RES MILITARIS
REVUE EUROPEENNE D ETUDES EUROPEAN JOURNAL OF MILITARY STUDIES

Social Science Journal

Diabetic Diagnosis Using Machine Learning as well as Deep Learning Techniques

Somendra Tripathi, Hari Om Sharan, C. S. Raghuvanshi

Faculty of Engineering & Technology, Rama University, Kanpur, Uttar Pradesh, India

Abstract:

A growing incidence of Type II diabetes worldwide has prompted the medical industry to explore solutions for enhancing their medical technology. The fields of machine learning and deep learning are now being actively researched for the development of intelligent and efficient methods for detecting diabetes. This research thoroughly examines and explores the effects of the most recent machine learning and deep learning methods on the detection and categorization of diabetes. The accessibility of diabetes statistics is noted to be restricted. The databases consist of measures obtained from laboratory-based tests and invasive procedures. To develop an efficient solution that is both cost-effective and high-performing, it is necessary to do research on anthropometric measures and non-invasive examinations. Multiple studies have shown the potential to develop detection models using anthropometric measures and non-invasive medical indications. This research examined the effects of oversampling strategies and data decrease in dimensionality via the selection of features.

Keywords: Diabetes diagnosis using machine learning, deep learning, and feature selection, Anthropometric measurements

1.0 Introduction

Diabetes, also called type 2 diabetes, is a persistent metabolic disorder that impacts the human body's ability to turn blood sugar into energy. Individuals diagnosed with diabetes are unable to effectively regulate their blood sugar levels, leading to elevated levels of both blood sugar & arterial pressure. If diabetes fails to be promptly recognized, diagnosed, and effectively treated in its early stages, it has the potential to result in several life-threatening conditions, including diabetic retinopathy and neuropathy, renal failure, and other coronary artery disease [33]. Despite substantial progress in the medical field over the last century, the prevalence of diabetes continues to increase in most cultures. The incidence of this phenomenon is increasing in all countries, irrespective of the income level of the population. Based on a study conducted by [50], it is projected that the percentage of adults worldwide diagnosed with diabetes will increase to 10.2% (578 million) by 2030 and further climb to 10.9% (700 million) by 2045. Therefore, it is crucial to create sophisticated technologies that assist medical professionals in diagnosing diabetes and providing decision support. Conventional laboratory-based approaches for detecting diabetes are typically characterized by being both time-consuming and costly. Typically, doctors use oral glucose tolerance, fasting blood sugar, or random blood sugar testing to make an approximate prediction and diagnosis of diabetes mellitus in patients [25]. The Glycated hemoglobin (A1C) test was released to the public in 1980 as a means to perform additional confirmation on the diagnosis of diabetes in patients. This test evaluates the percentage of blood sugar that is linked to the haemoglobin over a three-month period. The procedures involved in this test are intricate, laborious, and necessitate the presence of medical experts and specialized equipment on-site to conduct the laboratory tests. It is worth noting that if these resources are not available at the location, additional costs will be incurred for the collection and storage of blood samples, among other things [35].

Type 2 diabetes mellitus (T2DM) is the predominant form of diabetes, characterized by high blood sugar levels resulting from either the body's resistance to insulin or inadequate synthesis of insulin [8]. Several factors, including sedentary lifestyle, unhealthy food, tobacco use, excessive alcohol consumption, and obesity, have been identified as potential contributors to the development of Type 2 Diabetes Mellitus (T2DM) [28, 36]. Type 2 diabetes mellitus (T2DM) increases the likelihood of developing cardiovascular illnesses, including coronary heart disease, stroke, peripheral artery disease, and aortic disorders. These conditions are mostly caused by high blood pressure resulting from the presence of diabetes mellitus [28]. Furthermore, it is worth noting that Type 2 Diabetes Mellitus (T2DM) can be diagnosed based on genetic factors, as there is a genetic connection between insulin resistance, insulin secretion, and high blood sugar levels [18].

Several researchers have currently developed and analyzed various data-driven models for detecting diabetes [9, 64, 71]. However, the majority of researchers utilized datasets that were acquired using medical indicators based on lab tests for the purpose of training and validating the models [20, 44, 51]. Nevertheless, these prediction models are deemed redundant since lab-test-based data can already be utilized to determine and diagnose whether an individual has diabetes or not with encouraging precision [49]. There is a greater demand for a preliminary diagnosis solution that does not require any laboratory test measurements. Hence, it is imperative to do research on the characteristics of anthropometric measurements and their influence on models for detecting diabetes. Machine learning is a commonly used method in data-driven diabetes detection classification solutions. It is popular because it can classify data using statistical approaches, without needing a lot of computational power [10]. While the relationship between potential risk factors and diabetes is not linear, machine learning algorithms can still be used to classify linear functions. To address this issue, various kernel functions can be added to machine learning models, allowing them to predict nonlinear functions using methods based on statistics.

An alternative often used for this task is employing deep learning methodologies. Deep Neural Networks (DNNs) can achieve satisfactory classification results with minimal manual engineering optimization, thanks to their great processing power. In addition, Deep Neural Networks (DNN) have the ability to solve functions with nonlinear variables, unlike machine learning models which need to be combined with specific algorithms and functions in order to achieve this [10]. Therefore, its strong and simplified learning capabilities have positioned it as one of the top choices for resolving this categorization issue. Therefore, it is recommended to create a hybrid model that combines machine learning and deep learning methods in order to produce a straightforward and efficient model that can assess all the retrieved features from a dataset.

This specific diabetes detection challenge is chosen because diabetes mellitus is the primary factor that contributes to other potentially fatal cardiovascular illnesses. Thanks to the progress achieved in the field of data science over the last ten years, it has become increasingly feasible to classify diabetes using machine learning and deep learning methods. Thus, there is a need for a pre-diabetes classification tool that is trained using non-invasive laboratory measures data.

This is because the development of such a tool has the potential to significantly alleviate the burden on the current healthcare system from all perspectives. Furthermore, it is imperative to explore the potential of machine learning and deep learning techniques in evaluating the underlying factors contributing to diabetes mellitus.

This review study is crucial for exploring the potential of building a data-driven diabetes classification model. The model would be trained using datasets that contain anthropometric or non-lab-invasive medical metrics. Various machine learning and deep learning methods are examined to assess their individual strengths and weaknesses in addressing this diabetes categorization problem. The approaches encompass all aspects, including the selection of datasets, the process for imputing missing data, feature selection, sampling, and, most crucially, the classification algorithms employed to carry out the work.

2.0 Collecting data

Identifying the right data set is a crucial step in training machines machine learning and deep learning algorithms for diabetes identification. The accuracy of these models heavily depends on the quality of the dataset used [12]. Recent research have shown that the majority of data-driven models for detecting diabetes have been trained using a publically accessible dataset called the Pima Indians Endocrinology Database (PIDD) [14]. These models are specifically designed for machine training and deep learning applications. The dataset, which was published in 1988, contains information on nine characteristics of 768 female individuals who were at least 21 years old at the time of data collection. The recorded features include age, body mass index (BMI), diabetic pedigree functional (DPF), number of pregnancy, plasma glucose levels in a test for oral glucose tolerance, arterial pressure, triceps folds in the skin thickness, 2-per hour. serum glucose, and the presence of diabetic in the sample's physique.

Age, BMI, total number of their pregnancies diabetes lineage operation, and axillary cutaneous fold thickness in this PIDD dataset can be measured without the need for sophisticated laboratory equipment. However, the measurement of the other two features in the dataset, namely Blood Glucose Levels and Insulin Dose Taken, necessitates the use of specific instruments or equipment. The 2-hour serum insulin test and oral glucose tolerance test (OGTT) involve initially collecting blood samples from the individuals, followed by administering a specific quantity of glucose solution for consumption. Subsequently, blood specimens are collected from the individuals at intervals of 30 to 60 minutes.

In addition to measuring blood glucose levels and insulin doses, it is important to note the challenges associated with doing diabetic pedigree function measurements. According to the initial description of the PIDD dataset, this characteristic is derived by computing the probability of having diabetes based on one's family history [56]. Initially, in order to carry out this computation, it is necessary to obtain information regarding the family members of the subjects. However, it should be noted that this process is quite time-consuming. Furthermore, the outputs of the diabetes pedigree function measurement vary depending on the specific diabetes pedigree algorithms used, indicating the absence of a universally accepted norm.

PIDD is characterized by a sufficient number of recorded instances and may be easily included into any learning model, without the need for extensive data pre-processing. Thus, PIDD is extensively utilized in machine learning and deep learning techniques for the purpose of detecting diabetes. In addition to the widely used PIDD database, different databases are utilized for training built detection algorithms. Quan et al. [71] developed a prediction model utilizing the database from Luzhou, China. The database contains records of 14 medical examination characteristics for a total of 137,998 samples, consisting of 68,994 healthy samples and 68,994 diabetes samples [71]. Subsequently, these trained models are subjected to testing and validation using an additional dataset consisting of 13,700 samples and the identical set of characteristics. Quan el al. utilized the PIDD dataset to verify the accuracy of the developed detection models. The Luzhou, China database contains 14 recorded features, which include Age, pulse rate, breathing, left systolic pressure (LSP), right systolic pressure (RSP), left diastolic pressure (LDP), right diastolic pressure (RDP), height, weight, appearance index, fasting glucose, waistline, low-density lipoprotein (LDL), and high-density lipoprotein (HDL). In order to analyze this dataset, measures of systolic and diastolic pressure necessitate the use of an a breathing device [59], while both high and low cholesterol levels require a blood test known as a lipoprotein panel.

Exploring the efficacy of algorithms for classification based on datasets without invasive laboratory measurements is interesting, as both the PID and Luzhou, China datasets depend on these measurements. The study direction is essential because of the progress made in the medical area, where precise categorization of diabetes outcomes can now be accomplished utilizing the previously mentioned conventional methodologies and datasets. Unlike previous datasets, there is a present need for a diabetic classification tool that is both cost-effective and time-efficient. It is indisputable that datasets play a crucial part in meeting this demand. By utilizing non-intrusive datasets and employing machine learning and deep learning methods, there is the potential to create the necessary categorization tool. Consequently, numerous studies have prioritized the use of non-invasive datasets instead of traditional invasive datasets for their research.

In the study conducted by Swapna et al. [58], cardiac electrocardiograms of 40 individuals were recorded for a duration of 10 minutes while they were in a reclined and relaxed position. Out of the 40 participants, 20 had diabetes while the other 20 were considered healthy [58]. On the other hand, Sandhiya et al. [51] utilized a dataset that was acquired for the purpose of their research.

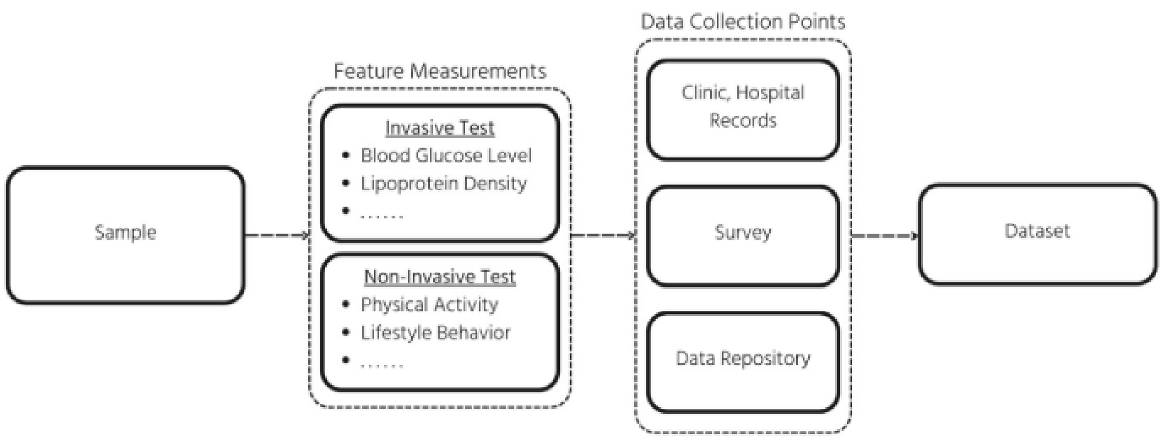


Figure 1: An extensive summary of collecting information

from the Machine Learning Research Archive , Department of Informatics and Computer Science (ICI). This diabetic collection captured information on both lab and non-lab test aspects, including insulin dosages administered, blood sugar levels collected prior to and following meals, the quantity of meals consumed, and the amount of exercise activity samples engaged in [51].

The dataset used by Anna et al. [9] to train their diabetes prediction algorithms includes non-labtest-based medical variables for 451 children between the ages of 6 and 18. Age, sex, weight, height, existence of Type 1 Diabetes, and physical exercise records, including step counts and the amount of time of sitting down, as well as moderate, light, and strenuous activity per month, are among the parameters that are documented [9]. The samples are fitted with accelerometers and pedometers to measure the duration and intensity of physical activity. The goal of the research using this dataset is to address a number of issues that children have during lab-test-based diabetes detection testing, such as needle phobia. Their objective is to investigate the dependability of diabetes detection algorithms when non-lab-test-based datasets are used for training.

Training data was used from the UCI Diabetic Repository by Vidhya et al. [62]. This dataset documents associated lifestyle practices and health indicators of samples, such as: age, drinking alcohol, binge-eating late at night, eating habit, blood glucose level, smoking habit, food habit, conducting regular exercise, gender, and the presence of diabetes in the family the course of time.

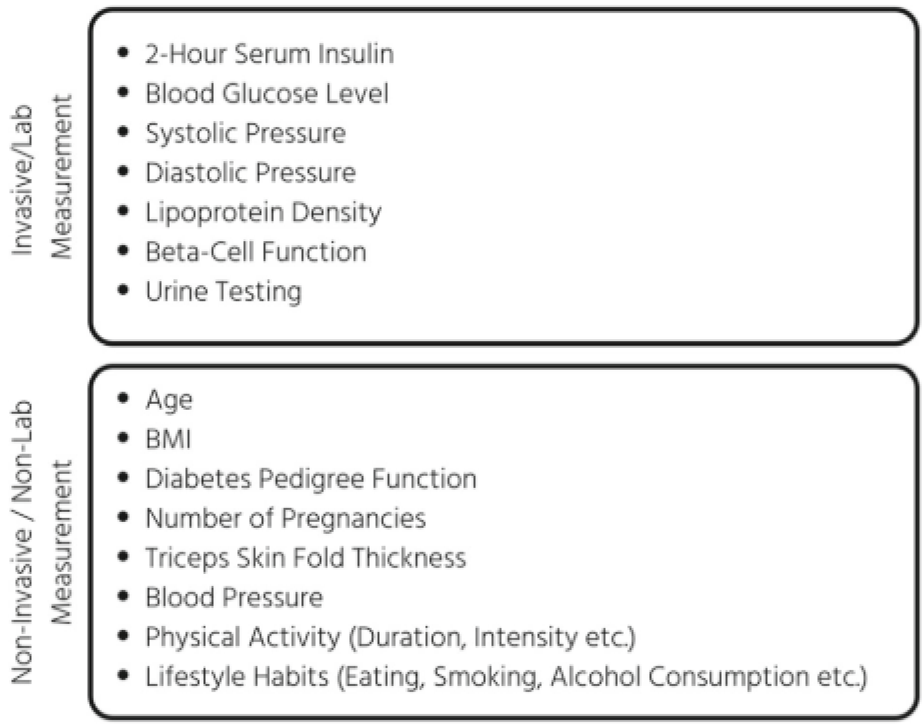


Figure 2: Well-liked characteristics derived for a system for detection based on deep learning and neural networks

Fig. 3 illustrates a pair of intrusive measures used in the diabetes categorization task: fingerstick gauges (right) and collecting blood (left).

Rani et al. [45] utilized a dataset that documented various medical indicators including age, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, which visual blurring, itchiness, irritability, postponed recuperation, biased paresis, muscle stiffness, alopecia, obesity, and the presence of diabetes in the samples [45]. The data is recorded as an analog value, which represents a binary choice between YES or NO, rather than using precise numerical values from medical tests. It is important to note that these signs can be detected by examining everyday living routines or physical appearances without the need for any laboratory testing, comparable to the research conducted by Anna et al. [9].

In most current diabetes diagnosis algorithms, the features used as input are invasive and necessitate costly equipment. Additionally, the techniques for collecting the data are laborious and intricate. Simultaneously, there is limited research that has demonstrated the potential use of basic physical characteristics and lifestyle factors, such as the duration of physical activity and the presence of smoking behavior, in the development of a diabetes detection model [9, 62]. However, the reliability of these factors has not yet been verified. Figure 1 and Fig. 2 classified numerous prevalent features employed in prior study within this domain. Figures 3 and 4 displayed many standard measurements used in the task of classifying diabetes. It is evident that intrusive measurements often need the extraction of the patient's blood, and patients are typically required to fast before the measurement.

Figure 5 depicts a pie chart that illustrates the distribution of datasets utilized in all the papers that were reviewed. Out of the 101 classification models analyzed in the reviewed publications, only 12 of them were trained using non-invasive datasets. This figure provides additional evidence that the potential of constructing a classification model using non-invasive datasets has been largely overlooked by most individuals. However, this should not be the case, as such an approach offers numerous advantages over conventional methods, including lower costs, greater accessibility, improved patient comfort and safety, and easier implementation in real-world scenarios. These benefits can be realized if the successful development and validation of such a model is achieved.

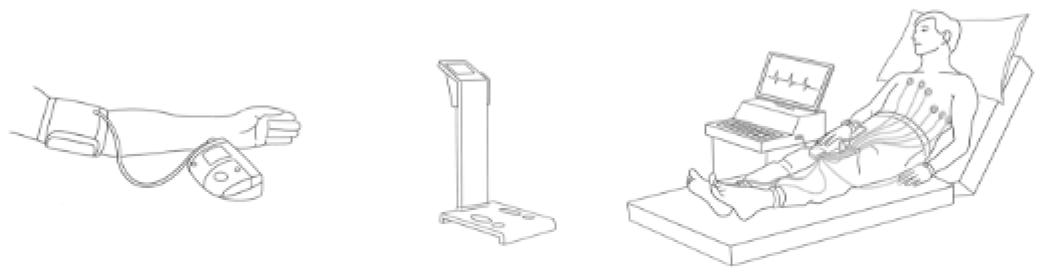


Fig. 4: Multiple channels of ecg machine (right), kern ical scales for mass measurement (center), and sphygmomanometer for arterial pressure measure (left) are instances of non-invasive measurements in the diabetes categorization assignment.

3.0 Data pre-processing

Although there are actually various types of datasets used to solve this problem, it is indisputable that the majority of them do not fulfil the quality requirements for training machine learning and deep learning models. This is due to the following reasons: Initially, it is inevitable that numerous databases include incomplete or inaccurate information during their construction. The presence of such data points can have an impact on the performance of the models, particularly when dealing with medical or healthcare jobs. It is important to avoid this. Furthermore, it is important to note that the features captured in the dataset may not exhibit any correlation with the goal diabetes outcome. Consequently, include these features to train the classification models would not only negatively impact the models' performance, but also result in an increase in computing cost and time required. In addition to these factors, datasets may also differ in terms of sizes, units, and distributions. This variability might result in the dominance of specific features throughout the learning process, resulting to inaccurate and biassed comparisons between distinct features. Ensuring class balance is a crucial step when creating an ideal dataset, as it helps avoid the model from being biassed towards the dominant class. Hence, prior to training the model with the dataset, it is necessary to carry out data pre-processing methods such as data imputation, feature selection, data normalisation, and class balancing in order to address the above problems. In addition, the encoding of data must be performed based on the type of the dataset to enable the model to successfully absorb and comprehend the categories of data.

3.1 Data denoise and imputation

Although PIDD [14] has numerous benefits, there are a few specific concerns when employing them for machine learning or deep learning applications. The PIDD dataset exclusively contains data on female samples, together with the corresponding number of pregnancies for each individual. Consequently, the models developed using the PIDD dataset may not be suitable for guys. The PIDD dataset contains a significant amount of missing or aberrant recorded data, necessitating data filtering to prevent biassed and erroneous predictions. However, removing these atypical data points from the dataset could lead to a reduction in the number of samples, perhaps resulting in an inadequate representation of the true distribution of data within a large population. Therefore, the model will not yield satisfactory outcomes upon its public release.

According to reference [70], it was noted that a significant drawback of PIDD is its limited amount of recorded samples and features. The problematic generalisation of machine learning and deep learning models arises as a result of this issue. In addition, the authors also highlighted that addressing data variability, data quality, feature processing, and result interpretability are crucial challenges in developing diabetes detection models. Therefore, both the data pre-processing and feature selection stages are crucial in building a prediction model that can attain precise performance.

As previously stated, the PIDD dataset contains missing or aberrant data. The Triceps Thickness Fold class has a total of 30% missing data, while the Insulin Dose Class has 49% missing data. The data is recorded as zeros. Given its significant proportion within the dataset, it is imperative not to overlook this topic, as it would have a direct impact on the accuracy of the categorization.

To address the issue of missing data [1, 17, 29, 47, 52], the vacant values were replaced with the mean value of the corresponding class. This technique is popular because it guarantees the continuation of the dataset. However, the data that has been filled in cannot accurately represent the true distribution. As a result, the classification will be erroneous when additional samples that are not included in the dataset are used for classification in the model.

An alternative method to address this problem is to exclude any recorded samples that include missing data [29, 44, 71]. As previously stated, this step will additionally reduce the quantity of available samples in the dataset. The accuracy of the boundary between diabetic and non-diabetic output will be compromised if the models are trained using such dataset.

In their work, Santosh et al. [29] investigated the impact of various treatment approaches on the performance of a model using the PIDD dataset. All classes in the dataset are used to train the models in this research. Subsequently, writers employed three distinct methodologies to address missing data: sample removal, substitution of missing values with mean values, and substitution with zero. Consequently, during the testing phase, the strategy of eliminating samples with missing data yielded the highest accuracy compared to the other two strategies, namely replacing with the mean value and replacing with zero. The result suggests that substituting missing data with the mean value will partially mitigate the bias problem caused by the missing data. While the best accuracy was attained in this research by deleting samples with incomplete data, it remains uncertain how the models would perform when new datasets are included.

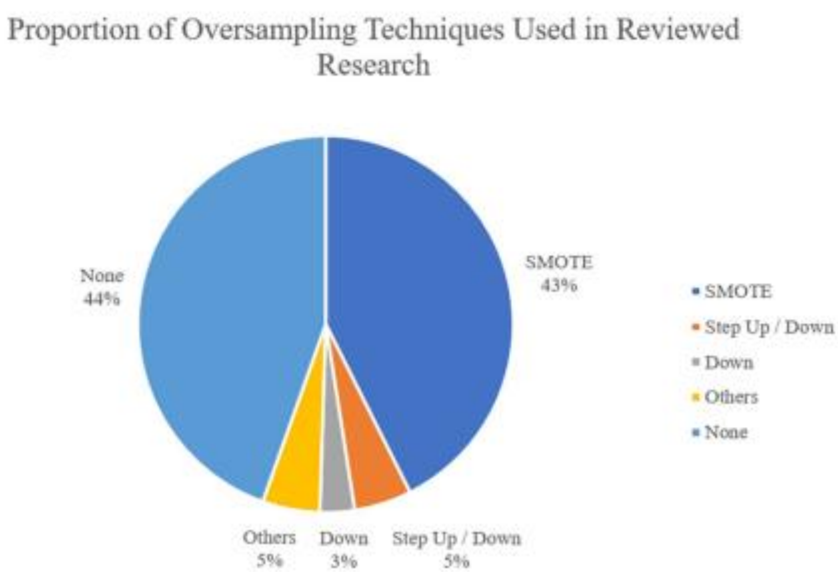


Fig. 7 Proportion of oversampling techniques implemented in reviewed papers

In summary, researchers have utilised both data imputation and missing data removal techniques to address the issue of missing data in datasets. Although both strategies have their own merits, the choice of solution ultimately depends on the characteristics of the dataset. If a small dataset like PIDD is chosen, it is advisable to use data imputation to fill in missing values. Removing more data will only reduce the number of available data points, which is particularly important for training deep learning-based classification models. Conversely, if a large dataset is chosen and the number of data points is not a concern, it is preferable to eliminate small noises in order to preserve the original data distribution. However, although it is acknowledged that eliminating noises from a short dataset can result in unsuccessful training because to the limited amount of data, oversampling the dataset can be a useful solution to address this problem.

In order to expand the number of available classes and samples, Maria et al. [17] employed a technique called Data Augmentation to enhance the PIDD dataset. The primary objective of this technique is to conduct oversampling, which involves augmenting the amount of elements in datasets. This allows the detection models to have a greater number of samples for training. The use of Sparse Autoencoder (SAE) results in the expansion of the original PIDD dataset, which initially consists of 8 characteristics (excluding the existence of diabetes in samples), to a dataset with 400 features. This expanded dataset is then transformed into a matrix with dimensions of 20 by 20. The Variational Autoencoder (VAR) is utilised to augment the number of samples to 449 for class 0 (non-diabetic) and 484 for class 1 (diabetic), effectively addressing the problems of insufficient sample size and imbalanced distribution in the dataset. Prior to feeding the dataset to the detection models, it is crucial to address issues such as missing data or outliers in order to achieve a perfect prediction model.

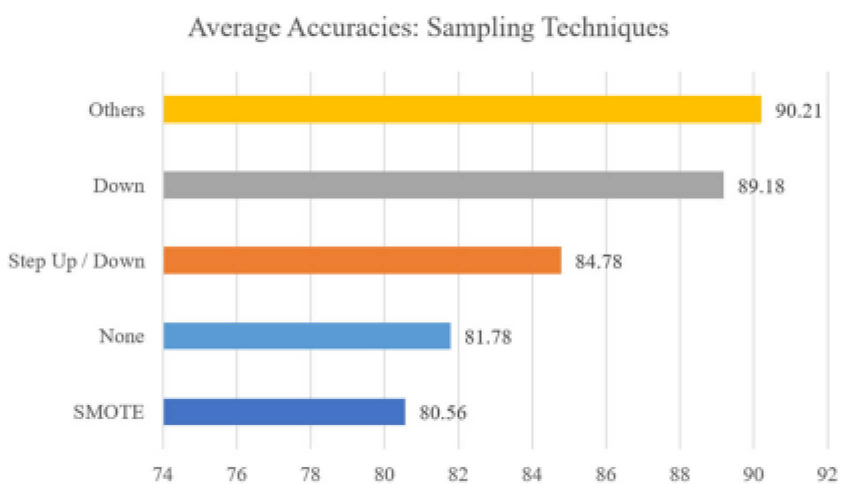


Fig. 8 Average accuracies oversampling techniques implemented

As depicted in Figures 7 and 8, indiscriminately oversampling the datasets would alone result in detrimental effects on the performance of the models. When comparing the proportion and average accuracies of models implemented with no oversampling technique and SMOTE, it is observed that both have a comparable proportion. Specifically, 44% of the models did not implement any oversampling technique, while 43% were implemented with SMOTE. Despite having a larger number of training data, models built using SMOTE achieved a lower average accuracy in performing the diabetes classification test. Several factors contribute to this outcome, with the quality of the dataset being the most crucial factor. Many researchers did not do any data pre-processing before implementing SMOTE oversampling. This resulted in the SMOTE method generating inaccurate data points due to a dataset packed with noise. Consequently, the model was trained incorrectly. Prior to feeding the dataset to the detection models, it is crucial to address issues such as missing data or outliers in order to achieve a perfect prediction model (Fig. 9).

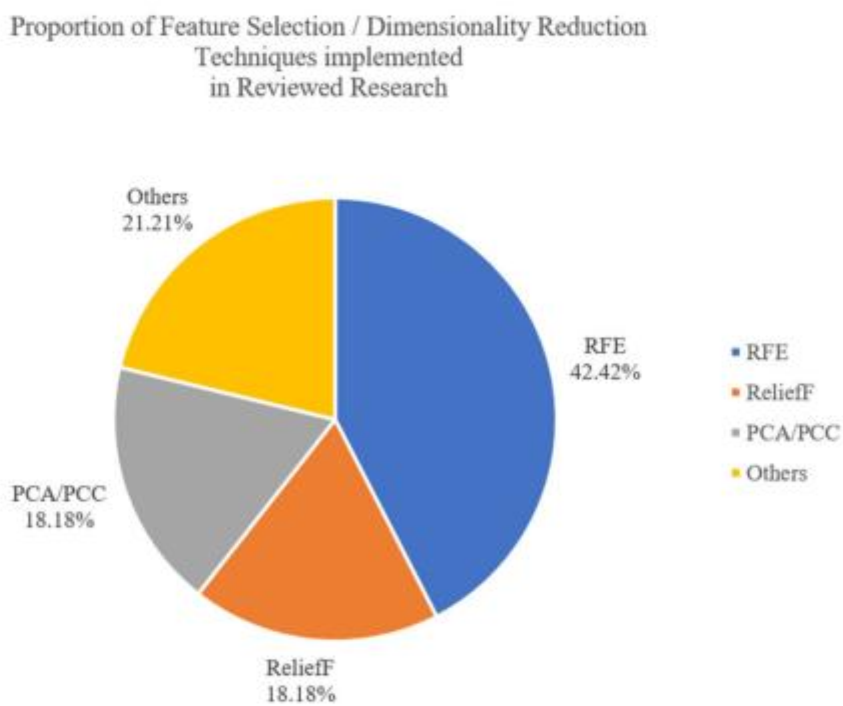


Fig. 9 Proportion of feature selection techniques implemented

In addition, several researchers choose for utilising machine learning-based algorithms to oversample their chosen datasets. As depicted in Figure 10, the models classified as "Others" had the highest average accuracy of 90.21%. This form of oversampling approach has the ability to provide more generalised and accurate data points. Although the utilisation of such technique may really result in an increase in the necessary processing time and cost, feature selection might help alleviate this problem.

3.2 Data normalization

Values of several reported classes in every medical database exhibit significant fluctuations. If two classes with significantly different data ranges are inputted, the machine will prioritise the characteristic with higher values as having a more substantial impact on the outcome. Not only will it impact the accuracy of the model, but it will also complicate the interpretation and analysis of the association between recorded features in the database. To enhance convenience, numerous studies employed the min-max normalisation method to scale each class in the dataset. Min-max normalisation is a widely used technique for normalising data. Its purpose is to convert all features so that they have a minimum value of 0 and a maximum value of 1. The min-max normalisation approach is susceptible to outliers, which are unusual extreme situations that can significantly affect the results. Equation (1) represents the mathematical formula for min-max normalisation.

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

In order to remove any potential abnormal data points in the dataset, a study conducted by Nadeem et al. [39] recommended using the Z-score normalisation method as an appropriate approach for this purpose. While Z-score normalisation can effectively handle outliers in a dataset, it does not result in an output where all features have the same scale. In contrast, min-max normalisation guarantees that all features are scaled between 0 and 1. Equation (2) is the mathematical representation of Z-score normalisation.

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

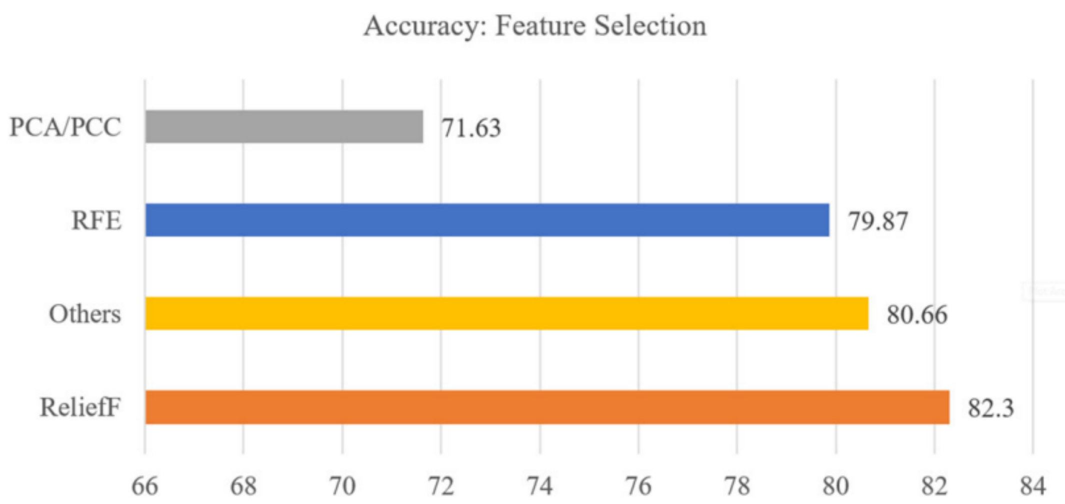


Fig. 10 Average accuracies of feature selection techniques implemented

The problem of high-dimensional characteristics with redundant parts still persists in the collection, despite the normalisation of the data. To reduce both computing time and power consumption, it is crucial to carefully choose significant features based on their functionality, as will be shown in the section that follows.

4 Strategies for Choosing Features and Dimensional Reducing

While hyperglycemia has been associated with the death of Beta-cells and an increase in TNF-alpha expression in the human body [37], the specific medical signs leading to these abnormalities remain uncertain [23]. Diabetes is frequently linked to the following characteristics: Hyperglycemia, obesity, abnormal systolic blood pressure, familial history of diabetes, immune system and genetic defects, dyslipidemia, liver and renal dysfunctions, and several other variables [66]. These may encompass additional physical and clinical information, which, when completely integrated, can result in a significant level of complexity in the implementation of features.

Like in any field of machine learning, it is necessary to remove unnecessary characteristics in order to make the model description as concise as possible. This aligns with the notion of "Occam's Razor" in computation, which asserts that when faced with two models of different complexity, the simpler model and description should be chosen. This may pertain to the complexity of a machine learning model, although it largely pertains to the inclusion of features in the model. In other words, the selection of features aims to minimise the number of features, resulting in an ideal model description. This is particularly apparent in the context of diabetic detection, since the features included are frequently of a higher level. Another correlated method is feature reduction. This strategy, as its name implies, utilises principles to choose projected features or alter features before applying them to machine learning models. In machine learning applications, there are often two popular approaches: feature selection, which aims to choose relevant and distinguishing characteristics, and dimensionality reduction, which involves expanding the feature space via projection methods. Nevertheless, years of implementation have demonstrated that these efforts result in various degrees of progress when put into practice. Choosing highly pertinent characteristics or feature projections in both scenarios might reduce the computational resources needed and lessen the time necessary for the detection procedure. Hence, it is crucial to calculate the connection and relative significance of each feature using certain methods. Within this perspective, one must consider: what are the most effective feature selection algorithms or feature reduction procedures utilised in this particular field? This section discusses two commonly used feature selection techniques: Principal Component Analysis (PCA) and Pearson Correlation Coefficient (PCC). PCA and PCC are frequently employed for dimension reduction and feature selection, respectively. Principal Component Analysis (PCA) is a feature selection approach that is extensively employed in the field of computer science, particularly for datasets with a large number of dimensions. PCA, or Principal Component Analysis, is primarily used to transform samples from their original space to a lower-dimensional representation. This allows for the representation of the data to still capture statistical characteristics, such as the covariance of the original data [57]. Two commonly used methods for reducing the dimensionality of features are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Both methods employ feature projection onto a new axis and then select the projection that captures the highest variances. PCA is commonly employed in image pre-processing and feature dimension reduction because of its resilience in identifying discriminant projections. Song et al. [57] employed Principal Component Analysis (PCA) to decrease the dimensionality of the images in the dataset prior to implementing them in their proposed models. Following the completion of feature selection, the dimensions of the photos are reduced to 40%-70% of their original size. The results indicate that the processed photographs obtain higher accuracy compared to the untreated images.

However, it is important to remember that the attribute reduction/feature picking might have negative consequences and therefore should be carefully considered on a case-by-case basis when used. In such instances, it is necessary to randomly divide the train and test data. The case study reported in Sivaranjani et al. [55] demonstrates that the PCA feature selection method resulted in marginal enhancements to the SVM-based model. The performance of the RF-based model declined following dimensionality reduction. Authors have highlighted that the reduction in dimensionality has led to a drop in the amount of data available in the PIDD dataset. Consequently, this limitation hinders the ability to make generalisations. Francesco et al. [38] found similar results to their research, showing that their machine learning models performed similarly. The PCA algorithm only provided small improvements to most of the models and actually lowered the performance of the RF-based model once PCA was introduced. In summary, the average accuracy of all models implemented with feature selection methods is 81.16%. Machine learning models achieved an average accuracy of 80.31%, while deep learning models earned an average accuracy of 84.51%. From the two figures provided, it is evident that both the PCA / PCC and RFE feature selection algorithms account for approximately 60% of the total proportion. However, these two algorithms have the lowest average accuracies compared to other feature selection algorithms. Specifically, the PCA / PCC algorithm achieved an average accuracy of 71.63%, while the RFE algorithm achieved an average accuracy of 79.87%. Based on the average accuracies of less than 80%, it can be inferred that not all feature selection techniques are appropriate for addressing the diabetes classification challenge. Introducing such algorithms may have negative effects on the performance of the models. Figure 11 effectively demonstrates the beneficial effects on machine learning models when suitable feature selection methods are applied. It should be noted that the three machine learning algorithms shown are the most common algorithms in all of the reviewed research publications.

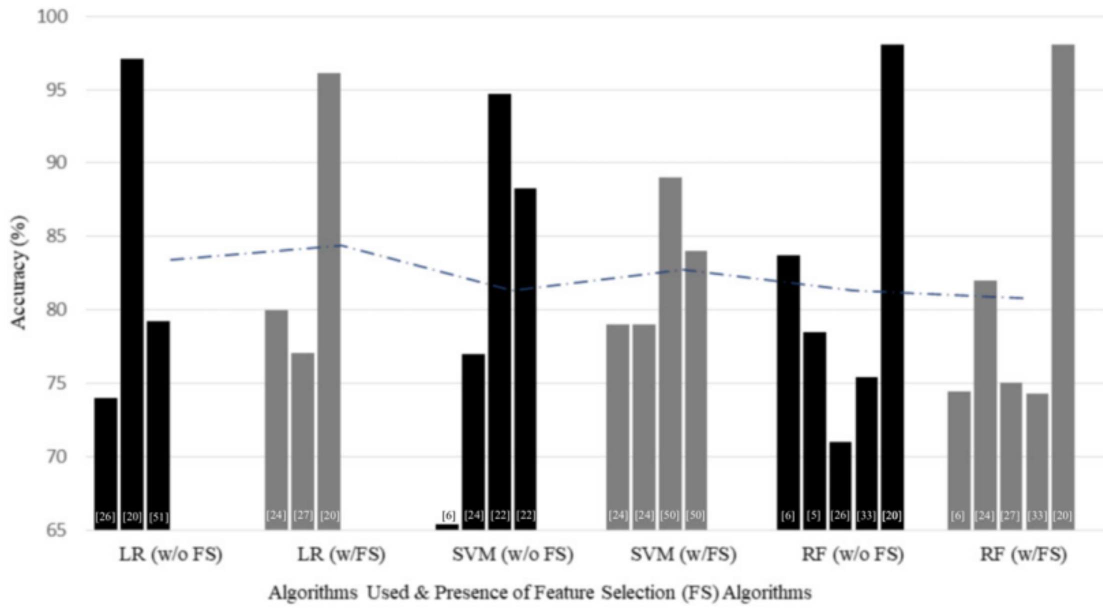


Fig. 11 Impact of feature selection on ml models

4.1 Feature selection

Most feature selection procedures employ a specific sort of feature assessment and scoring system. Less sophisticated methods involve systematically setting up a smaller collection of characteristics and then assessing them. The drawback of approach is that the fitness of combined feature sets is predominantly non-linear. Thus, employing a "greedy" technique for sequential selection may not result in optimal outcomes. Another effective strategy is assessing the similarities between features. An example of such an algorithm is PCC [53], which is widely used to calculate the correlation between two variables inside a single dataset. In numerous studies within this field, researchers frequently employ this approach to choose the relevant variables that exhibit a robust link with the diabetes result in the samples. In the study conducted by Nour et al. [1], all the correlations between each variable in the dataset were calculated in order to examine their individual statistical associations with one another.

Zhou et al. [69] utilised a dataset of 141 variables from 9765 samples, encompassing 5 categories of diabetic problems. Their study aimed to identify samples with distinct diabetic complications by analysing the 141 medical variables documented in the dataset. Extracting all variables in the dataset and studying their relative impact in detecting every diabetic complication is not feasible. Therefore, prior to inputting the data into the models, they applied the Pearson correlation coefficient (PCC) and calculated the correlation between each indicator and the occurrence of diabetic problems. For each diabetic complication, the researchers selected the top 10 indicators that had the highest correlation values according to the PCC method. As a result, all four models utilised in this research had excellent outcomes in accurately diagnosing every type of diabetes complications. This suggests that feature selection using PCC is viable in this particular field. Lukmanto et al. [34] utilised Fuzzy Support Vector Machine to analyse the PIMA Indian dataset and examined the feature selection method to assess the enhancements obtained. The authors emphasised that the results demonstrate a highly encouraging accuracy rate of 89.02% in predicting patients with DM through the utilisation of feature selection methodologies employing F-Score Feature Selection. This method is partially associated with Pearson's correlation coefficient (PCC), where correlation scores between characteristics were used to determine deletion and feature reduction. Hou et al. [21] utilised the fisher score method to assess a dataset consisting of 19 features. The accuracy, precision, sensitivity, F1 score, Matthews correlation coefficient (MCC), and area under the curve (AUC) were calculated. The results demonstrate that their approach can effectively be utilised for feature selection in a diabetes classifier, leading to enhanced performance. This will offer valuable assistance to physicians in promptly identifying cases of diabetes. The purpose of feature selection in data-driven solutions is twofold: to enhance the performance of the model and to analyse the impact of each feature on the outcome class. For example, in a diabetes classification task, feature selection can provide valuable insights to medical professionals by examining the correlation between the target feature and the diabetes outcome from a data science perspective. Fig.11 demonstrates the impact of feature selection on the accuracy of diabetes detection machine learning models (LR, SVM, RF), in contrast to models without feature selection. The data were gathered from multiple research articles that employ feature selection in their methodologies. The Blue line represents the mean accuracies produced by all the models before and after feature selections are conducted. The difference demonstrates the impact of feature selection. Most studies have shown that models with feature selection consistently outperform models without feature selection. Nevertheless, this literature review demonstrates that even without feature selection, deep learning-based models outperform machine learning-based models with feature selection, as depicted in Figure 12. However, it is worth noting that positive feedback can be achieved by implementing appropriate feature selection algorithms on the dataset. In fact, additional evidence indicates that the models performed worse once feature selection was used on the dataset. Although numerous cases have demonstrated the importance of feature selection in optimising the performance of models, many researchers have failed to recognise that the outcomes of feature selection are invariably linked to the quantity of available samples. If the initial dataset size is already limited, reducing its dimensionality more will simply result in a drop in the amount of available data. As a result, training models will be less effective. Fig. 12 clearly demonstrates that the implementation of feature selection with deep learning algorithms leads to a significant fall in average accuracies due to the reduced amount of accessible training data. While research has demonstrated the significance of feature selection in reducing computing load, it is equally vital to address the potential drawbacks by adding a larger dataset of higher quality. Although high-quality huge datasets are often difficult to come by, oversampling techniques, as mentioned in the previous section, are always effective in reducing the negative impacts caused by feature selection. Based on the literature study, it is important to highlight that feature selection and dimension reduction should be carried out with careful examination, despite the fact that many application areas consider these techniques as standard in machine learning. Do the dimensions of the features necessitate feature selection? Do the facts provide enough information to make generalisations? These factors are crucial and, thus, the conclusion ultimately hinges on the facts obtained.

5 Deep learning and machine-learning algorithms.

Machine learning and deep learning algorithms are commonly used in the majority of data-driven diabetes categorization tasks. Prior to examining the advantages and disadvantages of each model, it is crucial to have a clear comprehension of how the models are assessed in this particular work. In the context of categorising binary data, a confusion matrix can be used to assess the effectiveness of the model.

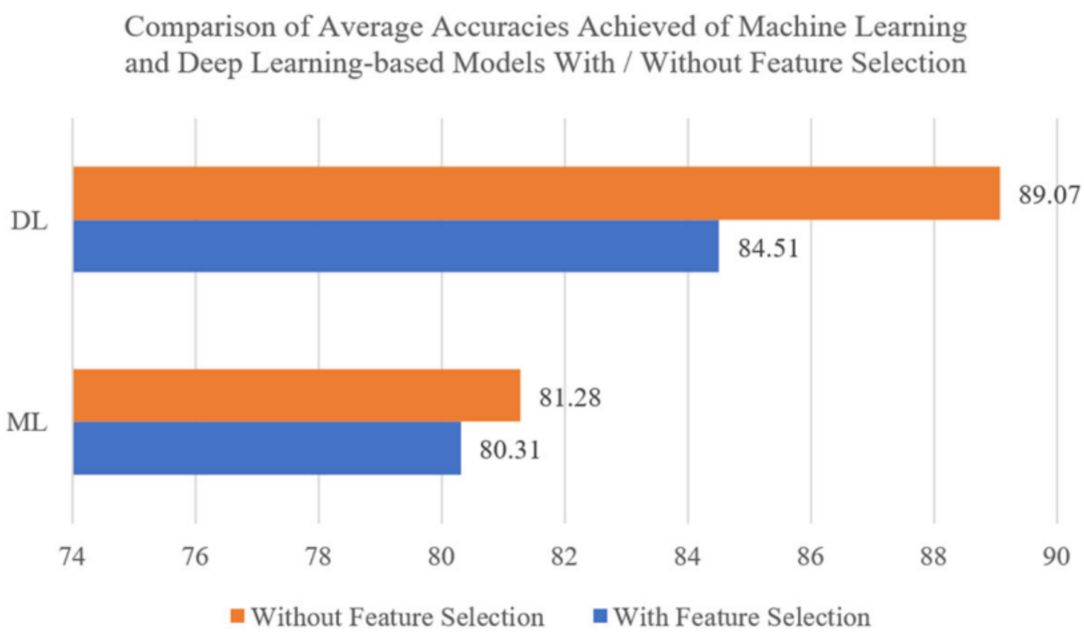


Fig. 12 Comparison of models' accuracies before and after feature selection

The confusion matrix provides a summary of the anticipated and actual classes, enabling researchers to compute several assessment metrics that offer insights into the performance of the model. Analysing these results is crucial for identifying the model's threshold and fine-tuning hyperparameters to obtain optimal performance. A confusion matrix presents the count of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for a given set of predictions in comparison to their actual ground truth labels, as depicted in Figure 13. Therefore, the evaluation metrics listed below can be calculated using the results: Accuracy refers to the ratio of accurate predictions to the total number of predictions produced. The calculation is performed as

$$Accuracy(Acc) = \frac{TP+TN}{TP+TN+FP+FN}.$$

Precision refers to the ratio of correctly identified positive instances to the total number of positive predictions generated by the model. The calculation is performed as

$$Precision(Pr) = \frac{TP}{TP+FP}.$$

Recoll: The ratio of correctly identified positive samples to all positive samples in the dataset. The calculation is performed as

$$Recall(Re) = \frac{TP}{TP+FN}.$$

5.1 Approaches for Diabetes Diagnosis Using Machine Learning Techniques

Machine learning encompasses the ability to acquire knowledge and identify patterns using statistical models and deep learning methods [11]. All of these accomplishments are made possible by including linearly sophisticated statistical methods into the models [10]. This technique has demonstrated efficacy and dependability in numerous diabetes detection solutions. The integration of machine learning models with other helpful algorithms like PCA and PCC has been a popular method due to the potential for enhanced possibilities. Scientists can investigate the hyperparameters of the created models, allowing them to utilise these models to solve complicated nonlinear problems in real-world applications. Hyperparameters in machine learning models refer to the intrinsic characteristics of the machine learning algorithms. These include the model architecture, learning rates, number of epochs (iterations), number of branches (specific to Decision Tree and Random Forest), number of feature clusters, and other related factors. Prior to commencing the training and validation process, these parameters are predetermined [65].

Amani et al. [64] utilised two machine learning models and one deep learning model to train on the PIDD dataset. The machine learning models utilised in this research are Support Vector Machine (SVM) and Random Forest (RF). The Support Vector Machine (SVM) algorithm is widely used to determine decision boundaries of linear functions after conducting data analysis [19]. On the other hand, Random Forest (RF) is another algorithm that constructs Decision Trees (DT) on various samples in a random manner [3]. Each DT follows a structure similar to a flowchart diagram, consisting of multiple nodes. Every node in the decision tree (DT) represents a categorization rule determined by the algorithm. Extending from the node, the branches ultimately connect to the leaf nodes, with each leaf node representing the decision result of this iteration [3]. Due to the inability of native SVM models to solve nonlinear functions, the authors utilised a kernel function called Radial Basis Function (RBF) to transform the nonlinear functional into a linear space. This transformation allows for simpler separation of the data. The Support Vector Machine (SVM) model achieved an accuracy of 73.94%, while the Random Forest (RF) model achieved an accuracy of 79.26%.

Quan et al. [71] conducted a study in which they examined the performance of machine learning and deep learning models for predicting diabetes. They specifically used J48 and RF algorithms for this comparison. J48 is a classifier that uses statistical methods to generate a decision tree based on input data [6]. Their study utilised Primary Immunodeficiency Diseases (PIDD) and a dataset documenting hospital physical examination data in Luzhou, China. Both algorithms concluded that the blood glucose level is the most influential factor in the function of detecting diabetes. Therefore, an additional test was performed on both datasets to evaluate the impact of excluding blood glucose values and utilising only blood glucose levels for training on the accuracy of the models. The RF and J48 models achieved average accuracies of 72.59% and 75.19% respectively in both the PIDD and Luzhou Diabetes Dataset. A notable discovery in this research paper is that when conducting PCA feature selection on the PID dataset, it resulted in a decrease in accuracy for all models used in the study. However, this issue was not observed when using the mRMR feature selection algorithm on the same dataset. On the other hand, when incorporating the Luzhou, China dataset, all feature selection algorithms led to a decline in the accuracy of the models. This suggests that not all feature selection algorithms are effective in enhancing the performance of models. This could explain why the RF model trained with the PID dataset in this study performed worse than the one mentioned in the earlier research by Amani et al. [64]. In their research, they opted to use all features without incorporating any feature selection algorithm. It is important to mention that in the study conducted by Quan et al. [71], they did not use oversampling to address the negative effects of feature selection.

In their work, Nour Abdulhadi et al. [1] conducted a comparison of six machine learning models: Random Forest (RF), Logistic Regression (LR), Voting Classifier, Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), and polynomial SVM. Logistic Regression (LR) is a classification method that calculates the probability of the output. It presupposes that each variable is independent of the others [13]. On the other hand, a Voting classifier is a technique that combines multiple basic classification models. The final decision output is determined by the predictions and results of the classifiers [30]. Like LR, LDA is effective in solving complex functions when there are more than two output classes in the dataset [61]. RF obtained the highest accuracy rate of 82%, while Logistic Regression and Voting Classifier achieved an accuracy rate of 80%. Linear Discriminant Analysis (LDA), Polynomial SVM, and SVM all achieved an identical accuracy rate of 79%. In contrast to the technique taken by Amani et al. [64], Nour Abdulhadi et al. filled the missing data in their implemented PID dataset with the mean values of the corresponding class. This is to prevent any further reduction in the amount of training data accessible in the dataset. In addition, they made the decision to incorporate all attributes when training the created models. Therefore, Nour Abdulhadi et al. models outperformed the models of Amani et al. [64] and Quan et al. [71] in terms of accuracy. This is because Nour Abdulhadi et al. models were trained with a larger dataset and more features from the PID dataset. Muhammed et al. [52] conducted their study by employing six machine learning methods, including SVM, K-Nearest Neighbour (KNN), Logistic Regression (LR), Naive Bayes (NB), DT, and RF, to develop diabetes diagnosis models. KNN achieves classification by calculating the proportion of samples from different classes surrounding the test subject in the data distribution, where "K" represents the number of these samples [60]. NB is a method that utilises Bayesian's Theorem to classify data based on the likelihood of attributes. The assumption made in the prediction procedure is that the feature being considered is statistically independent from the other features [46]. The authors of the study emphasised that in order to attain higher accuracy, a database with a larger number of samples and no missing values is necessary. In order to preprocess the data, the authors substituted all missing data with the mean values and used min-max normalisation to the refined dataset. The results indicate that the KNN and SVM models outperformed the other six machine learning models, with an accuracy of 77% in the trial. The LR and NB models followed closely behind with an accuracy of 74%. Lastly, the DT and RF-based models had the lowest accuracy of 71% in this study. When analysing the PID dataset, Muhammed et al. and Nour Abdulhadi et al. utilised a comparable method. The SVM-based models they developed yielded similar accuracies, with Muhammed et al. achieving 77% accuracy and Nour Abdulhadi et al. achieving 79% accuracy. However, there was a significant disparity in the performance of their RF-based models, with a difference of 11% in achieved accuracy. In this research, the tree-based DT model likewise demonstrated a relatively low accuracy of 71%. The factors contributing to the variations in performance of decision tree-based models will be addressed in the subsequent paragraph. In the research conducted by Amin et al. [20], they utilised different feature selection techniques, including Random Forest (RF) and Iterative Dichotomiser 3 (ID3) procedures, to create their decision tree (DT) models. The proposed DT+(DT-ID3) model attained the highest accuracy among all three techniques, with a 99% accuracy rate. The decision tree (DT) model achieved a high accuracy of 98.2% when trained on the dataset with all its characteristics extracted. Unlike the previous two studies, Amin et al.'s model, which is also based on decision trees, has the potential to outperform a model based on neural networks. Claiming that this conclusion is acceptable and can be easily applied to real-life problem-solving tasks is less persuasive, given that the model was trained using a low-quality PID dataset without oversampling and imputation. This suspicion arises due to the widespread occurrence of overfitting in many tree-based models when they are not appropriately calibrated. An excessive number of hyperparameters, such as the number of branches and maximum depth, can lead to unneeded complexity in models. This can result in the models capturing every minute aspect in the dataset, including extraneous noise and patterns, which in turn causes overfitting. It has been observed that authors seldom address and examine the problem of overfitting in models, and this issue should not be overlooked in any pertinent research.

The researchers in the study conducted by Sajratul et al. [47] utilised the k-mean algorithm with Greedy Stepwise Search to improve the process of feature selection in the PIDD dataset. The implementation of the Greedy Stepwise Search determined that a feature subset consisting of the number of pregnancies, blood glucose level, BMI, age, diabetes pedigree function, and k-mean cluster results in the lowest error for the outcome. The LR and RF-based models utilise these properties, resulting in accuracy rates of 77.08% and 75% respectively. In addition to utilising the Greedy Stepwise Search approach for feature analysis, they also incorporate a standard statistical approach by examining the histogram of features in relation to the diabetic result. Their research revealed a significant rise in the risk of diabetes when both glucose and B.M.I. levels in the PID dataset increased. Conversely, the opposite effect was observed in relation to Skinfold Thickness. After conducting a parallel verification with the results of Greedy Stepwise Search, any undesirable characteristics were eliminated. Consequently, the computational expense was greatly reduced, while still keeping satisfactory accuracy. This study has demonstrated that in addition to establishing specific algorithms for feature selection, researchers need also examine the outcomes of these algorithms using various approaches. This step is essential in constructing a persuasive decision support tool based on data analysis.

Huma et al. [42] suggested various sampling techniques, including Linear Sampling, Shuffled Sampling, Stratified Sampling, and Automatic Sampling, for feature extraction. The extracted characteristics are inputted into the Naive Bayes (NB) and Decision Tree (DT) models, resulting in accuracy rates of 76.33% and 86.62% respectively. In contrast to the above mentioned studies, Huma et al. applied oversampling techniques to the PID dataset prior to utilising it for model training. Their paper also emphasised the importance of pre-pruning in the construction of a tree-based forecasting tool. According to the investigation conducted in this paper, the tree-based DT model achieves a high accuracy of 86.62%. However, this result can still be enhanced by implementing data imputation to address the issue of missing data, which is not mentioned in their work. Additionally, feature selection can be employed to eliminate any feature that negatively impacts the model's performance. In their investigation, Francesco et al. [38] examined the features in the PIDD dataset to see whether all of the recorded features are potential risk factors for developing diabetes. Following the process of feature selection, there is only a marginal enhancement observed in all of the models. The highest accuracies obtained by each of the models are as follows: J48: 74.2%, Hoeffding Tree: 77%, Jrip: 75.5%, BayesNet: 74.9%, RF: 75.4%. Among these five machine learning models, it is noteworthy to notice that the RF-based model is the sole model that exhibited reduced accuracy following the implementation of feature selection on the training dataset. Fayroza et al. [26] employed three distinct machine learning models in their study: KNN, NB, and LR. The PIDD dataset underwent min-max normalisation, but no effort was made to address missing data. LR outperformed the other two models when assessed using three separate performance assessment techniques. Table 1 displays the comprehensive performance attained by each model in the paper. Rani et al. [45] developed their proposed model using a Multi-Layer Perceptron (MLP) technique and evaluated its performance by comparing it with models based on Logistic Regression (LR) and Random Forest (RF). Their feature selection techniques involve calculating feature scores in the dataset using a built-in class using a Tree Based Classifier. The top 10 features with the highest scores are then taken from the original dataset for model training and testing. After conducting feature selection, their logistic regression (LR) based model demonstrated a reduced accuracy of 96.153% (compared to 97.115% prior to feature selection). Conversely, the random forest (RF) based model maintained a consistent accuracy of 98.076% throughout the entire trial. In contrast to the PID dataset, the dataset implemented only includes features that are directly associated with the consequences of diabetes infection. Unlike certain features in the PID dataset, such as Triceps Skinfold Thickness and Number of Pregnancies, which have not been confirmed to have a direct relationship with the outcome of diabetes. Prior to generating a dataset, it is essential to do a preliminary research to identify the association between the target attributes and the desired conclusion. It is crucial to take this step since creating a dataset is costly, thus any potential misuse of resources in gathering data should be meticulously avoided. Nadeem et al. [39] developed a fusion model comprising Support Vector Machine (SVM) and Artificial Neural Network (ANN) algorithms for the purpose of diabetes detection. Subsequently, this model is compared to two alternative models, one based on Support Vector Machines (SVM) and another based on Artificial Neural Networks (ANN). They utilised the NHANES and PIDD datasets by merging them together for the purpose of conducting this research. Once data pre-processing is completed on the two datasets, which involves filling missing data with mean values and normalising using the Z-score normalisation approach, the data is then inputted into the machine learning models. The Fusion of SVM-ANN model demonstrated superior performance compared to the other two models in both datasets, with an average accuracy of 94.67%. In contrast, the SVM and ANN models achieved accuracies of 88.3% and 93.63% respectively. The fusion of NHANES and PID datasets has resulted in a dataset consisting of 10,627 records with 8 characteristics, which can be used for training and validation. The performance of the SVM-based model has shown considerable improvements compared to the SVM models discussed in the previously reviewed research. Instead of employing the technique of oversampling, Nadeem et al. made the decision to directly augment the size of the dataset by merging two datasets together

This approach is also a viable strategy for mitigating the scarcity of training and validation data. However, there is still room for advancement in this research. Specifically, in the dataset fusion, only 3,556 samples (33.46%) were classified as "diabetic," while the remaining samples were classified as non-diabetic. Class imbalance issues frequently arise in real-world scenarios. Addressing this problem enables more efficient training of the models. In addition, the authors did not do any feature selection or feature analysis in this research. Given the adequate amount of data, exploring feature selection could potentially improve the performance of the models. This aspect should be further examined. Harleen et al. [24] conducted their study utilising Linear Kernel Support Vector Machines (SVM), Radial Basis Function (RBF) Kernel SVM, K-Nearest Neighbours (KNN), and Multifactor Dimensionality Reduction (MDR). The researchers have also utilised the PIDD dataset in their study. In order to enhance the dataset, statistical approaches were employed to eliminate the outlier, while the missing values were replaced by employing the k-NN imputation algorithm for prediction. In addition, the Boruta wrapper method was employed to identify pertinent and significant characteristics from the dataset. The linear kernel Support Vector Machine (SVM) achieved the highest performance among all models, with an accuracy of 89% reached throughout the testing phase. The accuracy achieved by all five models is presented in Table 1. Gradient Boosting Machine (GBM) is a widely used method in current research for addressing this function. GBMs, or Gradient Boosting Machines, are capable of fitting new algorithm models, known as base classifiers, in response to the variables [41]. Similar algorithms to GBM include Extreme Gradient Boosting (XGBoost) and LightGBM. These three techniques are commonly employed in classification tasks. Leon et al. [27] conducted their study by employing five machine learning models: Language Model (LM), RF, XGBoost, Regularised Generalised Linear Models (Glmnet), and Light Gradient Boosting Machine (LightGBM). The Glmnet algorithm is a software that may be used to fit regression models, including Cox, Poisson, Multinomial, Logistic, and Linear Regression Models [67]. In contrast to these models. The elastic net approach combines the L1 and L2 penalties from the lasso and ridge methods in a linear manner [67]. These models are also utilised to determine the significance of each available characteristic in the dataset. The dataset utilised consists of 111 variables from 27,050 samples. Following feature selection and data pre-processing, 59 features from 3,723 samples are ultimately employed in the machine learning models. The models are utilised to train and evaluate utilising five subsets from the dataset, denoted as T6, T12, T18, T24, and T30. The performance of the five machine learning models employed in this research is summarised in Table 1, using AUC as the validation metric. Lai et al. (31) employed four machine learning models, namely GBM, LR, RF, and Rpart, for the purpose of diabetes detection. When using the PIDD dataset to train the models, the Random Forest (RF) model had the highest performance with an Area Under the Receiver Operating Characteristic (AROC) value of 85.5%. The Gradient Boosting Machine (GBM) model came in second place with an AROC of 85.1%. The LR and Rpart diabetes detection models both achieved AROC (Area Under the Receiver Operating Characteristic) scores of 84.6% and 80.5% respectively. Birjais et al. [7] employed Gradient Boosting Machine (GBM), Logistic Regression (LR), and Naive Bayes (NB) in their study. The PIDD dataset was also implemented by the researchers in this study. In line with the previous study paper, the authors employed KNN imputation to forecast and complete the absent values in the PIDD dataset, aiming to mitigate bias. Despite the absence of feature selection, the machine learning-based diabetes diagnosis models they developed achieved impressive accuracy rates of 86%, 79.2%, and 77% using GBM, LR, and NB correspondingly. In summary, all of the Machine Learning-based models examined have demonstrated strong performance.The research analysed several Machine Learning models and found that they attained an average accuracy of 80.6%. This suggests that machine learning algorithms are effective in handling the problem of diabetes diagnosis and categorization. Out of the 75 machine learning-based models that were examined, LR, SVM, and RF-based models are the most commonly used in this field, accounting for 53% (40 models) of all the models. According to Figure 14, the models based on LR, SVM, and RF produced average accuracies of 79.2%, 81.4%, and 81.9% correspondingly. To accomplish

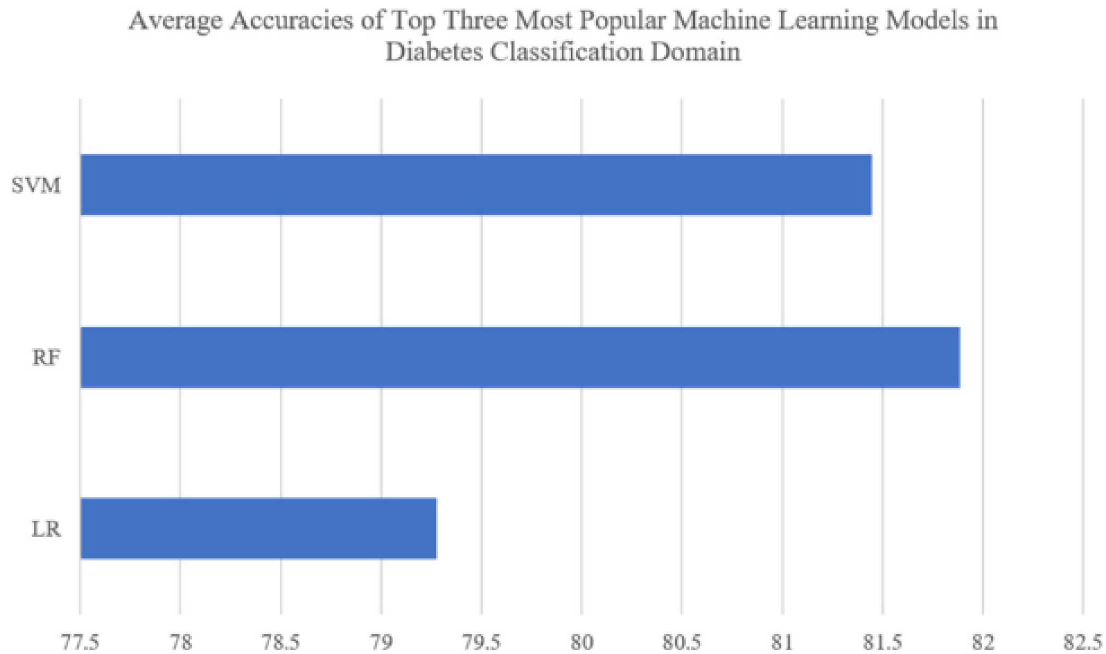


Fig. 14 Average performances (accuracy) of three popular ML-based models

To attain a high degree of precision, it is important to conduct thorough hyperparameter tuning and feature selection. Moreover, the quality of the datasets is essential in achieving these exceptional outcomes. A total of 27 models were trained using the PID dataset, including LR, RF, and SVM models. The average accuracy achieved by these models was 79.2%. Conversely, the remaining 13 models were trained using more extensive datasets of superior quality, resulting in an average accuracy of 83.3%. On the other hand, deep learning models do not necessitate intricate modifications to attain good outcomes, as will be elaborated upon in a subsequent section.

5.2 Models for predicting diabetes using deep learning techniques

Deep learning algorithms aim to replicate the cognitive and learning processes of the human brain. Deep learning algorithms offer several advantages, including the inclusion of built-in features like feature selection and feature extraction [54]. Consequently, running such devices requires less manoeuvres. Deep learning models require a higher volume of data in order to achieve more accurate training results. In addition, the operation of these models necessitates equipment with greater processing capacity than that required by machine learning models [10]. Out of all the research that has been evaluated, the most popular algorithms utilised in the subject of diabetes detection are Convolutional Neural Network (CNN), Deep Belief Network (DBN), and Deep Neural Network (DNN). CNN is widely recognised for its effectiveness in data-driven classification tasks, particularly in the areas of image processing and classification [4, 32]. When input data is provided to the algorithm, it will be transmitted to the Convolutional Layer. In this layer, convolution will be executed to produce a feature map that groups and condenses the information contained in the data [2]. The feature map is subsequently transferred to the Pooling Layer, where the data size is reduced using a selected pooling method [2]. This layer is crucial for reducing the amount of compute power and time needed. Activation functions are present between the Fully Connected Layer and the Output Layer to address classification difficulties. For binary-class classification problems such as diabetes detection, it is typical to use either the Sigmoid or Softmax functions [2]. DBN, however, is a deep learning technique that comprises many restricted Boltzmann machines (RBMs) arranged in a stacked manner [63]. RBMs have the ability to learn the probability distribution of a given dataset. In this configuration of Restricted Boltzmann Machines (RBMs), data is sent into the initial layer of a Deep Belief Network (DBN), and the outputs from this layer are then used as input for the subsequent layer in the stack [63]. A deep neural network (DNN) is created by integrating numerous hidden layers [22]. Within a single hidden layer of the Artificial Neural Network (ANN), there exists a sophisticated amalgamation of mathematical and statistical functions that assesses the data provided to the programme. DNN, short for Deep Neural Network, is composed of numerous hidden layers. The output of one layer is used as the input for the next layer, similar to how a Deep Belief Network (DBN) operates. The intricate architecture of the DNN allows it to exhibit a resilient and accurate performance when addressing classification problems [22], at the cost of demanding significant processing effort and power. Maria and her colleagues (Maria et al.) utilised a Convolutional Neural Network (CNN) technique to develop their predictive model. The authors utilised the PID dataset to train and validate the model. Nevertheless, it has been noted that the existing medical datasets available online for machine learning and deep learning applications lack an adequate number of samples, which is essential for achieving a higher level of accuracy in deep learning models. Therefore, the authors augmented the sample size by employing

Variational Autoencoder (VAE) and expanded the number of characteristics by utilising Stacked Autoencoder (SAE). To address missing or anomalous data in the dataset, they employ the strategy of substituting the mean value of the respective feature. In this study, the researchers have applied the Multilayer Perceptron (MLP) algorithm, along with the aforementioned data augmentation techniques. The CNN model, when combined with the SAE data augmentation method, outperformed all other state-of-the-art models, achieving an accuracy of 92.31%. As stated in the machine learning section, researchers consistently emphasise that the quantity of training data available in the PID dataset is a crucial factor in determining the success of models. Although the dataset is available, the author proposes that using different data imputation algorithms can solve this problem when it is not practical to increase the dataset size in real-world scenarios. Deep learning models can achieve good performance when correctly calibrated, even when dealing with significantly bigger datasets that require less data pre-processing. The study conducted by Sandhiya et al. [51] employed two feature selection techniques: Conditional Random Field and Linear Correlation Coefficient based Feature Selection (CRF-LCFS). The selected dataset was acquired from the UCI repository. Prior to feature selection, their proposed CNN model attained an accuracy of 82.5%. However, after conducting feature selection, the accuracy was enhanced to 84%. P.Prabhu et al. [43] developed the model using DBN as a foundation. Their proposed model achieved a Recall score of 1, Precision score of 0.6791, and F1 score of 0.808, outperforming previous machine learning-based predictive models. The study conducted by Safial I. A. et al. [5] compared the accuracies of their DNN-based prediction model with the five-fold k and ten-fold cross-validation approaches used in the PID dataset. Upon concluding their investigation, they determined that the five-fold cross-validation method outperformed the ten-fold cross-validation method. The DNN-based model attained a precision of 98.04% using the five-fold cross validation technique and 97.27% using the ten-fold cross-validation technique. Nadesh et al. [40] employed a Deep Neural Network (DNN) methodology in their research. The feature selection process involved the utilisation of an algorithm known as Feature Importance (FI), which is also referred to as Extremely Randomised Trees. This algorithm chooses four features from the database depending on the scores obtained in the FI entropy calculation. In analysing their DNN-based diabetes prediction model, they have also employed the 10-fold cross-validation method. The DNN-based model is trained using different train/test splits, including 60/40, 70/30, and 80/20. Increasing the size of the training part leads to improved accuracy in the DNN model. Specifically, when using an 80/20 train/test split, the model achieves an accuracy of 98.16%. When using the 10-fold cross-validation approach to this model, it obtained the lowest accuracy of 96.10% compared to the accuracy of 96.77% achieved by the 60/40 split method. Rakshit et al. [44] employed a Two-Class Neural Network model to differentiate between persons with and without diabetes using the dataset. Upon reviewing multiple research studies on primary immunodeficiency diseases (PIDD), the authors utilised all the features documented in the dataset, with the exception of Triceps skinfold thickness. They subsequently modified the scaling of each feature to optimise the model's learning process. 80% of the data was allocated for training the model, while the remaining 20% was reserved for testing the model. The Two-Class Neural Network-based diabetes predicting model attained an accuracy of 83.3%. In their study, Nesreen et al. successfully applied the model in the Just Neural Network (JNN) environment, as documented in reference [15]. The JNN-based prediction model has the capability to calculate the relative significance of each feature derived from the dataset, allowing researchers to conduct analytical tasks with greater precision. The calculated outcome is subsequently utilised to assist in the process of selecting features in order to enhance the final result. The predictive accuracy of this model for diabetes is 87.3%, with 76% of the data used for training and 24% for testing. In addition to traditional invasive datasets like the PID dataset, Vidhya et al. [62] conducted research on diabetic complication models utilising non-invasive datasets, employing DBN, SVM, and ANN. This study aimed to identify the most influential risk factors for diabetic complications in individuals with diabetes using the aforementioned categorization algorithms. I obtained a dataset from the UCI diabetes repository, which included of 13 features that were captured based on the life behaviours of the samples. The dataset was pre-processed using a Restricted Boltzmann Machine (RBM). The results are as follows: The DBN model achieved an accuracy of 81.19% in the training step and 80.99% in the validation stage. The SVM model achieved an accuracy of 72.72% in the training stage and 62.81% in the testing stage. The ANN model achieved an accuracy of 76.52% in the training stage and 57.61% in the testing stage. When choosing the features, the authors utilised unsupervised learning and pre-training of Restricted Boltzmann Machine prior to extracting the required features. Given the demonstrated feasibility and high accuracy of classifying diabetes using non-invasive datasets, it is imperative that this research topic is not overlooked and be further explored. In another study conducted by Ryu et al. [48], they developed a diabetes prediction model utilising a deep neural network (DNN) technique. The study primarily concentrated on the screening of samples including undiagnosed cases of type 2 diabetes, utilising the NHANES dataset. It is noteworthy that they also prioritise non-invasive aspects that do not necessitate lab tests or blood samples. The suggested DNN-based model achieved an AUC of 80.11. The works conducted by Vidhya [62] and Ryu [48] et al. have demonstrated that non-invasive datasets can be effectively utilised in a deep learning strategy for the purpose of solving the issue of diabetes categorization. However, in order to attain optimal outcomes, the presence of a significantly large dataset is crucial. Deep learning algorithms have demonstrated their efficacy, resilience, and accuracy in image processing, making them well-suited for detecting diabetes retinopathy. This condition, induced by diabetes, leads to damage in the retina of the human eye. Lam et al. [32] and Arcadu [4] utilised convolutional neural network (CNN) techniques for diagnosing diabetic retinopathy. On the other hand, Gadekallu et al. [16] employed a deep neural network (DNN) methodology, together with the PCA-Firefly Feature Selection algorithm. All three studies have demonstrated the feasibility of employing deep learning methods for diagnosing diabetic retinopathy. The model developed by Gadekallu et al. achieved the greatest accuracy of 97% [16]. Within the set of numbers [9, 29, 38, 42, 45, 47, 64, 71], in addition to utilising machine learning methods, they have also employed deep learning models for the purpose of identifying diabetes. Subsequently, the researchers examined which strategy can provide superior performance for this particular function. Comprehensive data regarding these research findings are shown in Table 2. In summary, the Deep Learning-based models outperform the Machine Learning-based Diabetes Detection Models, achieving an average accuracy of 86.7% in this diabetes classification function. According to Figure 15

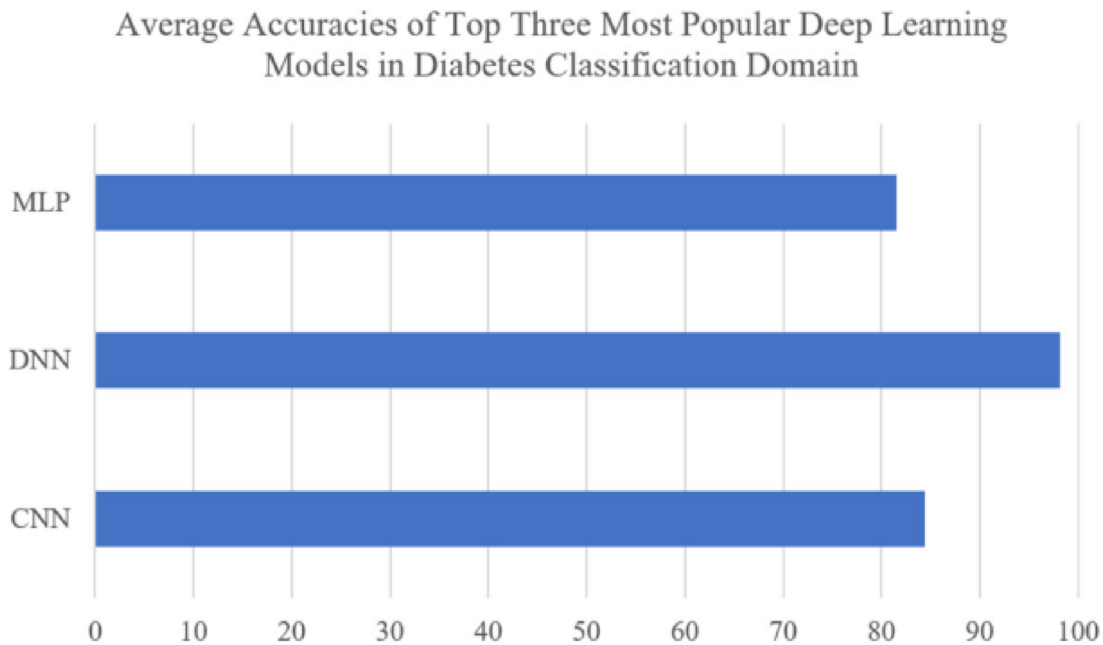
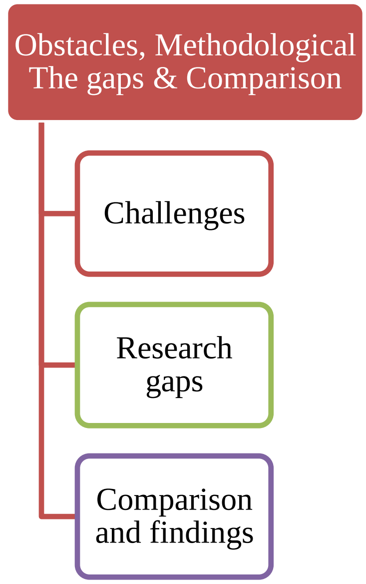


Fig. 15 Performance of three popular DL-based models

The table above displays the performances of three widely used deep learning models: CNN, DNN, and MLP. Overall, their respective accuracies were 84.37%, 98.1%, and 81.49%.

6 Obstacles, Methodological The gaps & Comparison

Despite recent advancements in research, the classification of diabetes using machine learning or deep learning methods still poses some unresolved challenges. In addition, this extensive literature study reveals additional potentials and opportunities for improvement that can be further explored. The following points will be addressed in this part of the article.



7. Conclusions and future endeavours

Ultimately, the creation of a data-centric diabetes detection model is essential due to the increasing prevalence of diabetes worldwide. The reliabilities of diabetes detection models that utilise non-lab tests and non-invasive measurements should be further examined to explore the potential for reducing medical costs and labour associated with diabetes detection and treatment. To accomplish this, a dataset of higher quality is necessary, meaning it should have more recorded characteristics and samples, with no missing or aberrant values. After conducting an extensive analysis of over 50 machine learning and deep learning models, it is determined that each type of algorithm (machine learning and deep learning) offers distinct advantages in certain domains.

Research has demonstrated that incorporating feature selection into machine learning models for diabetes detection is advantageous. Therefore, it is crucial to apply feature selection algorithms to the dataset and subsequently conduct a cross-test using a dataset that has not undergone feature selection. This will help determine if feature selection has any detrimental effects on the tested models. Moreover, while most deep learning algorithms already include built-in feature extraction and selection functions, research suggests that conducting a pre-feature selection process can help analyse their influence on classification. However, this aspect is rarely explored in existing studies. Furthermore, it is imperative for researchers to thoroughly examine the most economically efficient characteristics when generating datasets for this objective. This is considered a key concern in tackling the aforementioned cost issue. The most crucial factors to consider when developing a data-driven solution for diabetes classification are the cost, ethical considerations, and medical analysis.

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) 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

7. 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

8. 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

9. 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

10. 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

11. 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

12. 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

13. 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
14. 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
15. 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 202:10596
16. 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
17. Ghosh S, Dasgupta A, Swetapadma A (2019) A study on support vector machine based linear and nonlinear pattern classification. In: 201
18. 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
19. Intelligent machine learning approach for effective recognition of diabetes in e-healthcare using clinical data. Sensors 20(9):2649
20. 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
21. 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
22. 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
23. 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
24. 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
25. 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

26. 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

27. 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

28. 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

29. Lam C, Yi D, Guo M, Lindsey T (2018) Automated detection of diabetic retinopathy using deep learning. AMIA Summits Transl. Sci. Proc. 2018:147

30. 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

31. 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.

32. 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

33. 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

34. Mathis D, Vence L, Benoist C (2001) ß-cell death during progression to diabetes. Nature 414(6865):792– 798

35. Mercaldo F, Nardone V, Santone A (2017) Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. Procedia Comput. Sci. 112:2519–2528

36. 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

37. 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

38. Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Front Neurorobot 7:21

39. Naz H, Ahuja S (2020) Deep learning approach for diabetes prediction using pima indian dataset. J Diabet Metabol Disord 19(1):391–403
40. 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
41. 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
42. 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
43. 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
44. 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
45. 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
46. 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 DiabetesMellitus. 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
47. 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
48. 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
49. 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

50. Schober P, Boer C, Schwarte LA (2018) Correlation coefficients: appropriate use and interpretation. Anesth Analg 126(5):1763–1768
51. Shrestha A, Mahmood A (2019) Review of deep learning algorithms and architectures. IEEE Access 7:53040–53065
52. 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
53. 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
54. 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
55. Swapna G, Vinayakumar R, Soman K (2018) Diabetes detection using deep learning algorithms. ICT Express 4(4):243–246
56. Tao K-M, Sokha S, Yuan H-B (2019) Sphygmomanometer for invasive blood pressure monitoring in a medical mission. Anesthesiology 130(2):312–312
57. 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
58. Tharwat A, Gaber T, Ibrahim A, Hassanien AE (2017) Linear discriminant analysis: A detailed tutorial. AI Commun 30(2):169–190
59. 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
60. 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
61. 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
62. Yang L, Shami A (2020) On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing 415:295–316
63. 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

64. 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

65. 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

66. 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

67. 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