Journal of Pioneering Medical Sciences
Received December 19, 2023 Accept February 05, 2024 Publish April 29, 2024.
Volume 13, Issue 2, Pages 60-73

DOI https://doi.org/10.61091/jpms202413209

Deep Learning Models for Type 2 Diabetes Detection in Saudi Arabia

Noha Alsulami[1,*], Miada Almasre[1], Shahenda Sarhan[2] and Wafaa Alsaggaf[1]
[1]IT. King Abdulaziz University, Jeddah, Saudi Arabia.
[2]CS. Mansoura University, Mansoura, Egypt.

Corresponding author: Noha Alsulami (e-mail: nmuttlaqalsulami@stu.kau.edu.sa).

Abstract One of the predominant health issues affecting Saudi Arabia and leading to many complications is Type 2 diabetes (T2D). Early detection and significant preventative measures lead to curbing and controlling the health issue. There are fewer datasets in the literature for the detection of T2D in the Saudi population. Past studies using Saudi data have favored machine learning algorithms to classify T2D. Although the application of this data in machine learning is evident, no studies exist in the literature that compare this data, especially those related to deep learning algorithms. This study's objective is to use specific Saudi data to develop multiple deep learning models that could be used to detect T2D. The research uses a Deep Neural Network (DNN), an Autoencoder (AE), and a Convolutional Neural Network (CNN) to create predictive models and compare their performance with a traditional machine learning classifier used on the same dataset that outperformed other machine learning algorithms such as a Decision Forest (DF). Various metrics were used to evaluate the effectiveness of the models, such as accuracy, precision, recall, F1 score and area under the ROC curve (AUC) where the ROC acts as a receiver operating characteristic curve. There are two cases in this paper: (i) uses all features of the dataset and (ii) uses six of the ten features, such as DF. In case (i), the results were shown that AE outperformed other models with the highest accuracy for imbalanced and balanced data 81.12% and 79.16%, respectively. The results for case (ii) showed that AE scored the highest 81.01% accuracy with imbalanced data compared to DF and DF achieved the highest accuracy of 82.1% with balanced data. As a result, both cases explored in this study revealed that AE has a constant superior performance if imbalanced data is used. In contrast, DF demonstrated the highest accuracy when a balanced dataset was used with a feature set reduction. They help to identify the undiagnosed T2D, and they are essential for professionals in Saudi Arabia in the health sector to promote health connections, identify risks and contain or improve their diabetes management.

Key Words Diabetes, Type 2 diabetes, Deep learning, Detection model, Deep neural network, Autoencoder, Convolutional neural network.

1. Introduction

Insulin resistance and deficiency are the main causes of type 2 diabetes (T2D), which raises blood sugar levels [1]. The illness has spread around the world in recent decades and impacted a large number of people [2]. According to estimates from the World Health Organization (WHO), 3 million Saudis have prediabetes, or the risk of developing diabetes. According to WHO data, 7 million Saudis suffer from diabetes, which has a serious impact on their lives [3]. Furthermore, with an estimated 24% of the adult population suffering from T2D, Saudi Arabia is rated seventh among nations with a high prevalence of T2D [4], [5]. Factors including urbanization, obesity rates rising quickly in the area, and sedentary lifestyles are to blame for the large number of patients [5]. Early detection and the right kind of intervention can help manage T2D. This will postpone or avoid the disease's consequences, like retinopathy, heart problems, and renal failure. Early identification hence improves treatment outcomes and lessens the load on healthcare systems [6]. Healthcare problems including diagnosis and disease identification have improved thanks to the inclusiveness of deep learning and machine learning approaches [7]. Furthermore, the methods' computation has affected the early detection of T2D; for this reason, analyzing the medical records is necessary. However, a number of studies and papers have been published on the superiority of deep learning and machine learning models over antiquated statistical techniques in indicating the risk of diabetes [8].

Motivation. Because early detection of type 2 diabetes aids in the management and prevention of issues like nephropathy, retinopathy, and cardiovascular disease, it is crucial. Certain conventional screening techniques, such blood glucose testing and oral glucose tolerance tests, take a lot of time and don't always yield reliable results [9]. Thus, there has been interest in creating a disease detection technique that is accurate, dependable, and economical. Due to their ability to analyse massive datasets, machine learning and deep learning techniques have been shown to be effective [7]. Because the approaches analyse complex data automatically rather than by hand, they are more efficient than standard statistical methods and have been applied in a variety of healthcare applications [10]. Numerous studies have looked into the use of deep learning and machine learning methods for T2D prediction in recent years. Deep learning algorithms are being used to diagnose type 2 diabetes in Saudi population, although there is a dearth of research on this topic. There aren't many datasets available in related literature for Saudi population T2D detection. One of these datasets was gathered in [3], and the dataset sub-section of this work will detail it. Despite this, the authors in [3] obtained the greatest accuracy of 78.9% using a decision forest classifier for unbalanced data and 82.1% with balanced data using the Synthetic Minority Oversampling Technique (SMOTE). They did, however, only employ a few machine learning algorithms. Furthermore, comparable research regarding the application of deep learning models to T2D datasets is currently lacking.

Deep learning is one of the most widely used techniques in literature for analyzing large-scale data. In most of the studies in which it has been applied, deep learning has demonstrated excellent performance, which is why it is a widespread technique. Below are several instances of applications where deep learning has made a significant impact.

1) **Image Recognition:** Deep learning models, particularly Convolutional Neural Networks (CNNs), have excelled in image classification and object detection tasks [11].
2) **Natural Language Processing (NLP):** Models like BERT have set new benchmarks in a variety of NLP tasks, including question answering and sentiment analysis [12].
3) **Autonomous Vehicles:** Deep learning algorithms are vital for interpreting sensory input and making driving decisions in self-driving cars [13].
4) **Game Playing:** Deep reinforcement learning models, such as AlphaGo, have defeated world champions in complex games like Go [14].
5) **Healthcare and Medicine:** Deep learning is utilized in predictive analytics, disease identification, and drug discovery [15].
6) **Speech Recognition:** Deep learning models are extensively used in developing accurate speech recognition systems [16].
7) **Fraud Detection:** Deep learning assists in identifying fraudulent activities through pattern recognition in transaction data [17].
8) **Customer Service:** Chatbots and virtual assistants employing deep learning offer enhanced customer service by understanding and responding to user queries effectively [18].
9) **Facial Recognition:** Deep learning is used in facial recognition systems for security and authentication purposes [19].

Contributions. There are two gaps related to Type 2 Diabetes (T2D) detection in Saudi Arabia that this research aims to fill to contribute to the field. Some of the ways that this study will contribute to the field include:

1) **The development of deep learning models:** In this study, various deep learning models are tailored for T2D detection in the Saudi population, which are developed and implemented. These deep learning models are Deep Neural Network (DNN), Convolutional Neural Network (CNN), and Autoencoder (AE). All these three models are advanced computational techniques leveraged in the study to determine how effective they are in identifying complex medical patterns among T2D patients. It is important to note that previous Saudi Arabian studies in literature have failed to explore this area thoroughly.
2) **Comparative performance analysis:** This paper further contributes to the existing literature by comparing the deep learning models that we developed with Decision Forest, a traditional machine learning classifier. The DNN, AE, and CNN comparisons with Decision Forest were performed using a Saudi-specific dataset, which has been referenced [3]. These comparisons were done with and without the Synthetic Minority Oversampling Technique (SMOTE) application, which is essential for addressing class imbalance. The comparisons of this study were made up of two scenarios, which used available features and reduced six out of ten features, which is an approach similar to the one taken in the Decision Forest [3].

The datasets used in this work were both balanced and imbalanced data. These datasets are essential because they provide important information regarding data preprocessing techniques such as SMOTE, which influence the performance of deep learning and traditional models. This research examined the performances of both a complete feature set and a reduced one, which enabled us to evaluate their capabilities and techniques that can be adopted to optimize them for detecting T2D in Saudi Arabia. As a result, this study contributed innovatively to how the current gaps in literature can be bridged by introducing deep learning methods to the Saudi T2D detection landscape and proposing a rigorous comparative framework, which can be used by future studies and in clinical decision-making processes. The results of this study can be used to enhance both T2D screen and diagnosis effectiveness and precision in Saudi Arabia. This paper is structured into distinct sections: Section 2 delineates the models employed for diabetes detection. Section 3 presents materials and methods used. Section 4 exhibits the experimental results, and Section 5 provides a conclusive summary.

2. Related work

The methods used to detect diabetes, particularly type 2 diabetes, are displayed in this section. With the aid of machine learning techniques, Farooq Ahmad et al. [20] elaborates on the factors that predict T2D. Three thousand patients' worth of records from several Saudi hospitals were used by the researchers. Afterwards, they used preprocessing methods and discussed their importance, resulting in a 162 case decrease. Modelling using the techniques of ensemble majority voting (EMV), random forest (RF), logistic regression (LR), decision tree (DT), and support vector machine (SVM). SVM obtained 82.10% in the first dataset, RF with nine features achieved 88.27% accuracy, and RF with eight features obtained 87.65% accuracy in the second dataset. Additionally, Syed and Khan [3] developed an application that might be utilized to identify T2D in Saudi Arabia. Data for their study were obtained from King Abdulaziz University and included 3906 non-diabetic subjects and 990 diabetic cases. To identify important features, binary LR and the Pearson chi-squared test were employed. The dataset underwent pre-processing using an 80:20 ratio, dividing it into training and testing sets. The Synthetic Minority Oversampling Technique (SMOTE) was used to attain equilibrium. After utilising nine different binary classification methods, it was discovered that the Decision Forest (DF) approach outperformed other models.

Gollapalli et al. [4] created an efficient model for diabetes that was able to predict and detect three types of diabetes: Type 1 Diabetes, T2D, and prediabetic. They did this by applying a variety of machine learning classifiers on a dataset that was obtained from King Fahad University Hospital (KFUH), a hospital in Saudi Arabia. There were 897 instances and 10 different attributes in the sample. The writers or researchers used stacking techniques, SVM, Bagging, DT, K-nearest neighbour (KNN), and RF as their main classifiers. To maximize the outcomes, four trials were carried out, with experiments 2, 3, and 4 employing SMOTE to balance the dataset. Their novel stacking model combined KNN with a KNN meta-classifier, Bagging DT, and Bagging KNN to obtain an accuracy of 94.48%. Sex, education, antiDiab, and nutrition were the five factors that were found to have a significant impact on the accuracy of the model through a critical study of feature importance. Alassaf et al. [21] presented a method intended to proactively diagnose diabetes in an uncharted area. They received data from Khobar's KFUH, which was the first time the information was used to support a diagnosis. The authors performed pre-processing and identified key traits before classifying the data. In addition, they employed recursive feature elimination and the correlation coefficient for feature selection. Following that, four categories of algorithms Naïve Bayes (NB), Artificial Neural Networks (ANN), Support Vector Machines (SVM), and KNN—were evaluated with an emphasis on classification accuracy, F1-measure, precision, and recall. ANN fared better than the other models, with 77.5% accuracy.

Alanazi and Mezher proposed a model combining RF and SVM classifiers to predict diabetes. They obtained a real dataset from the primary health care unit of the security forces in Tabuk, Saudi Arabia. Their model employed RF showed a remarkable performance with an accuracy of 98% and 99 % receiver operating characteristic curve (ROC) boast. It signifies that the RF method was much better than SVM in accuracy [22]. In [23], the research project leveraged real healthcare datasets comprising 18 attributes, sourced from the Ministry of National Guard Health Affairs (MNGHA) database. The primary objective was to construct a predictive model for identifying diabetic patients within the adult population of Saudi Arabia. Three distinct algorithms, namely the Self-Organizing Map (SOM), C4.5, and RF, were applied for this purpose. Comparative analysis against various classifiers revealed that RF consistently delivered superior performance.

In the investigation of T2D, Jaber and James in [24] employed diverse classifiers, including the NB Algorithm, the LR Algorithm, and the RF Algorithm. The Pima Indian Diabetes Dataset (PIDD) served as the foundational dataset. The findings underscored the supremacy of the RF Algorithm when compared to alternative approaches. Various machine learning algorithms including linear discriminate (LD), linear SVM, quadratic SVM, cubic SVM, Gaussian SVM, fine KNN, weighted KNN and neural pattern recognition (NPR) were harnessed to construct classification models for diabetes detection in [25]. The result of a rigorous performance analysis showed that the weighted KNN showed commendable predictive accuracy in estimating the prevalence of diabetes in both male and female datasets, with an impressive average accuracy of 94.5% and shorter training time compared to other classification methods.

In [26], the study systematically employed pertinent features to forecast diabetes and elucidate the intricate interplay among these variables. Essential tasks such as attribute selection, grouping, prediction, and association rule mining for diabetes were executed employing diverse analytical tools. Principal component analysis facilitated the identification of salient attributes. The findings underscored a significant correlation between body mass index and glucose levels, elucidated through the Apriori approach. Diabetes prediction was executed through the deployment of ANN, RF, and K-means clustering approaches, with the ANN approach achieving the highest accuracy at 75.7%, thereby potentially assisting medical practitioners in refining treatment decisions. The overarching aim of the study in [27] was to harness pertinent features, develop predictive algorithms using machine learning techniques, and ascertain the optimal classifier to generate results that closely align with clinical outcomes. The proposed approach was centered on the identification of pivotal attributes crucial for early diabetes detection. Various machine learning algorithms were evaluated, including SVM, RF, NB, DT, and KNN. Notably, the examination of diabetic data revealed that the DT and RF yielded maximum specificity rates of 98.20% and 98.00%, respectively. Moreover, the NB approach achieved the highest accuracy rate of 82.30%. To further enhance classification accuracy, this research also incorporated feature selection techniques to identify the most influential variables within the dataset.

The identified areas for improvement identified in the related work with this study that could lead to improvements included:

- **Limited Deep Learning Exploration:** Most of the studies that have been conducted in the literature are those that provide insights regarding machine learning and T2D. The literature analysis indicates a massive gap regarding the impact of deep learning methods in Saudi Arabia. Common deep learning models overlooked in studies conducted in Saudi Arabia include DNN, AE, and CNN, which tend to have significant success in domains such as image recognition, natural language processing, and bioinformatics. DNN, AE, and CNN are deep learning methods that can assist in exposing intricate patterns in complex T2D detection medical data. However, the sad reality is that these deep learning methods in Saudi Arabia are untapped areas.

- **Absence of Comparative Research:** Currently, no comparative studies have been conducted in Saudi Arabia using the same dataset used to compare deep learning methods and the traditional machine learning models for T2D detection. This is an issue since comparative studies are vital for a better understanding of the relative advantages and limitations of various analytical approaches. The lack of such studies makes it extremely difficult to conclude whether deep learning methods can help address machine learning models' accuracy, speed, and diagnosis weaknesses in detecting T2D.

- **Unavailability of Public and Free Datasets for T2D Detection in Saudi Arabia:** The lack of available and free dataset within the Saudi population, which is accessible to researchers, cause a substantial hiatus in the present research landscape for T2D. The ability of the researchers to develop models and carry out thorough studies that can be generalized for the entire population is limited, because the datasets available are usually kept within the hospitals and not generally released to the public. This in turn obstructs progress by restricting the extent of possible research and impedes the development of strong diagnostic tools that can benefit the larger community. This problem in Saudi Arabia could be bridged by establishing an open-access, anonymized dataset, potentially, catalyzing advancements in detecting and managing T2D.

3. Materials & Methods

This study made use of deep learning approaches to compare T2D classifications. The two cases that this study considered are those that use all features and those that used 6 out of 10 features as used with the Decision Forest in [3]. Both cases use balanced data (with SMOTE) and imbalanced data (without SMOTE). The whole system for detecting the T2D disease is depicted in Figure 1. The dataset is first pre-processed. The pre-processed dataset was then separated into two pieces: (1) training set and (2) testing set. The system uses different deep learning models such as DNN, AE and CNN. Finally, a set of performance measurement metrics are used to assess each model's effectiveness such as accuracy, precision, recall, F1-score and area under the ROC curve (AUC), in which ROC acts as the receiver operating characteristic curve.

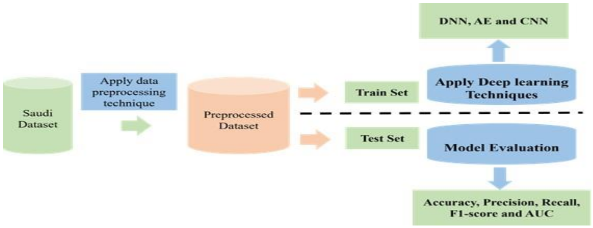

DNN, AE and CNN
Apply data preprocessing technique
Saudi Dataset
Preprocessed Dataset
Train Set
Apply Deep learning Techniques
Test Set
Model Evaluation
Accuracy, Precision, Recall, F1-score and AUC

Figure 1: The whole system for detecting the T2D disease.

A. *Dataset*

The dataset used was collected by Syed and Khan in [3]. This dataset is named Saudi Dataset (SD) to be easier to follow. SD is collected by filling out a cross-sectional survey. Participants from the King Abdulaziz University (KAU) were issued with a fill-in form to complete. The survey utilized close ended questions to gather data regarding diabetes risk factors for the participants from KAU to predict the occurrence of diabetes in Saudi Arabia's western part. The researchers extracted the most common attributes of diabetes prediction from models of diabetes papers that had recently been published when preparing the survey. The researchers used non-invasive tests and direct observation techniques to address attributes in the survey. However, the researchers first obtained the necessary permissions from the KAU Deanship of Graduate Studies before conducting the study at the university. The study participants from KAU were students, staff, and faculty members. The survey contains eleven questions as follows [3]:

1) Region.
2) Age group.
3) Sex.
4) Body Mass Index (BMI).

5) Waist size (navel level measurement).
6) Do you do a minimum of 30 minutes of physical activity every day?
7) Fruits and vegetables daily intake.
8) Are you currently undergoing hypertension treatment?
9) Does your family tree have a history of diabetes?
10) Do you smoke?
11) Has your blood glucose been high for any reasons (such as pregnancy, illness, etc.)?

The total number of subjects in this study was 4896, of which 990 were those at high risk, and the remaining 3906 were those at low risk of having diabetic complications. The diabetes related data attributes of the participants are shown in Table 1, where the researchers used the Label Encoding method as a pre-processing step to encode the attributes. Q1-Q10 represent either the explanatory variables or predictors of this study. They used the survey's last question as a categorical response variable retrieving data on the level of High Fasting Blood Glucose. It was agreed that those who answered "Yes" to the categorical response variable question, would be considered that they were at a high risk of developing diabetes. For those who answered "No" to the categorical response variable question, would be considered that they were at a low risk of developing diabetes [3].

B. Deep Learning Models

This part defines the used deep learning models including DNN, AE and CNN.

• **Deep Neural Networks:** Deep Neural Networks (DNN) are neural architectures where neurons are structured in a connected series across multiple layers. Neurons in each layer receive activations from the previous layer, creating a continuous series of interconnected neurons. These neurons collectively perform computations on input data, which include weighted summation followed by nonlinear activation functions. Consequently, DNN exhibits complex and nonlinear mappings from input to output. Through the backpropagation technique, DNN has the capacity to learn these intricate mappings from data by adjusting the weights of each neuron [28].

• **Autoencoder:** The autoencoder (AE), also referred to as an auto associator, belongs to the family of Artificial Neural Networks (ANN) and operates as an unsupervised learning algorithm. Its primary objective is to encode datasets, effectively reducing dimensions. Historically, auto-associators have been the focus of extensive research within the field of ANN. Recently, autoencoders have found significant application in the realm of learning generative data models. The typical workflow begins with inputting data, which is subsequently transformed into an abstract representation. Following this, the encoder mechanism comes into play, converting the abstract representation back to its original format. The encoder possesses the capacity to encode the input into a distinct representation, with the ultimate goal of ensuring seamless reconstruction of the input data from that representation. Throughout this process, the autoencoder diligently endeavors to establish an identity function. A noteworthy attribute of autoencoders lies in their ability, during the propagation phase, to systematically eliminate extraneous data features while retaining valuable information. This coding process frequently involves transforming the input vector into a lower-dimensional representation, thereby enhancing the overall efficiency of the learning process [29], [30].

• **Convolutional Neural Networks:** Commonly known as CNN, Convolutional Neural Networks are neural architectures organized hierarchically, featuring layers that alternate with subsampling layers. These neural networks draw inspiration from the simple and complex cells found in the human visual cortex. Collectively, these hierarchical neural networks comprise fully connected layers analogous to Multilayer Perceptrons (MLP). CNN operates in a manner akin to the human visual system, adept at identifying patterns and structures in visual data. Since their inception, CNN has gained substantial popularity and has become one of the prime choices for a wide range of deep learning tasks, including object recognition in large images [31]. In this paper, CNN with one dimension is used.

No.	Attributes	Type	Description	Labels
1	Region	Integer	Subject's Region	10=Yambu, 1=Abwa, 3=Khulays, 9=Thual, 2=Jeddah, 8=Sabar, 4=Medina, 7=Rabigh, 5=Masturah, and 6=Mecca.
2	Age	Integer	Subject's Age	0:Age<40 Years, 1:Age>=40 & Age<=49 Years, 2:Age>=50 & Age<=59 Years, and 3:Age>=60 Years
3	Gender	Integer	Subject's Gender	0=Female and 1=Male
4	BMI	Integer	Body Mass Index of Subject in (weight in kg/(height in m)^2)	0:BMI<25 Kg/m2, 1:BMI>=25 &BMI<=30 Kg/m2, 2:BMI>30 Kg/m2
5	Waist Size (WS)	Integer	Subject's Waist Size in cm for Male and Female	0_{male}:WS<94cm OR 0_{female}:WS <80cm, 1_{male}:WS>=94 & WS<=102cm OR 1_{female}:WS>=80 & WS<=88cm, 2_{male}:WS>102cm OR 2_{female}:WS>88cm.
6	Physical Activity	Integer	The Subject's Physical Activity is defined as 30 minutes of exercise daily.	0:Yes and 1:No

7	Diet	Integer	The Subject's Healthy Diet is Defined as regular consumption of fruits and vegetables.	0:Everyday and 1:Not Everyday
8	BP	Integer	Does the Subject Take Blood Pressure Medicine or No?	0:No and 1:Yes
9	Family History	Integer	The subject's Family History is Defined as having any member diagnosed with diabetes.	0:Family has no history with Diabetes, 1:Grandparents have Diabetes, and 2:Parents have Diabetes
10	Smoking	Integer	Subject's Smoking Habit.	0:Non-Smoker and 1:Smoker
11	Class	Boolean	This is the subject's response variable based on fasting plasma glucose exposure=5.6mmo/L in the examination of health or during a health examination or expectant.	0:Low Risk and 1:High Risk

Table 1: T2D diagnosis features of SD [3]

C. Preparing the data

The Tensorflow library was installed to ensure that the proposed model could be used in the Keras library. Keras was most preferred because it makes it simple to create neural networks capable of operating on Central Processing Units (CPUs) and Graphics Processing Units (GPUs) simultaneously. Such a system allows for faster computations and seamless parallel processing. The key component of deep learning network construction is the organization of the simple layers that require fewer steps for using Keras to construct complex networks [32]. The first process of the proposed solution involves importing the necessary libraries for handling, modelling, and processing, such as numpy, pandas, TensorFlow, and sklearn. After importation, it preprocesses input features by normalizing input features. This process is important as it ensures that all the input features are on the same scale, improving the model's training. The technique that was used as an oversampling technique is the Synthetic Minority Oversampling Technique (SMOTE), which was used to balance the dataset. The next step is splitting the dataset into 70% and 30% for training and testing, respectively as shown in Table 2.

Dataset	Subjects number	Train	Test
SD	4896	3427	1469

Table 2: Data Processing (Separation)

D. Building and Training Deep Learning Model

The deep learning model was constructed using three different types of layers as follow:

- The input layer is the layer to which the dataset features would be transferred. There are no computations that take place in the input layer. This layer only allows features of the datasets to be transmitted through it to hidden levels.
- The hidden layers are the layers found between the input and output layers. The hidden layers are used for computations and transferring the data results to the output layer.
- The output layer is the neural network layer. The results of the dataset are usually displayed after training the newly generated model. The output layer is the layer that is responsible for generating output variables [33].

(i) Case: Using all Features of Dataset

- **DNN:** In DNN, the first input layer is usually size 10, containing the first hidden layer comprising 128 units with ReLU. Then it is followed by the batch normalization layer, whose job is to improve the neural network's speed, performance, and stability. After that, it is followed by the dropout layer, rated 0.5, whose job is to prevent overfitting. The dropout layer achieves this during training, setting half of the input units to zero. The other layers of the model are the second and third hidden layers, each having 256 units and ReLU activation, batch normalization, and dropout with a rating of 0.5. In the end, the model adds an output layer containing one unit and an activation function known as a sigmoid necessary for binary classification. To avoid binary classification problems, the classifier has a loss function developed by the RMSprop optimizer and binary cross-entropy.
- **AE:** AE is made up of the training of an autoencoder and the training of a classifier. The first step usually involves defining and training an autoencoder. An autoencoder is a neural network that learns how input data is reconstructed. The input layer of an autoencoder is 10, comprising of a 64-size encoding layer containing ReLU and activation normalization, and batch normalization. The autoencoder also holds a size ten decoding layer containing a linear activation. The next step is leveraging the Adam optimizer for 200 epochs to train the autoencoder using mean squared error loss. Then, the encoder part can be extracted through the training, allowing input mapping to the encoded representation's lower dimensional. Then, the encoded data is used to

define and train a classifier. The classifier comprises a size 64 input layer that matches the encoded data size. Other classifier layers are a size 32 hidden layer with ReLU activation and batch normalization and a size one output layer with sigmoid classification for binary classification. An RMSprop optimizer has a 0.001 learning rate, and the binary cross-entropy loss are used to comply with the classifier. The encoded representation of the training data is used to train the classifier for 200 epochs.

- **CNN:** CNN is constructed and trained as one-dimension Convolutional Neural Network (ID CNN). First, the input data is reshaped to meet the ID CNN's requirements. Some dimensions in the reshaped data include several samples and features. The building process of the Sequential model classifier usually starts with a ReLU activation function, a kernel size of 3, and a 1D convolutional layer with 64 filters. The ID convolutional layer is then followed by a size two max pooling layer and a 0.5-rated dropout layer. The 0.5-rated dropout layer's job is to prevent overfitting and achieves this by setting half of the input units to 0 randomly during training. The convolutional layer's output is then flattened to a single dimension, enabling it to serve as input to the dense layers that follow it. The convolutional layer is followed by two dense layers having 128 units and 64 units, respectively, with each one of them followed by a 0.5-rated dropout layer. The model has an output layer containing a single unit with an activation function known as a sigmoid for binary classification purposes. The RMSprop optimizer and the binary cross-entropy are used to compile the model acting as the loss function. A size 128 batch trains the model on reshaped training data for 200 epochs. The performance evaluation on the reshaped test data is done after training the model.

(ii) Case: Using 6 features out 10

In this case, Region, Gender, BMI, Diet, BP and Smoking are used as defined in Table 1, resembles the technique used by Decision Forest as depicted in [3]. For all deep learning models, the experiment of using all features case was repeated in this case with changing the size of input layer from 10 to 6.

E. Tuning the Algorithms

Deep learning models differ from machine learning models in that deep learning models have been practically staffed with hyperparameters. These hyperparameters control the number of hidden units, known as the network's structure [33]. In this study, we went further to perform a hyperparameter search using the grid search method (GridSearchCV) discussed in the sklearn module. However, this study made use of a large number of parameters, and we decided to tune DNN, AE, and CNN by selecting those parameter values that had the best accuracy. Eight processes were used in determining the best parameter values for this study, which are [34]:

- **Optimization algorithm training:** This process involved using tools to update model parameters and reduce loss function value, which had been identified during the training set evaluation.
- **Epochs:** This involved determining how often the learning algorithm could work during the training session.
- **Activation function:** This involved determining the activation output of the neuron based on the inputs fed to the models.
- **Learning rate:** This involved leveraging the hyperparameter used to control weights being adjusted with respect to the loss of gradient.
- **Batch size:** This refers to the training examples used in one iteration.
- **Hidden layers number:** It refers to the number of hidden layers contained in each model.
- **Neurons number found in hidden layers:** It refers to the neurons contained in each of the network's hidden layers used in this study.
- **Dropout regularization:** This refers to the process in the networks that results in the random dropping out of nodes during training.

The DNN's fine-tuning parameters, the AE's fine-tuning parameters and the CNN's fine-tuning parameters are shown below in Table 3, 4 and 5, respectively.

Parameter	Value
Training optimization algorithm	RMSprop
Epochs	100
Activation function	Sigmoid for Output layer ReLU for hidden layers
Learning rate	0.001
Batch size	64
Hidden layers number	4
Number of neurons in the hidden layers	H1=128, H2=256, H3=256, H4=128
Dropout regularization	0.5

Table 3: Fine-Tuning parameters of DNN

Parameter	Value
Training optimization algorithm	RMSprop and Adam
Epochs	200
Activation function	Sigmoid for Output layer ReLU for hidden layers
Learning rate	0.001
Batch size	32
Hidden layers number	1
Number of neurons in the hidden layers	32

Table 4: Fine-Tuning parameters of AE

Parameter	Value
Training optimization algorithm	RMSprop
Epochs	200
Activation function	Sigmoid for Output layer ReLU for other layers
Learning rate	0.001
Batch size	128
Convolutional layers count	1
Filters of each convolutional layer count	64
The size of kernel in each convolutional layer	3
The size of pooling after each convolutional layer (if applicable)	2
Dense layers count	2
Neurons of each dense layer count	128 and 64
Dropout regularization	0.5

Table 5: Fine-Tuning parameters of CNN

F. Pseudocode

The pseudocode for the proposed solution that involves using SD to detect diabetes is as follows:

1. Importing of the SD and libraries used for data processing.
2. The input data is pre-processed by normalizing input features, using SMOTE and dataset division into training and test sets.
3. The necessary Python packages, including Tensorflow and Keras library, are imported and configured into the Python environment.
4. The DNN, AE and CNN are initialized.
5. The first hidden layer and input layer are added.
6. Additional layers are added, and the activation functions are used.
7. The activation function and output layer are added.
8. Compilation of deep learning models.
9. The deep learning models are fitted to the training set.
10. Prediction of the test results.
11. Evaluation of the deep learning models.
12. Deep learning models are tuned and improved.

4. Experimental Results

The algorithms were executed using Python programming language. They can be run on Google Colab, providing it with access to GPUs, which help to speed up the training process significantly. When the algorithm is run on Google Colab, it freely accesses the GPUs, developing and running the Python code directly without requesting additional configurations [35]. At the backend, TensorFlow is used to build and train the programs' networks written in Python 3.

A. Performance measures

Evaluation of the performance of the proposed models was done using various measurements such as Accuracy (Acc), precision, recall, F1-measure, and area under the ROC curve (AUC), in which ROC acts as the receiver operating characteristic curve. The model's accuracy is defined as the patient proportion that the models diagnosed properly. The accuracy of the model's formula is shown in (1) [36]:

$$Accuracy\ (Acc) = \frac{TN + TP}{TN + TP + FN + FP} \quad (1)$$

True positive (TP) is a term used to refer to the population of patients classified as positive and are truly positive. True negative (TN) represents the predicted patients that are negative and actually negative. False positive (FP) in the equation represents the patients classified as positive, while in the real sense are negative. False negative (FN) refers to the number of patients that are categorized as negative but are actually positive. To determine the proposed model's classification quality, the parameters of the access are regularly estimated [36]. Precision is the second performance evaluation metric that refers to the sum of true positive and true negative. The formula for calculating precision is presented in (2) [37]:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall refers to the attributes that are classified correctly and is also known as sensitivity or the True Positive Rate. Recall is expressed as the total number of positive predictions divided by the total number of class values. The formula for calculating recall is presented in (3) [37]:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

The F1-measure is also known as the F1-score, and it is a performance metric since it can provide data between the recall and precision. The formula used to calculate F1-measure is presented in (4) [37]:

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

The Area under the Curve (AUC) value can be used to measure the discriminative power of classification algorithms. AUC is used to assess models' performance with values ranging from 0 to 1. Values of 1 or near 1 mean the model is excellent at finding the balance between recall and precision. Such values indicate models capable of producing superior classification performance [3].

B. Results

This paper contains two cases, as it was indicated earlier as follows:

(i) Case 1: Using the entire dataset's features.

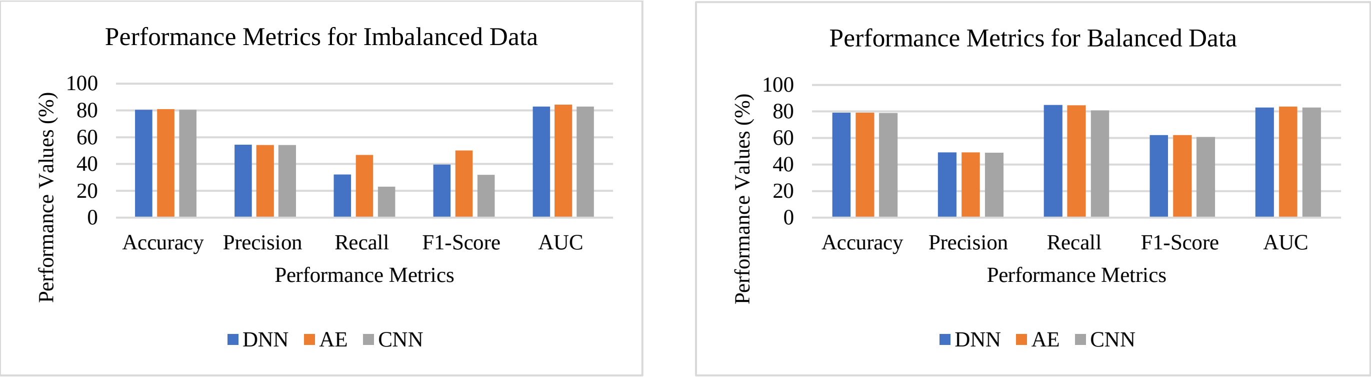
In the first case, we used the dataset's entire features as outlined in Table 1 above. This study went a step ahead to explore how three deep learning models compare with one another in detecting T2D in data that was imbalanced (does not use SMOTE) and balanced (makes use of SMOTE). SMOTE is currently one of the methods most used to address imbalanced classes. Thus, this study made use of SMOTE to obtain balanced data. As shown in Figure 2, the metrics that the researchers used to evaluate the developed deep learning methods are accuracy, F1 score, precision, AUC, and recall. Figure 2 (a) below shows the imbalanced data for this study:

- **DNN:** This is one of the models that demonstrated good accuracy and AUC of 80.61 % and 83.02 %, respectively. However, it was evident that DNN had a moderate precision of 54.44 % and low recall and F1-score of 32.25 % and 39.66 %, respectively. These results indicate that DNN is unreliable for identifying positive cases in imbalanced data, as some challenges exist.
- **AE:** This model demonstrated that it could deliver superior performances in recall and F1-scores of 46.88 % and 50.05 %, respectively. AE also recorded excellent performance in accuracy and AUC of 81.12 % and 84.3 %, respectively. The scores of this model indicate that it is reliable for identifying diabetic cases, especially in imbalanced datasets.
- **CNN:** The results of this model were slightly similar to DNN in specific parameters and significantly differed in others. For example, the accuracy score of this model and AUC were 80.41 % and 83.02 %, respectively. However, CNN had lower scores than DNN in precision, recall, and F1-score, with 54.14 %, 23.06 %, and 32.04 %, respectively. As a result, it can be concluded that CNN is a model that is not entirely reliable in detecting diabetes cases in patients.

Based on the scores of the three models developed in this study, it can be argued that AE is the best model for detecting cases of diabetics. This is because AE has the highest score among the three models in recall, F1 score, AUC, and accuracy. Although DNN outperformed the other models, especially in precision, AE had the highest scores in all the other metrics. Figure 2 (b) shows the results of the three models for balanced data. The summary of the scores is as follows:

- **DNN:** This model recorded strong performances in recall of 85.03% but moderate accuracy, precision, and F1-scores of 79.1%, 49.12%, and 62.27%. Nevertheless, the excellent recall score reveals that DNN is an effective model for identifying positive cases in balanced data.
- **AE:** Again, AE had some of the best performances when compared to the other two models for detecting diabetic cases in balanced data. For example, the scores of this model were: accuracy (79.16%), precision (49.22%), F1 score (62.28%), recall (84.8%), and AUC (83.7%). These scores indicate that AE is quite reliable in distinguishing diabetic classes.
- **CNN**: When compared generally to metrics, it shows lower scores, with accuracy at 78.94%, precision at 48.85%, recall at 80.87%, and F1-score at 60.9%. Its AUC reflects DNN (83.02%).

Again, AE is the most balanced model compared to the other models, as evidenced by its metrics scores. AE is the model that leads in most metrics, such as accuracy, precision, F1-score, and AUC. Although DNN had an excellent recall score, AE performed better in other metrics, revealing that it is more accurate and reliable for detecting people with diabetes in imbalanced datasets. Based on the results of the imbalanced and balanced datasets, it is evident that AE is the best model for identifying cases of diabetes in both of these datasets. This is because out of the three models, AE had superior performances in accuracy, F1-score, and AUC. In contrast, DNN and CNN seemed to lag, recording some of the poorest performances in recall and F1-score metrics, especially on imbalanced data.



(a) Imbalanced data

(b) Balanced dataset

Figure 2: Results of using all features case

(ii) Case 2: Using six features out of ten features

Another case that this study used to compare the three deep learning models that were developed and the Decision Forest (DF), popularly known for outperforming most machine learning algorithms in [3]. All these four models were used to detect T2D for imbalanced and balanced data. The imbalanced data was the data that did not use SMOTE, while the balanced data was the data that was balanced using SMOTE. These four models for detecting diabetic cases were evaluated using five metrics, which are accuracy, AUC, precision, F1-score, and recall, as shown in Figure 3. Figure 3 (a) shows the DNN, AE, CNN, and DF results when using imbalanced dataset. The summary of the results of these models are as follows:

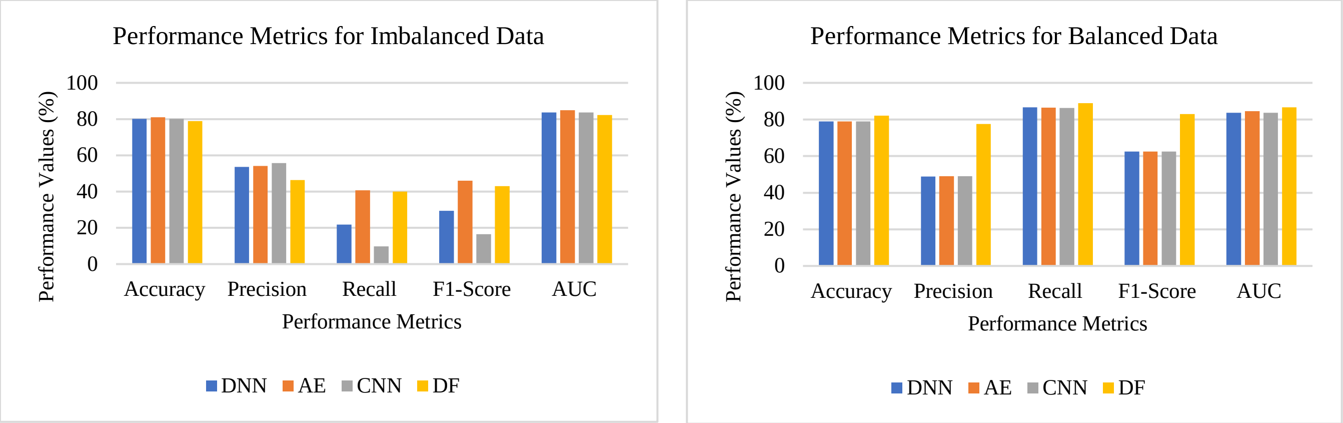
- **DNN:** This deep learning model had average scores in accuracy and AUC of 80.22 % and 83.63 %, respectively. However, DNN had poor scores in precision and recall, at 53.54 % and 21.71 %, respectively. These scores indicate that DNN has limited effectiveness in accurately identifying positive diabetes cases.
- **AE:** This deep learning model was the most promising as it recorded some of the highest scores on multiple metrics such as accuracy (81.01 %), recall (40.71 %), F1-score (46.02 %), and AUC (84.97 %). The scores of AE reveal that it has an excellent balanced performance, and it effectively identifies true positives in imbalanced datasets.
- **CNN:** This model had one of the best precision scores among the three models at 55.8 %. However, CNN recorded the worst scores in recall and F1-score of 9.73 % and 16.52 %, respectively. The results of this model demonstrate challenges with using CNN to classify positive cases of diabetes correctly.
- **DF**: Amongst the models, DF displayed the least performance in AUC (82.2%), accuracy (78.9%), precision (46.4%), F1-Score (43), and recall (40%). Although there was some balance between precision and recall it mostly lagged behind the deep learning models.

Generally, the results of the four models reveal that AE is the best for detecting diabetic cases in imbalanced datasets. This is because AE had the highest scores in all the performance metrics used in this study. As a result, it can be concluded that AE is the most reliable model for identifying true positive cases of people with diabetes and balancing class differentiation. On the contrary, DNN scores indicate that although it has an overall moderate performance, it struggles a lot with recall, making it slightly unreliable in detecting positive cases. CNN is another model with varying scores across the metrics used in this study. While this model has excellent precision performance, it had some of the worst recalls and F1 scores. The results of the recall and F1-score reveal that CNN is not a reliable classification model. DF was a model with balanced performances in precision and recall but failed to record promising scores in the other metrics. The performance of DF across all the metrics suggests that deep learning methods are more reliable, especially AE.

Figure 3 (b) shows the DNN, AE, CNN, and DF results when using balanced datasets. The summary of the scores of these four models, when balanced data is used, are as follows:

- **DNN:** This model had some of the best scores, evident in recall (86.61 %) and accuracy (78.88 %). However, DNN had moderate scores in precision and F1-score of 48.84% and 62.46%, respectively. These results suggest that DNN is best suited for identifying positive cases.
- **AE:** This model performed better than DNN in most metrics. In fact, the only metric that DNN had a better score than AE is recall, in which its performance score was 86.45 %. AE's accuracy score was 78.98 %, precision score was 48.79 %, F1-score was 62.52 %, and AUC was 84.47 %.
- **CNN:** This model had the second highest accuracy score of 79.01 %. It also had good scores in other metrics, scoring 49.02 % in precision, 86.38 % in recall, and 62.54 % in F1-score. The results of this model reveal that it is accurate in identifying T2D and classification.
- **DF:** This model outperformed all three deep learning models, recording some of the highest scores across all metrics. DF had the highest accuracy score of 82.1 %, precision score of 77.6 %, recall of 89 %, F1-score of 82.9 %, and AUC of 86.7 %.

The results of these four models reveal that DF is the superior model, recording the highest accuracy, precision, recall, F1-score, and AUC scores. This shows that DF is the best model for detecting cases of diabetes in the balanced data context. Conversely, although DNN has an excellent recall score, it has moderate scores in accuracy and precision. However, while DF outperforms AE and CNN, the two deep learning models have better scores than DNN. Compared to DNN, AE and CNN have more balanced performance profiles, making them more ideal. Overall, it can be concluded that DF is a better model for correctly identifying positive cases of diabetes and ensuring high accuracy in predictions when using balanced datasets. In analyzing the three deep learning models, AE is a better model as it had better accuracy, recall, F1-score, and AUC scores for imbalanced data. On the other hand, DF is the better model for balanced data as it performs better than the three deep learning models.



(a) Imbalanced data

(b) Balanced data

Figure 3: Results of using 6 features case.

5. Discussion

This paper uses two cases to evaluate the performances of four different models for detecting diabetes. These cases are:

(i) Case 1: Using the dataset's all features.

The full feature of the dataset case was used to develop essential insights regarding developing T2D detection systems for diabetes. The two datasets that were used were imbalanced and balanced. For the imbalanced dataset, the results of this study indicate that AE was the best for detecting diabetic cases as it was the top-performing model. The metrics it excelled in were recall and F1-score, suggesting that its biggest strength is identifying true positive cases. Such strength will be crucial in medical diagnostic scenarios characterized by high costs of false negatives. AE's performance across the used metrics further reveals that this model can capture patterns in data that are complex and non-linear. This is crucial in instances where there is a need to identify minority classes in imbalanced data. AE also had one of the best AUC scores, revealing its ability to further distinguish different classes under varying threshold settings.

DNN had excellent accuracy and precision scores but failed to record such scores in recall and F1-score. These findings suggest that DNN is a more reliable model for predicting the majority class rather than the minority class. The issue that might be leading to this trend in DNN is its inability to capture minority class characteristics due to insufficient representative capacity. This limitation can be addressed by adding regularizations or class-weighted loss functions.

CNN performed slightly similarly to DNN, especially in accuracy and AUC metrics. However, this model is somewhat outperformed by DNN in metrics such as recall and F1-score. The overall scores of CNN further reveal that it has local connectivity and weight-sharing properties, making it more suitable for various spatial data processing, such as image recognition. However, CNN's overall performance further suggests that feature dependencies in tabular diabetes data do not align well with local connectivity and weight-sharing properties.

The results of the entire balanced data of the three deep learning models reveal that they have improved performances across all the metrics. However, AE remains the superior model, recording the highest accuracy, precision, F1-score, and AUC scores. This indicates that AE is quite adaptive even in scenarios where class distribution adjustments have been made. However, the excellent recall score of DNN must be noticed, although it fails to translate the same increase in other metrics, specifically precision, and F1-score. The overall performance of DNN indicates that although it is effective in identifying positive diabetes cases, there is a high chance that it is also likely to misclassify negative cases.

(ii) Case 2: Using six out of ten features.

The process of taking six features out of the ten features makes the model performances more dynamic. In this second case, AE demonstrated superior performances on imbalanced data. These performances reveal that although the encoding-decoding mechanism reduces dimensionality in datasets, it also enhances the ability of the model to capture essence. That is why AE is often used in high-dimensional data, as it can help reduce complexity and overfitting. DNN had a moderate performance profile that provides various insights about this model. For example, when using DNN with reduced feature sets, it will be essential to include additional sophisticated feature engineering or network architecture optimization. On the other hand, CNN had some of the worst performances, which seem to be declining. The overall performance of CNN suggests that its architecture is not well suited for this tabular data.

DF had the best scores in all metrics on balanced data compared to the three deep learning models. This is specifically evident in its precision and F1-score performances in which it recorded its highest scores. This indicates that the decision trees found in DF effectively use balanced distribution to ensure the model accurately identifies class distinctions. Also, the superior performances of DF can be attributed to its decision trees' feature selection ability and ensemble techniques.

DF's excellent performance on balanced data also indicates essential insights about why it is important for models to address class imbalances. It was evident from the results of DF that the performance of traditional machine learning models can best be enhanced with SMOTE. By using the SMOTE technique, researchers were able to improve DF's performance to the extent that it outperformed AE, DNN, and CNN.

Overall, this study's findings justify why model selection is vital in data characteristics. From the findings, AE was best suited for handling imbalanced data, while DF can perform better than deep learning models if balanced datasets are used. However, one thing evident from the results is the need to use tailored approaches when selecting models for analyzing datasets. This is especially necessary if there is a need to consider data characteristics and performance metrics. Misdiagnosis of T2D in clinical settings tends to have a lot of consequences for all stakeholders. Therefore, it is essential for clinical professionals to use the best models that can help them avoid the problem above. While AE can be a helpful model for detecting T2D in imbalanced data, DF is more reliable when using balanced data. However, more research needs to be conducted on hybrid models or ensemble techniques, especially those that are trying to leverage the strengths of deep learning and traditional machine learning.

One of the most intricate and meaningful endeavors in the medical field is exploring machine learning models used to identify T2D and trying to determine performance disparate. The models used for detecting T2D were constructed, and their performance tested within the intricate medical data landscape are AE, DNN, CNN, and DF. Such examination can significantly help determine factors contributing to performance variation among these four models. Some of the critical areas of this model that this work focused on include the interaction between their complex structures and features quantity within the dataset, how SMOTE impacts data balancing employment, and the importance of hyperparameter optimization and model selection strategies. Also, this study explored how the performances of the four models were affected when the entire or subset of dataset features were used. It was believed that a solid understanding of the circumstances that influence the optimal or underperformance of AE, DNN, CNN, and DF would be provided by focusing on such areas. As a result, this study moved beyond performance metrics to explain the reasons behind operational success and challenges in T2D detection.

1- Model Complexity

- **Using All Features of the Dataset:** This was one of the cases used in the study, and it was apparent that AE was the best model when all features, balanced and imbalanced data, were used. This success is because AE can construct better-nuanced data representations. Such capability is vital if models are to detect complex patterns of T2D accurately. AE is designed to encode input data in its lower-dimensional space, after which it reconstructs it. This is an essential process since it enables models to capture a data's underlying structure effectively. Contrary to AE, DNN, and CNN lack the ability to capture subtle patterns in datasets if they lack the same dimensionality reduction and reconstruction level. In CNN, this model heavily relies on spatial relations between data features. As a result, if there are no strong spatial correlations between features in the data being used to detect T2D, CNN will reduce effectiveness.
- **Using Six Out of Ten Features:** In this case, the features of a dataset are reduced from ten to six. Despite these feature reductions, AE still demonstrated some of the best performances. The model that struggles most to perform when features are reduced is DNN. The performance struggles by DNN resulted from overfitting or generalization difficulties due to less information. CNN also recorded poor performances, which might be due to this model containing convolutional layers relied upon to extract meaningful patterns. Overall, DF had the best performance in the balanced dataset compared to these three other models. This might be because features reduction is more informative and quieter. These factors enable DF to capitalize on its ability to handle structured tabular data effectively without the complexities introduced by irrelevant features.

2- Data Balancing

- **Using All Features of the Dataset:** In the case of using all the features, SMOTE was used to balance the dataset of the study. This technique significantly improved DNN's recall performance since minority class examples to learn were increased, and bias towards the majority class was reduced. AE is another model that greatly benefitted from using SMOTE but to a lesser extent because it was already doing a fantastic job handling class imbalance. However, using SMOTE did not benefit CNN, possibly due to architectural bias. CNN is a model designed to effectively handle datasets whose class distribution dominates the learning process, which can result in a bias towards feature-rich datasets.
- **Using Six Out of Ten Features:** In this case, using SMOTE to balance data also positively impacts the recall score of DNN. DNN's performance, revealed that this model would perform optimally if data balancing techniques such as SMOTE were used. However, AE continued to demonstrate its strong ability to handle imbalanced data effectively. Regardless of these strong performances by DNN and AE, the use of SMOTE had the hugest impact on DF. Using SMOTE to balance data increased DF's performance drastically in all metrics. This demonstrates that using balancing techniques such as

SMOTE enhances the ability of DF to make more accurate predictions.

3- **Hyperparameter Tuning and Model Selection**

- **Using All Features of the Dataset:** Compared to all the other models, AE was best designed to handle dataset complexities, which is why it performed the best. On the contrary, DNN and CNN had lower performances largely because their hyperparameters were not adequately optimized for T2D detection challenges.
- **Using Six Out of Ten Features:** The three deep learning methods had ineffective hyperparameters for reduced feature cases, and, therefore, alternative tuning approaches need to be determined. However, DF has a more effective hyperparameter tuning process in the reduced feature scenario or is well designed for such datasets.

4- **Feature Selection**

- **Using All Features of the Dataset:** The model that benefited most from utilizing all dataset features is AE. The reason behind this is that AE can greatly reduce internal dimensionality. As a result, it can extract relevant patterns in a dataset without the less informative features overwhelming it. Conversely, CNN and DNN are not designed to adequately handle all features if some do not contribute to their performance.
- **Using Six Out of Ten Features:** Reduction of some features results in noise removal from datasets that benefit DNN and CNN. Whereas AE uses the encoding process as its internal feature selection, which enables this model to perform optimally even with reduced features. However, DF is better used for reduced features as this process reduces data complexity, allowing it to use a more focused set of features.

Overall performances of the developed models reveal that AE is well designed to perform optimally in various scenarios. However, this is not the case for DNN and CNN, as these models require feature selection to be conducted carefully and class balances to be performed optimally. For the DF, this model is best suited for a balanced dataset that contains fewer features. Therefore, traditional machine learning models can be more effective by ensuring they are tuned appropriately, and the data being used is well-prepared.

6. Conclusions

Early detection of Type 2 Diabetes (T2D) is critical for implementing appropriate treatment strategies and lifestyle adjustments. These will help decelerate and prevent the advancement of the condition. Studies in machine learning and medical datasets have overlooked the Saudi population on T2D detection. This study addresses this research gap by developing and comparing three deep learning methods and the traditional classifier. The deep learning methods developed are Deep Neural Network (DNN), Autoencoder (AE), and Convolutional Neural Network (CNN), while the traditional classifiers used are the Decision Forest (DF). As a result, this study contributes to the literature by providing a comparative analysis of different models, an aspect that was missing. The performance metrics used to evaluate these four models are accuracy, precision, recall, F1-score, and AUC. The findings of this study revealed that AE outperformed other models in the case of using all features of the dataset for imbalanced and balanced data with the highest accuracy of 81.12% and 79.16%, respectively. For the case of using 6 features, the results showed that AE achieved the highest accuracy of 81.01% with imbalanced data compared to DF and DF achieved the highest accuracy of 82.1% with balanced data. Findings suggest that AE is superior when imbalanced and balanced datasets with all features are used. However, DF is a more reliable model when using balanced data with fewer features. Overall, AE is more useful in detecting T2D cases that are yet to be detected. These results signify that the model can be a valuable tool for healthcare practitioners in Saudi Arabia. The models developed in this research may be employed as a screening instrument in public health initiatives. It will also support awareness goals about T2D. This in turn will promote healthier lifestyles among different populations. Future investigations should be expanded by incorporating a larger and more diverse dataset. This should encompass additional variables like genetic factors and dietary patterns. Alternative deep learning architectures and integrating advanced techniques such as data augmentation and transfer learning will lead to improved results. Future research studies will help authenticate the models using more extensive datasets and assess their practical viability within clinical contexts.

References

[1] American Diabetes Association. (2019). Standards of Medical Care in Diabetes-2019 Abridged for Primary Care Providers. *Clinical Diabetes, 37*(1), 11-34.

[2] Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., ... & IDF Diabetes Atlas Committee. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes Research and Clinical Practice, 157,* 107843.

[3] Syed, A. H., & Khan, T. (2020). Machine learning-based application for predicting risk of type 2 diabetes mellitus (t2dm) in saudi arabia: A retrospective cross-sectional study. *IEEE Access, 8,* 199539-199561.

[4] Gollapalli, M., Alansari, A., Alkhorasani, H., Alsubaii, M., Sakloua, R., Alzahrani, R., ... & Albaker, W. (2022). A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: Pre-diabetes, T1DM, and T2DM. *Computers in Biology and Medicine, 147,* 105757.

[5] Alwin Robert, A., Abdulaziz Al Dawish, M., Braham, R., Ali Musallam, M., Abdullah Al Hayek, A., & Hazza Al Kahtany, N. (2017). Type 2 diabetes mellitus in Saudi Arabia: major challenges and possible solutions. *Current Diabetes Reviews, 13*(1), 59-64.

[6] Tabák, A. G., Herder, C., Rathmann, W., Brunner, E. J., & Kivimäki, M. (2012). Prediabetes: a high-risk state for developing diabetes. *Lancet, 379*(9833), 2279-2290.

[7] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347-1358.

[8] Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2017). Artificial intelligence in precision cardiovascular medicine. *Journal of the American College of Cardiology, 69*(21), 2657-2664.

[9] American Diabetes Association. (2020). 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2020. *Diabetes Care, 43(Supplement 1),* S14-S31.

[10] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal,*

15, 104-116.

[11] Li, Y. (2022, January). Research and application of deep learning in image recognition. In 2022 IEEE 2nd international conference on *Power, Electronics and Computer Applications (ICPECA)* (pp. 994-999). IEEE.

[12] Otter, D. W., Medina, J. R., & Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *Ieee Transactions on Neural Networks and Learning Systems, 32*(2), 604-624.

[13] Miglani, A., & Kumar, N. (2019). Deep learning models for traffic flow prediction in autonomous vehicles: A review, solutions, and challenges. *Vehicular Communications, 20*, 100184.

[14] Justesen, N., Bontrager, P., Togelius, J., & Risi, S. (2019). Deep learning for video game playing. *IEEE Transactions on Games, 12*(1), 1-20.

[15] Mittal, S., & Hasija, Y. (2020). Applications of deep learning in healthcare and biomedicine. *Deep Learning Techniques for Biomedical and Health Informatics*, 57-77.

[16] Lee, W., Seong, J. J., Ozlu, B., Shim, B. S., Marakhimov, A., & Lee, S. (2021). Biosignal sensors and deep learning-based speech recognition: *A review. Sensors, 21*(4), 1399.

[17] Roy, A., Sun, J., Mahoney, R., Alonzi, L., Adams, S., & Beling, P. (2018, April). Deep learning detecting fraud in credit card transactions. In 2018 *Systems and Information Engineering Design Symposium (Sieds)* (pp. 129-134). IEEE.

[18] Nuruzzaman, M., & Hussain, O. K. (2018, October). A survey on chat- bot implementation in customer service industry through deep neural networks. In 2018 IEEE 15th *International Conference on e-Business Engineering (ICEBE)* (pp. 54-61). IEEE.

[19] Ríos-Sánchez, B., Costa-da-Silva, D., Martín-Yuste, N., & Sánchez-Ávila, C. (2019). Deep learning for facial recognition on single sample per person scenarios with varied capturing conditions. *Applied Sciences, 9*(24), 5474.

[20] Ahmad, H. F., Mukhtar, H., Alaqail, H., Seliaman, M., & Alhumam, (2021). Investigating health-related features and their impact on the prediction of diabetes using machine learning. *Applied Sciences, 11*(3), 1173.

[21] Alassaf, R. A., Alsulaim, K. A., Alroomi, N. Y., Alsharif, N. S., Aljubeir, M. F., Olatunji, S. O., ... & Alturayeif, N. S. (2018, April). Preemptive diagnosis of diabetes mellitus using machine learning. In 2018 21st *Saudi Computer Society National Computer Conference (NCC)* (pp. 1-5). IEEE.

[22] Alanazi, A. S., & Mezher, M. A. (2020, September). Using machine learning algorithms for prediction of diabetes mellitus. In 2020 *International Conference on Computing and Information Technology (ICCIT-1441)* (pp. 1-3). IEEE.

[23] Daghistani, T., & Alshammari, R. (2016). Diagnosis of diabetes by applying data mining classification techniques. *International Journal of Advanced Computer Science and Applications, 7*(7), 329-332.

[24] Jaber, F. A., & James, J. W. (2023). Early prediction of diabetic using data mining. *SN Computer Science, 4*(2), 169.

[25] Almutairi, E. S., & Abbod, M. F. (2023). Machine learning methods for Diabetes prevalence classification in Saudi Arabia. *Modelling, 4*(1), 37- 55.

[26] Alam, T. M., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Baig, T. I., ... & Abbas, Z. (2019). A model for early prediction of diabetes. *Informatics in Medicine Unlocked, 16*, 100204.

[27] Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big data, 6*(1), 1- 19.

[28] Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing, 73*, 1-15.

[29] Pathirage, C. S. N., Li, J., Li, L., Hao, H., Liu, W., & Ni, P. (2018). Structural damage identification based on autoencoder neural networks and deep learning. *Engineering structures, 172*, 13-28.

[30] Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing, 234*, 11-26.

[31] Kiranyaz, S., Ince, T., & Gabbouj, M. (2015). Real-time patient-specific ECG classification by 1-D convolutional neural networks. *Ieee Transactions on Biomedical Engineering, 63*(3), 664-675.

[32] Muni Kumar, N., & Manjula, R. (2019). Design of multilayer perceptron for the diagnosis of diabetes mellitus using keras in deep learning. In Smart Intelligent Computing and Applications: *Proceedings of the Second International Conference on SCI 2018, Volume 1* (pp. 703-711). Springer Singapore.

[33] Zhou, H., Myrzashova, R., & Zheng, R. (2020). Diabetes prediction model based on an enhanced deep neural network. *EURASIP Journal on Wireless Communications and Networking, 2020*, 1-13.

[34] Tabares-Soto, R., Orozco-Arias, S., Romero-Cano, V., Bucheli, V. S., Rodríguez-Sotelo, J. L., & Jiménez-Varón, C. F. (2020). A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data. *PeerJ Computer Science, 6*, e270.

[35] Pandiya, M., Dassani, S., & Mangalraj, P. (2020). Analysis of deep learning architectures for object detection-a critical review. 2020 *IEEE- HYDCON*, 1-6.

[36] Rahman, M., Islam, D., Mukti, R. J., & Saha, I. (2020). A deep learning approach based on convolutional LSTM for detecting diabetes. *Computational Biology and Chemistry, 88*, 107329.

[37] Naz, H., & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes & Metabolic Disorders, 19*, 391-403.