Kamal N, Hamid W, Shah MA (2018) Prediction of diabetes using machine learning algorithms in healthcare. In: 2018 24th International Conference on Automation and Computing (ICAC). IEEE, p 1–6
53. Schober P, Boer C, Schwarte LA (2018) Correlation coefficients: appropriate use and interpretation. Anesth Analg 126(5):1763–1768
54. Shrestha A, Mahmood A (2019) Review of deep learning algorithms and architectures. IEEE Access 7:53040–53065
55. Sivaranjani S, Ananya S, Aravinth J, Karthika R (2021) Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction. In: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), vol. 1. IEEE, p 141–146
56. Smith JW, Everhart JE, Dickson W, Knowler WC, Johannes RS (1988) Using the adap learning algorithm to forecast the onset of diabetes mellitus. Am Med Inform Assoc 261

57. Song F, Guo Z, Mei D (2010) Feature selection using principal component analysis. In: 2010 International Conference on System Science, Engineering Design and Manufacturing Informatization, vol. 1, IEEE, p 27–30
58. Swapna G, Vinayakumar R, Soman K (2018) Diabetes detection using deep learning algorithms. ICT Express 4(4):243–246
59. Tao K-M, Sokha S, Yuan H-B (2019) Sphygmomanometer for invasive blood pressure monitoring in a medical mission. Anesthesiology 130(2):312–312
60. Taunk K, De S, Verma S, Swetapadma A (2019) A brief review of nearest neighbor algorithm for learning and classification. In: 2019 International Conference on Intelligent Computing and Control Systems (ICCS). IEEE, p 1255–1260
61. Tharwat A, Gaber T, Ibrahim A, Hassanien AE (2017) Linear discriminant analysis: A detailed tutorial. AI Commun 30(2):169–190
62. Vidhya K, Shanmugalakshmi R (2020) Deep learning based big medical data analytic model for diabetes complication prediction. J. Ambient Intell. Humaniz. Comput. 11(11):5691–5702
63. Wang G, Qiao J, Bi J, Li W, Zhou M (2018) Tl-gdbn: Growing deep belief network with transfer learning. IEEE Trans. Autom. Sci. Eng. 16(2):874–885
64. Yahyaoui A, Jamil A, Rasheed J, Yesiltepe M (2019) A decision support system for diabetes prediction using machine learning and deep learning techniques. In: 2019 1st International Informatics and Software Engineering Conference (UBMYK). IEEE, p 1–4
65. Yang L, Shami A (2020) On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing 415:295–316
66. Yang G, Qian T, Sun H, Xu Q, Hou X, Hu W, Zhang G, Fang Y, Song D, Chai Z et al (2022) Both low and high levels of low-density lipoprotein cholesterol are risk factors for diabetes diagnosis in chinese adults. Diabet Epidemiol Manag 6:100050
67. Yuan G-X, Ho C-H, Lin C-J (2011) An improved glmnet for l1-regularized logistic regression. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p 33–41
68. Zaccardi F, Dhalwani NN, Papamargaritis D, Webb DR, Murphy GJ, Davies MJ, Khunti K (2017) Nonlinear association of bmi with all-cause and cardiovascular mortality in type 2 diabetes mellitus: a systematic review and meta-analysis of 414,587 participants in prospective studies. Diabetologia 60(2):240–248
69. Zhou L, Zheng X, Yang D, Wang Y, Bai X, Ye X (2021) Application of multi-label classification models for the diagnosis of diabetic complications. BMC Med. Inform. Decis. Mak. 21(1):1–10
70. Zhu T, Li K, Herrero P, Georgiou P (2020) Deep learning for diabetes: a systematic review. IEEE J. Biomed. Health Inform. 25(7):2744–2757
71. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H (2018) Predicting diabetes mellitus with machine learning techniques. Front Genet 9:515