Artificial Intelligence Evolution
http://ojs.wiserpub.com/index.php/AIE/

UNIVERSAL WISER PUBLISHER

Research Article

Diabetes Prediction Tool under System on Chip Using Machine Learning Method

Nour El-Houda Benalia[1*], Soumaya Ferhat Taleb[2], Amel Tibhirt[2], Karima Chenikha[2], Rabah Sadoun[2], Ahlem Bentrah[3]

[1]Preparatory classes Department, Polytechnic National School, El-Harrach, Algiers, Algeria
[2]Electronics Department, Polytechnic National School, El-Harrach, Algiers, Algeria
[3]Mostefa Ben Boulaid, Batna 2 University, Batna, Algeria
Email: nour_el_houda.benalia@g.enp.edu.dz

Received: 31 August 2022; **Revised:** 15 October 2022; **Accepted:** 16 October 2022

Abstract: Extraordinary advances in biotechnology and health sciences have brought significant generation of data, such as genetic data and clinical information, generated from huge electronic health records. Data analysis is a process of studying and identifying hidden patterns from large amounts of data and drawing conclusions. In health care, this analytical process is carried out using machine learning (ML) algorithms to analyze medical data to build Machine Learning models to transform all available information into valuable knowledge. Nowadays, diabetes has become a common disease among young people, elderly people, and even children. According to the World Health Organization (WHO) report, by 2025, this number is expected to exceed 380 million. This research work performs a comparison of 5 classification techniques, namely Naive Bayes (NB), Bagging (J48), Decision Tree (J48, C4.5), K Nearest Neighbors (KNN), and Support Vector Machine (SVM), to detect diabetes at an early stage. The performances of the five algorithms are evaluated and compared on various measures like accuracy, precision, and recall. The experiments were conducted based on the diabetes database, the source from the National Institute of Diabetes, Digestive and Kidney Diseases, showing the effectiveness of using the Decision Tree (DT) technique. This effectiveness led to the choice of this method. After obtaining the DT model of the problem, the main task facing us in this work is the hardware implementation of this model. Indeed, this forecasting system can be used in other complementary works as a processing unit in a cloud, to be able to manage numerous requests. The considered solution is the hardware implementation on a Field Programmable Gate Array (FPGA) board.

***Keywords*:** decision tree, diabetes, machine learning, Scikit Learn, System on Chip

1. Introduction

Diabetes mellitus is a chronic disease with severe complications that affects millions of people around the world. It is found in almost all populations and is emerging as a growing problem in developing countries. The cost of illness, in terms of suffering, health care, and loss of life, is high. Better knowledge of the causes and mechanisms of the main types of diabetes mellitus now forms the basis of prevention activities [1].

DOI: https://doi.org/10.37256/aie.3220221878

We can categorize the latter into two classes. The traditional one is fundamentally based on an approach based on recommendations (healthy diet, correct weight, regular physical activity) coupled with awareness and screening campaigns and the other one is based on the use of new information technologies and communication, both in terms of data collection, storage and above all processing. In the latter case, Artificial Intelligence (AI) presents itself as an emerging solution for the analysis and exploitation of data. Its application to the field of diabetes treatment is proving to be relevant.

AI is defined more generally as the ability of a machine to act on its own or under human control to reproduce actions or functions that are usually performed by humans. Today, we find it in several areas; we cite computer equipment, connected objects, network applications, transport, and others. The medical field is proving to be a major field of application. The application of AI to medicine offers a more than the interesting prospect of application of these new technologies, whether it is to strengthen the link between patients and doctors, to make faster and more precise diagnoses, or to optimize the creation of new treatments. The purpose is to refocus on human health (the comfort of life, prevention, longevity, etc.) as an individual but also, by extrapolation, on the reduction of economic costs for society in terms of taking into the patient load.

In this sense, a great deal of work has been done to propose prediction systems for many diseases, which is the main task of these efforts. In the paper of Ba-Alwi and Hintaya [2], a comparative analysis of many data mining algorithms for the diagnosis of hepatitis disease was presented. They found that the Naive Bayes is the best classification algorithm used in the rough set technique as it offers high accuracy in the least possible time.

Another work was done by Ibrahim [3]. The proposed paper suggested a system in which Artificial Neural Network is used for forecasting the defervescence day of fever in patients of dengue. For detection, the proposed approach relies solely on clinical signs and symptoms. The data are gathered from 252 hospitalized patients, in which 4 patients are having DF (Dengue fever) and 248 patients are having DHF (dengue hemorrhagic fever). MATLAB's neural network toolbox is used. In this experiment, the Multi-layer feed-forward neural network (MFNN) algorithm is applied. Day of defervescence of fever is accurately predicted by MFNN in DF and DHF with 90% correctness.

Sharma et al. [4] are the owner of our third work, where they developed a data mining model to predict heart disease efficiently. It primarily assists medical practitioners in making effective decisions based on the parameters provided. The author used the Cleveland dataset from UCI, as well as age, gender, sex, resting blood pressure, chest pain, serum cholesterol, fasting blood sugar, etc. as attributes. They've also separated the datasets into two halves, one for testing and the other for training. They used a 10-fold method to find the accuracy.

Recently, Deepika et al. [5] proposed predictive analytics to prevent and control chronic disease with the help of machine learning techniques such as naive Bayes, support vector machine, decision tree, and artificial neural network and they have used UCI machine learning repository datasets to calculate the accuracy. Among these, the Support vector machine gives the best accuracy of 95.55%.

Finally, we can cite the work of Sisodia et al. [6]. The first motivation of this study was the design of a model that can predict the possibility of having diabetes with maximum accuracy. For this, three machine-learning classification algorithms have been used to detect these diseases. The obtained results show Naive Bayes outperforms with the highest accuracy of 76.30% compared to Decision Tree and SVM algorithms. These results have been verified using Receiver Operating Characteristics (ROC) curves in a proper and systematic manner.

In this work, we are focusing on diabetes prediction, where our approach to building our diabetes prevention system revolves around two phases, namely, obtaining the model and implementing it on a technological target. The first phase involves the combination of the database with the choice and the use of an appropriate algorithm as a platform to generate it. In the second phase, the choice of the technical target remains dependent on the performance of the goal according to the recommended reaction time (therefore alert in the case of prevention). A flexible technical target associated with an adapted development approach could be a judicious choice for the development of any AI solution, in the cloud, at the edge, and any end node (i.e., the patient through his smartphone or any on-board sensor) as well.

The management of diabetes prevention by using new information and communication technologies is based on the process described in the following two figures.

Figure 1 explains the role and importance of data in the search for a prevention model [7].

Data mining proposes to use a set of algorithms from various scientific disciplines, such as statistics, artificial intelligence, or computer science, to build models from data, that is to say, find interesting structures or patterns

according to criteria fixed in advance, and extract as much knowledge as possible.

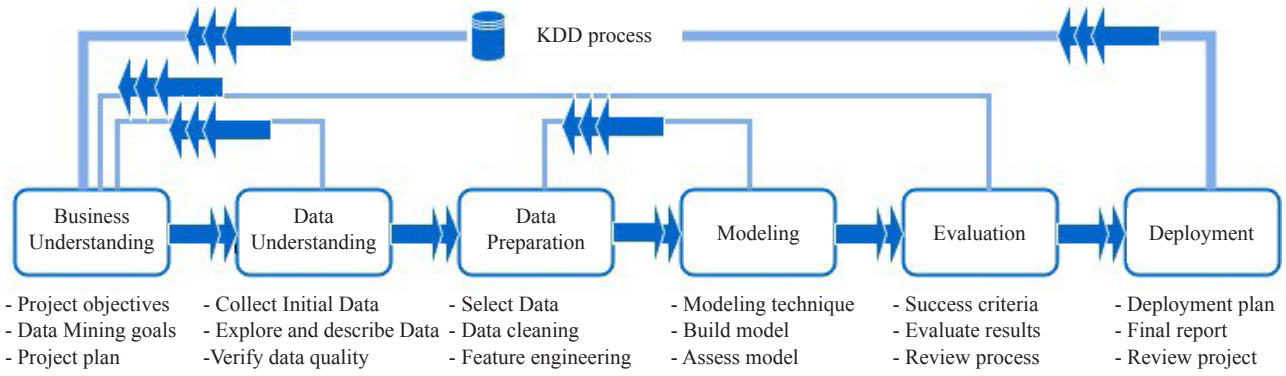

Figure 1. Explanatory diagram of the KDD process [7]

Figure 2 highlights the importance of treatment in the phases of both model research and deployment.

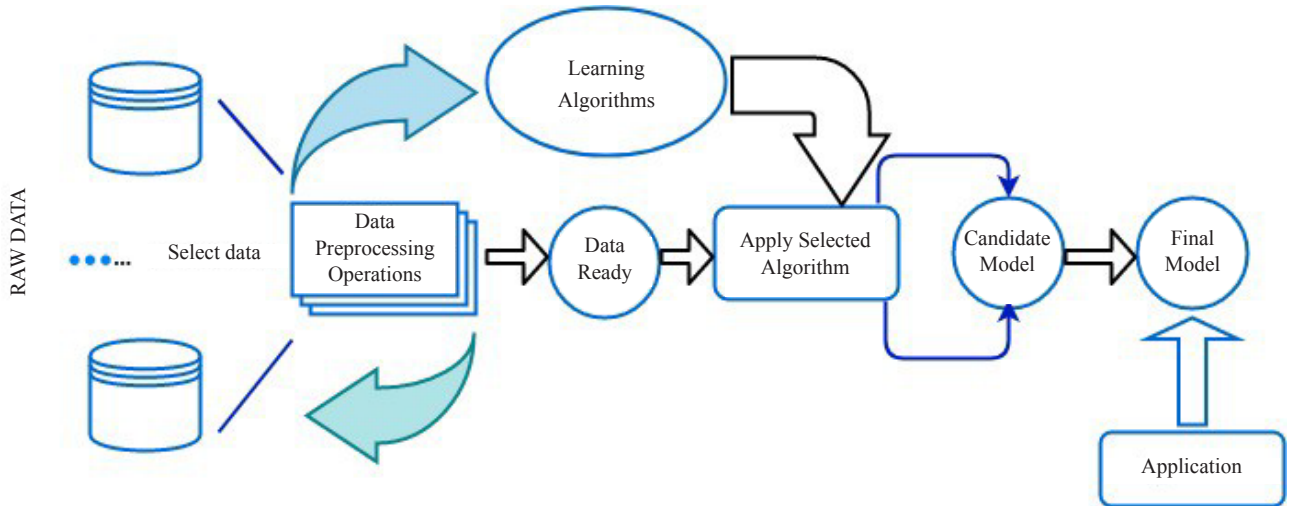

Figure 2. Explanatory diagram of the learning algorithm process

The deployment can be located both near the sensors (and therefore the patient) and at the level of data centers. Real-time monitoring of diabetes prevention depends on the amount of data to be processed (during the learning phase) and the detection phase may require substantial processing times depending on the size of the data to be processed and on the complexity of the model to be implemented.

The novelty of our work lies in the combination of the methodology described by the two figures above (KDD) with that dedicated to the development of circuits.

Coupled with the open source solutions well chosen based on the objective evaluation criteria, we are certain that the application of prototyping applied to Machine Learning (ML) will be of proven effectiveness. This will be at the heart of the question that we propose to resolve in this work.

To describe our work and the solution we implemented, we organized our paper into five sections.

The second section describes diabetes, methods of preventing diabetes, and the importance of diagnosing this

disease early. This section will be preceded by an introductory section.

The third section concerns the methodology followed to obtain the most precise prevention model. We have imposed two degrees of freedom on ourselves, namely, the algorithm used and the best ML platform that carries it.

In the fourth section, we present a method for obtaining the model describing the behavior of diabetes as a function of medical predictor variables using the ML algorithm deduced from the previous section. Our design flow for the implementation of the model obtained on a hardware target will occur; a SoC is our choice. This SoC could equally well carry our model either in the form of a hardware accelerator, or exclusive software implementation.

We will end this work with a conclusion that analyzes the obtained results and provide the prospects that we envisage.

2. Diabetes and its prevention method: an overview

Diabetes is a chronic disease also known as the "silent killer". Diabetes is triggered when there is an absolute or relative deficiency of insulin production in the body. In the absence of diabetes, insulin is able to do its job well and cells get the energy they need to function. In the absence or insufficiency of insulin, glucose cannot be used as fuel for the cells. It then accumulates in the blood and increases the blood sugar (hyperglycaemia) [8]. The effects of diabetes mellitus include long-term damage, dysfunction and failure of various organs. This has caused a significant increase in mortality among patients.

2.1 *Diabetes types and diagnosis*

There are several types of diabetes, ranging from the most common to the rarest:

1) Type I diabetes, also known as insulin-dependent diabetes, occurs when the body does not permanently produce sufficient insulin. So when the body stops secreting insulin, blood sugar levels are higher than normal. Type 1 diabetes will be diagnosed if his fasting blood sugar is above 1.26 g/l or if his blood sugar, at any time of the day, is above 2 g/l [9].

2) Type II diabetes, unlike type 1 diabetes, which in most cases, occurs in childhood and young adults, type 2 diabetes sets in a "mature" adult. Since it is rarely accompanied by symptoms in its early stages, it can go unnoticed for several years. Type 2 diabetes can be distinguished by the less sensitive of cells to the action of insulin (insulin resistance), which increases the level of glucose in the blood. Thus, fasting blood sugar becomes chronically greater than 1.26 g/l.

3) Gestational Diabetes. This type of diabetes begins during pregnancy (usually in the 2nd trimester) and then disappears in the weeks following childbirth. It is characterized by hyperglycemia (excess blood sugar) due to temporary resistance to insulin. This resistance is linked to the hormones produced by the placenta during pregnancy.

The main way to know if you have diabetes is through a blood test. 50% of people with diabetes do not know it and most of them are affected by diabetes type 2. We know that the faster the diagnosis is made, and the management of diabetes begins earlier, the greater the associated risks of this disease are reduced It is important to know that there is no specific and exclusive cause, but rather a set of factors that promote the development of diabetes. It can be of genetic origin; a history of diabetes of the same type is often present in the family, or behavioral factors, an unbalanced diet, and lack of physical activity contribute to the development of diabetes type 2 [10].

2.2 *Methods of diabetes prevention*

Even as research progress and new drugs emerge, the adage that "prevention is better than cure" remains key to diabetes management.

Avoiding diabetes is first of all to find the right way to remedy it, in this part we will discuss two methods, starting with traditional means, such as a healthy lifestyle which includes a healthy diet, correct weight, and regular physical activity, but also awareness campaigns, screening, and information. The second method concerns technological manners, based on the use of ML which emerges to be the new era of modern medical science.

2.2.1 *Traditional methods for preventing diabetes*

The reality of diabetes is unrecognized, underestimated, and even ignored in a sadly and dangerously shared indifference. More than ever, informing populations, training caregivers, and access to healthcare, are fundamental issues, ignored by public opinion and neglected by governments, in many countries around the world.

The first steps in living with diabetes are food-related. Hygienic dietary rules are essential. Physical activity also has an important role to play. And, above all, there are different lifelong treatments that are effective.

The standard treatment for diabetes, which should be started before any other, is lifestyle modification, including [11] weight loss when it's needed, regular physical activity, a balanced diet, management of the emotional factor (Stress), and smoking. Several studies show that smokers are more likely to develop type 2 diabetes than non-smokers or people who have quit smoking.

But also awareness campaigns that should encourage people to be screened, you should know that in several cases of diabetes the reason can be genetic. The sooner the disease is detected, even before symptoms appear, and the sooner you intervene to restore normal blood sugar levels, the lower the risk of complications will be. The classic means are numerous, but the goal remains the same.

2.2.2 *Machine learning for diabetes prevention*

The application of ML methods in bio-sciences is more than ever essential to intelligently transform all available information into useful knowledge. The application of ML and methods in diabetes research is a key approach for using the large amounts of available diabetes data to extract knowledge. In this article, we focus on the prevention of diabetes as a classification problem using machine learning.

3. Machine learning methodologies for diabetes prevention

In machine learning, a classification problem takes the form of a set of data, containing examples from the observation of a phenomenon. Each example consists of a description and a label. A learning algorithm analyzes its data to build a classifier. This classifier then has the task of labeling new examples based on their description.

In order to study the classification problem we are interested in, we will first define what machine learning is. Then we will present the state of the art and we will compare certain works concerning the application of learning methods in the medical field, in particular in the field of diabetes. Finally, we will present the database used and, based on this, we will compare the performance of several ML algorithms using several tools in order to choose the most appropriate one for our study case.

3.1 *Related works and the choice of the appropriate algorithm*

ML is a technology that is increasingly used in all industries. In the transport industry for the development of a driver-less navigation system, in the banking sector where one seeks to estimate the capacity of a person to repay a loan or in the medical sector where learning machines help diagnose cancer, and diabetes, of which the prediction of the latter constitutes the main point in this work. In the context of the prediction of diabetes, several works have been carried out in this field. In [12], Song et al. described and explained the different classification algorithms using different parameters such as glucose, blood pressure, skin thickness, insulin, BMI, diabetes, and age. In this research, the researchers used only small samples of data to predict diabetes. In this article, the algorithms used five different algorithms: GNNs (Graph Neural networks), ANN (Artificial Neural Network), SVM (Support Vector Machine), EM (expectation-maximization), and logistic regression. Finally, the researchers concluded that the ANNs provided high accuracy for predicting diabetes. In the study by Pradeep and Naveen [13], the performance of machine learning techniques was compared and measured according to their accuracy. The precision of the technique varies before pretreatment and after pretreatment, as identified in this study. This indicates that in disease prediction, preprocessing of the data set has its own impact on the performance and accuracy of the prediction. Researchers have shown that the decision tree technique provides better precision in this pretreatment study to predict diabetes disease. Whereas the Random forest and the SVM techniques provide better prediction after pretreatment in this study using the diabetes dataset. Meng et al. [14] used different data

mining techniques to predict diabetes disease using real-world data sets by collecting information from a distributed interrogator. The researchers compared three techniques: ANN, logistic regression, and J48. Finally, it was concluded that the J48 machine learning technique provides efficient and better precision. In [15], three machine learning classification algorithms, namely, Decision Tree, SVM and Naive Bayes, are used in this experiment to detect diabetes at an early stage. The experiments are performed on the Pima Indians Diabetes (PIDD) database which comes from the UCI machine learning repository. The performance of the three algorithms is evaluated on various metrics such as precision, precision, F-measure, and recall. Precision is measured in correctly and incorrectly classified instances. The obtained results show that Naive Bayes outperforms with the greatest precision of 76.30% compared to other algorithms. A comparison was made too among the J48, CART, SVM, and kNN algorithms for the prevention of diabetes in [16], where the authors used a database containing 10 attributes for 545 patients and the results with the WEKA tool showed that the J48 algorithm gave better results with an accuracy of 67.15%. Kaur and Kumari [17] classified the patients into diabetics and non-diabetics. They developed and analyzed five different predictive models using *R* data manipulation tool. To this end, they used supervised machine learning algorithms, namely a linear kernel support vector machine (SVM-linear), a radial basis function (RBF) kernel support vector machine, a nearest neighbor (k-NN), an artificial neural network (ANN) and a multi-factor dimensionality reduction (MDR). Hossain et al. [18] have done a study that extracted some social network-based features from the final disease network and some demographic characteristics directly from a dataset. These risk factors were then used to develop six machine-learning prediction models to assess the risk of CVD in patients with T2D. The accuracy of the classifiers ranged from 79% to 88%, showing the potential of the network- and machine learning-based risk prediction model utilizing administrative data. The proposed risk prediction model could be useful for medical practice as well as stakeholders to develop health management programs for patients at a high risk of developing chronic diseases.

The hardest part of solving a ML problem is often finding the right estimator for a given application or problem. Different estimators are better suited for different types of data and for different problems. To choose the most appropriate algorithm for our problem, we first studied the previous work reported in the literature on the prevention of diabetes with ML, then we strengthened our choice by making a comparison between different classification algorithms, using our database on different platforms.

Table 1. Distribution of related works according to classifier type used

Ref.	Methodologies	Findings
[14]	ANN, logistic regression, and J48	It was concluded that the J48 machine learning technique provides efficient and better precision.
[13]	Machine learning techniques was compared and measured according to their accuracy.	It has been concluded that the decision tree technique provides better precision in this pre-pretreatment study to predict diabetes disease. Whereas, the Random forest and the SVM techniques provide better prediction after pretreatment in this study using the diabetes dataset.
[16]	J48, CART, SVM, and kNN algorithms	The results with the WEKA tool show that the J48 algorithm gives better results with 67.15% accuracy.
[12]	GNNs (Graph Neural Networks), ANN (Artificial Neural Network), SVM (Support Vector Machine), EM (expectation-maximization), and logistic regression.	It has been concluded that the ANNs provided high accuracy for predicting diabetes.
[15]	Decision Tree, SVM, and Naive Bayes	It was concluded that Naive Bayes outperforms with the greatest precision of 76.30% compared to other algorithms.
[17]	Supervised machine learning algorithms, namely a linear kernel support vector machine (SVM-linear), a radial basis function (RBF) kernel support vector machine, a nearest neighbor (k-NN), an artificial neural network (ANN), and a multifactor system. Dimensionality reduction (MDR)	It was concluded that on the basis of all the parameters SVM-linear and k-NN are two best models to find that whether patient is diabetic or not.
[18]	They have done a study that extracted some social network-based features from the final disease network and some demographic characteristics directly from a dataset. These risk factors were then used to develop six machine learning prediction models to assess the risk of CVD in patients with T2D.	The classifiers accuracy ranged from 79% to 88% shows the potential of the network- and machine learning-based risk prediction model utilizing administrative data.

From the cited works, we noticed that machine learning techniques are strongly recommended for dealing with classification problems. Moreover, most previous studies (in Table 1) were based on the idea of building a global model for the treated classification problem. Therefore, after a big study in this sense, we found that the application of the DT algorithm to classify data of patients can overcome the previous gaps. This algorithm is known as very powerful in dealing with big data problems and doesn't require feature standardization or normalization. In the next sections, we will detail the steps of DT formulation and explain the results with a case study.

3.1.1 *Performance evaluation of different ML techniques: MATLAB, WEKA, Python*

The dataset source is from the National Institute of Diabetes and Digestive and Kidney Diseases. The purpose of the dataset is to predict whether a patient is diabetic or not, based on certain diagnostic metrics included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients on this basis were women at age 21 and older.

The datasets include 768 instances, with nine attributes that present eight medical predictor variables and a target variable, which indicates whether the patient has diabetes or not. The predictor variables include:

1) Number of times pregnant.
2) Plasma glucose concentration in 2 hours in an oral glucose tolerance test.
3) Diastolic blood pressure (mm·Hg).
4) Triceps skin fold thickness (mm).
5) Serum insulin 2 hours (mu U/ml).
6) Body mass index (kg/m^2).
7) Diabetes pedigree function.
8) Age (Years).
9) Class variables (0 or 1).

3.1.1.1 *Performance evaluation with MATLAB (Classification Learner)*

This application allows you to explore supervised machine learning using various classifiers, explore data, train models, and evaluate results. The classifiers included in this application are DT, SVM, logistic regression, KNN, and ensemble classification [19].

Learning is performed by providing a known set of input data and known responses for those inputs (classes in our study). This data is used to train a model. To use the model with new data, the application allows you to export the model with workspace and test new data.

Table 2. Comparison between classification algorithms in MATLAB

Algorithm	Accuracy (%)	AUC
DT (Depth = 2, CART)	73.3%	0.79
DT (Depth = 5, CART)	76.0%	0.81
DT (Depth = 12, CART)	71.5%	0.73
Bagged Trees	75.5%	0.80
SVM	75.9%	0.81
KNN	70.4%	0.72

We took 80% of the data for training and 20% for testing. Table 2 gathers the results of the performance evaluation in terms of accuracy and recall for several classifiers under MATLAB. We have considered classifiers of type DT (Decision Tree) and one of its sets (bagging), the SVM (Support Vector Machine), the KNN (K Nearest Neighbors).

Comparison results under the Learner Classification application show that all algorithms perform well in terms of both accuracy and AUC parameters. The performance marked using decision trees with a depth of five is slightly better than other algorithms.

3.1.1.2 *Performance evaluation with WEKA*

WEKA [20], is a ML software suite written in Java and developed at the University of Waikato in New Zealand and which is available under the GNU General Public License.

We took 80% of the data for training and 20% for testing. Table 3 summarizes the performance evaluation results in terms of accuracy, recall, and precision for several classifiers under WEKA. We have considered classifiers of type DT and its sets (bagging), SVM, KNN, and Naive Bayes.

Table 3. Comparisons among classification algorithms under WEKA

Algorithm	Accuracy (%)	AUC
Naive Bayes	76.30%	0.819
Bagging (J48)	75.20%	0.810
Decision Tree (J48, C4.5)	74.21%	0.764
KNN	70.18%	0.650
SVM	69.27%	0.570

We note that with the use of the Diabetes Database under WEKA, all algorithms perform well in terms of accuracy and AUC.

3.1.1.3 *Performance evaluation with Python (Scikit Learn)*

Scikit-learn [21], is a library in Python which provides numerous unsupervised and supervised learning algorithms. Python is a high-level programming language. Created by Guido van Rossum and first published in 1991. Python offers a dynamic type system and automatic memory management. It supports multiple programming paradigms including object-oriented, imperative, functional, and procedural, and has an extensive and comprehensive standard library.

We took 80% of the data for training and 20% for testing. Table 4 summarizes the results of performance evaluation in terms of accuracy, and recall for several classifiers with the Scikit Learn library.

The comparison results with the use of the Scikit Learn library show that all the algorithms present good performance in terms of the two parameters, accuracy and AUC. The performance marked by the use of decision trees with a depth of five is slightly superior to other algorithms.

Table 4. Comparisons among classification algorithms under Scikit Learn

Algorithm	Accuracy(%)	AUC
DT (Depth = 3, CART)	86.36%	0.852
DT (Depth = 5, CART)	88.31%	0.873
DT (Depth = 12, CART)	78.50%	0.74
Bagged Trees	83.76%	0.77
SVM	83.76%	0.742
KNN	81.16%	0.714

3.1.2 *Comparison of the results obtained for DT*

Table 5 summarizes the results of the performance evaluation in terms of accuracy and AUC for the DT-type classifier under the three platforms. The results show that the Scikit Learn assessment gives better results in terms of AUC and accuracy.

We conclude that DTs give a better prediction compared to other algorithms, hence the use of this algorithm to derive the diabetes prevention model.

Table 5. Performance comparison for DT on the three platforms

Platforms	Matlab (DT-CART) (%)	WEKA (J48-C4.5)	Scikit Learn (DT-CART))
Accuracy (%)	75.40%	74.21%	88.31%
AUC	0.8	0.764	0.873

3.2 *Decision tree for diabetes prevention*

After the study made in the previous section, which concerns a comparison of different classifiers and techniques used by researchers in the field of classification and learning in order to prevent diabetes, our choice was fixed on the DTs. Learning by using DTs is a method commonly used in data mining. The goal of such a method is to create a model that predicts the value of a target variable by function of several input variables (attributes).

In this part of the present paper, we will present some theoretical notions of decision trees as well as the advantages of their uses. We will then justify the platform, by which the model will be obtained. Finally, this part will be concluded by presenting the drawn model and its parameters.

3.2.1 *Description*

The DTs are a very effective method of supervised learning [22]. This involves partitioning a set of data into groups as homogeneous as possible from the point of view of the variable to be predicted. We take a set of categorical data as the input and output a tree that looks much like an orientation diagram, where each endpoint (leaf) represents a decision (a class) and each non-end (internal) node represents a test. Each leaf represents the decision to belong to a data class that verifies all the tests of the path leading from the root to that leaf (Figure 3). The most known benefits of using DT are:

- Simplicity of understanding and interpretation;
- Number of tests limited by the number of attributes;
- Efficient construction using optimization learning;
- Efficient on large data sets: the method is relatively economical in terms of computing resources;
- The model can handle both numeric values and categories.

3.2.2 *Decision Tree learning algorithms*

The different ways of distributing attributes on the DT (fractionation criterion) and choosing the root node are the subject of many DT algorithms.

1) ID3 Algorithm: This algorithm is based on the calculation of Shannon's entropy and information gain.

2) C4.5/C5 Algorithm: C4.5, successor to ID3, uses an extension to gain information called "Gain Ratio", C4.5 was replaced in 1997 by a C5.0/See5 (C5.0 for Unix/Linux, See5 for Windows) [23]. C5.0 is the classification algorithm that applies in large databases. The C5.0 is better than C4.5 in efficiency, memory, and speed.

3) CART Algorithm: this algorithm constructs a strictly binary DT with exactly two branches for each decision node. CART uses the Gini index [24] as a criterion for dividing the attributes and their distribution on the tree.

3.2.3 *The drawn model and its parameters*

Although there is no history of comparing ML platforms, in a recent study [15], researchers at Arizona State University found that comparison on the basis of precision and F-score shows that all platforms work equally well.

The Iman-Davenport [15] test statistic, calculated from the mean ranks of AUC values, rejects the previous hypothesis when considering their best classification accuracy for all algorithms.

Holm's test shows that Azure ML [15], Python (Scikit Learn), and SAS perform best of the five platforms when the best AUC score on a set of algorithms is used as a performance metric. From the results obtained in the article on the comparison of platform performances, and the great ability to adjust the use of Python with the Scikit Learn libraries, we choose the latter to obtain the DT model for our database of data.

3.2.3.1 *Parameters chosen for obtaining the DT*

As presented above, the Scikit Learn library was chosen for the large documentation that the latter presents, its performance ensured by a very recent study but also its great ability to adjust, such as the depth of the DT that we have set at 5 (depth giving the best performance) and the CART algorithm as the DT construction algorithm, which uses the Gini Index as a division criterion.

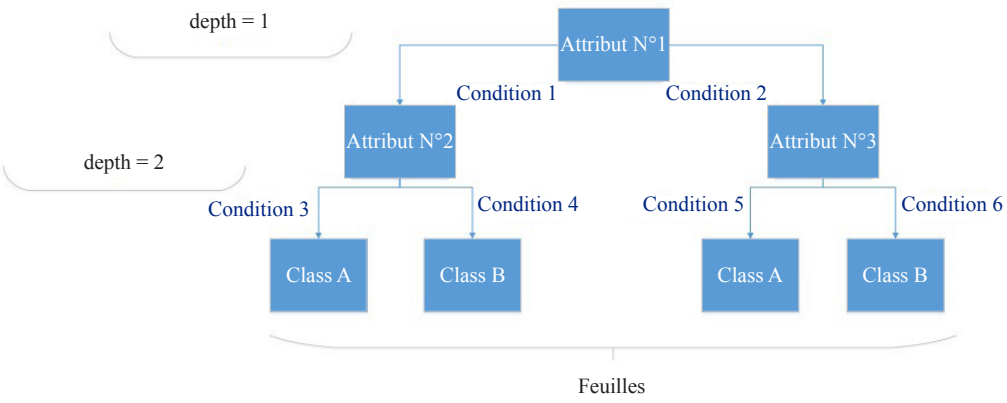

Figure 3. Diagram of the structure of a DT (depth = 2)

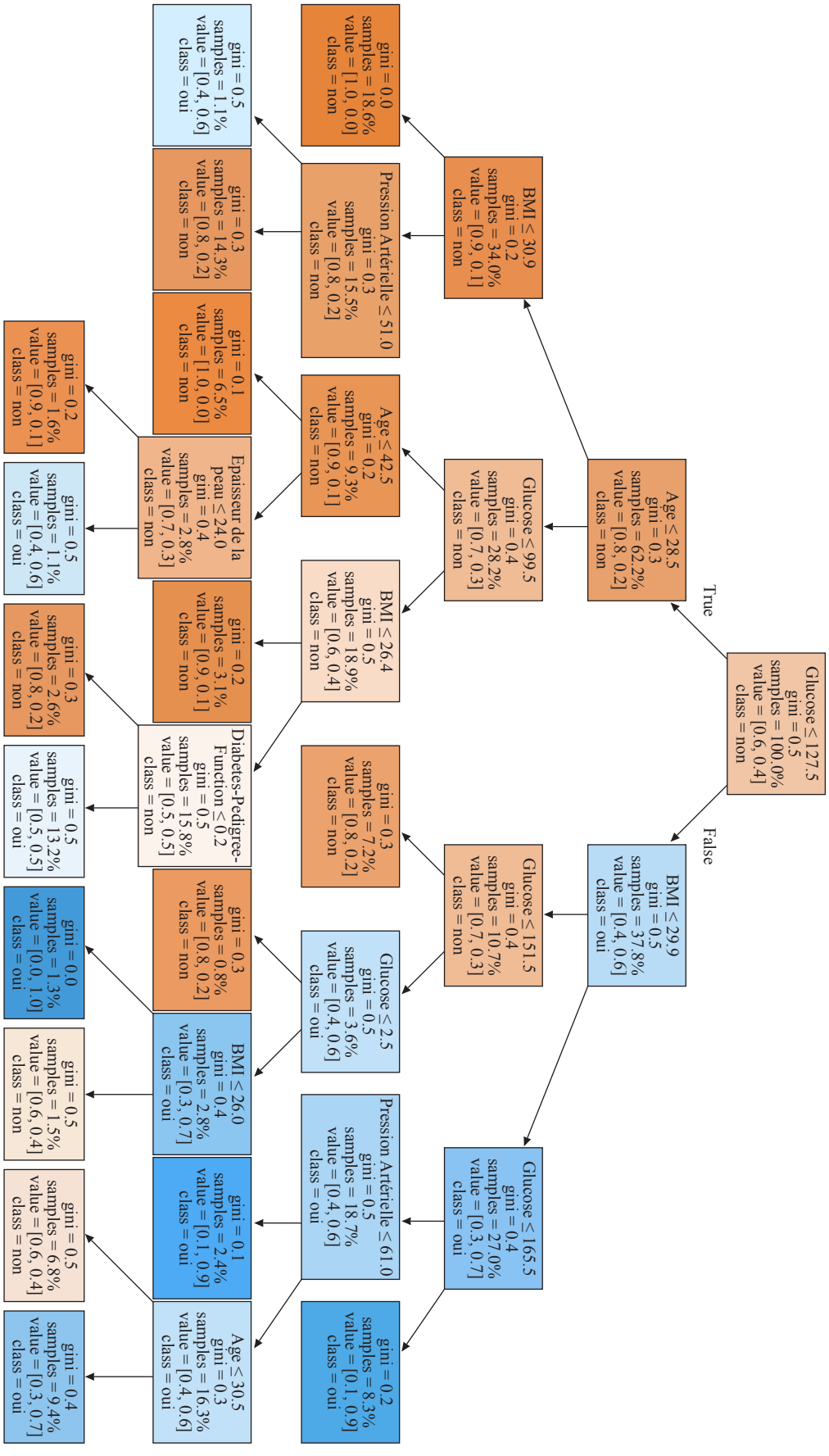

Figure 4. The DT obtained using Scikit Learn

To obtain the DT model (Figure 4), the Scikit Learn library gives us the possibility to make several adjustments that concern the characteristics of the tree, we quote a few:

1) Tree depth: this specifies how many levels our tree should contain. We set this parameter to 5.

2) Division criterion: this is the criterion that decides in which direction the tree should propagate, and the attributes in each node.

3) Clarification: this precision concerns the values of attributes, in our case we have fixed it to 1 (one digit after the decimal point).

The next step is to design a hardware accelerator for the resulting model.

4. Hardware methodology of implementation and its technical side

In the previous section, we got our DT model, corresponding to the diabetes prevention system. currently, we are faced with another task which is the hardware implementation of this model. The envisaged solution is the hardware implementation on an FPGA card.

In this section, we will first follow a methodology for the choice of the design process of our model, by which we will obtain the IP core that will be implemented on a SoC target, in order to test it using a software development environment.

4.1 *Behavioral description of DT in C*

First, we wrote the C source code corresponding to the DT model of diabetes prevention. The code consists of a function called "DT" with five "double" type variables, 2 integer variables, and the integer "diabetes" variable, which takes a "0" indicating that the diabetes test is negative or a "1" otherwise. The DT takes the form of a succession of If and Else statements leading to the result "1" or "0". The test starts from the root and continues throughout the tree, following precise directions in the branches of the tree according to the values of the attributes, until reaching the level of the leaves where the decision will be made.

A second C file is needed for testing in Vivado HLS which is the test bench also described in C. The test bench takes the form of the main C function which executes the "DT" function and self-checks the results.

4.2 *Performance obtained from synthesis and implementation with and without optimization*

The sync summary (Table 5) shows the estimated target clock frequency. If the estimated clock rate is higher than the target, the hardware will not run at that clock rate.

The Interface section displays the I/O ports and protocols created by the interface summary:

- The design requires a clock cycle, so a clock and a reset were added to the design: ap_clk and ap_rst. Both are single-bit inputs.

- A block-level I/O protocol has been added to control the RTL design: ports ap_start, ap_done, ap_idle, and ap_ready.

The design has four data ports:

- The blood glucose, BMI, EP, PA, and DPF input ports are 64-bit inputs whose I/O protocol is ap_none, and the age and pregnancy inputs are 32-bit, with the same protocol.

- The design also features a 32-bit diabetes output port.

4.3 *Physical implementation of the system on a SoC target*

The ZedBoard (Figure 5) is an evaluation board from Xilinx for using the Zynq-7000 chip (consisting of an ARM processor and an FPGA).

The Zedboard consists of two main parts [25]:

- A processing system (PS): based on an ARM Cortex A9 dual-core processor capable of supporting an operating system such as LINUX. We call PS the processor part and the associated peripherals that include the two ARM cortex

A9 cores, the AMBA and AXI buses, the DMA, the GPIOs, I2C, UART, CAN, SPI, the QuadSPI, NAND and NOR memory controller, and the RAM controller.

- A logic programming system (programmable logic-PL): This part contains the basic logic elements, RAM, DSPs, and standard inputs/outputs.

In terms of connectivity, the ZedBoard is quite complete and allows the interfacing of many elements:

- HDMI (Audio and video)
- VGA (Video)
• Ethernet (10/100/1000 Mbps)
- ARM Debug Access Port (DAP)
- USB 2.0 OTG port (Device, Host, and OTG)
- USB-JTAG programming port-USB-UART port
- FMC connector.

There are also quite a number of HMI components:

- 9 LEDs
- 8 interruptrices glissières
- 7 boutons poussoirs + 2 boutons Reset

Integrated memories, video, and audio inputs and outputs, dual role USB and Ethernet interface, as well as the SD slot, simplify the design phase as much as possible without additional hardware.

4.3.1 *Solution design flow*

The diagram below (Figure 6) encompasses the methodology followed for the design of the system on a chip (SoC) for the prevention of diabetes.

4.3.1.1 *The AXI interface*

For the implementation of our DT on a SoC target, we should establish a connection between the IP designed with HLS and the processing system of the ZedBoard card. To do this, Xilinx standardizes IP interconnections with the AXI4 protocol. To create this interface, we specified an AXILite-like interface for the seven parameters “blood sugar, age, BMI, EP, PA, DPF and pregnancies” in the behavioral description of DT in C.

The estimated performance with Vivado HLS after the addition of the AXI interface has changed. These have been summarized in Table 6.

Table 6. Hardware resources and estimated performance on Vivado HLS

Material resources and performance	Without directives (%)	Directive “Pipline”
BRAM 18K	0	0
DSP48E	76.00%	0.81
DT (Depth = 12, CART)	0	0
FF	1582	1582
LUT	6043	6032
Clock (ns)	8.371	7.884
Latency (clock cycle)	1	1

AXI is a point-to-point interconnect designed for high-performance and high-speed micro-controller systems and is part of AMBA ARM's specification. The AXI protocol is based on point-to-point interconnection to avoid bus sharing and therefore allow higher bandwidth and lower latency. AXI is arguably the most popular AMBA interface interconnect.

The essence of the AXI protocol is about providing a framework for the communication of the different blocks inside each chip. It proposes a procedure before any transmission so that the communication is clear and uninterrupted.

4.3.1.2 *Hardware design with Vivado IP Integrator*

A common use of the HLS design is to create an accelerator for a processor in order to move the code that runs on the CPU into the FPGA to improve performance.

a- Implementing the system on the Zynq SoC using Vivado

Figure 7 illustrates the hardware block design when using Vivado IP Integrator. Note that when creating the block in IP Integrator the Zynq PS and HLS, IPs are added manually, while the two additional IPs are automatically added to the design (Processor System Reset and AXI Interconnect) when executing the block and automating the connection. All the interconnections between the different blocks are also made.

- Processing System ZYNQ7

(Processing system7 0): The processing system is added manually, it is only for configuration and will not be implemented on the PL, it actually corresponds to the fixed PS on the Zynq chip, and everything related to the PS must be specified in this IP.

- DT 0: It was the IP that was designed, simulated, and exported in HLS format.

- AXI Interconnect (ps7 0 axi periph): It is used to interconnect a memory-mapped AXI master device, including the PS, with a memory-mapped slave device, which is the AXI-Lite compatible HLS core called "DT" in this solution.

- Processor System reset (rst ps7 0 100M): It allows the design to be adapted to its application by defining certain parameters to activate/deactivate the functions.

b- Synthesis, Soc implementation, and Bitstream generation

The binary stream generation creates the hardware image, which can be exported to a software environment in which we will create a software application using the custom hardware.

4.3.1.3 *Software design with SDK*

The Software Development Kit (SDK) provides an environment for building software platforms and applications for Xilinx integrated processors. The SDK works with hardware designs created with Vivado.

In this environment we have followed the following steps to verify the correct functioning of our IP generated by HLS:

1) Creation of a software application written in C code:

By using the functions available on the driver of our IP (xdt.c) provided by HLS, mainly the functions dedicated to the insertion of the three variables (blood sugar, age, BMI, EP, PA, DPF, and pregnancies), and the function for retrieving the variable (diabetes) indicating the result provided by our PI.

- Define a software model of the HLS hardware functionality (DT C code) with which we can compare benchmark results.

- Write the main function, which calls the hardware function (IP provided by HLS) and the software function (C code of DT), and which makes a comparison between the results of the two functions and sends a "Results match" response, if both functions give the same result and "Results Mismatch" otherwise.

2) We have made the necessary connection between the ZedBoard and the computer.

3) We used the BitStream generated by our SoC system from the IP Integrator tool as the Hardware platform.

4) Finally, we loaded the code on the ZedBoard card, and observe the result on the SDK terminal (Figure 8).

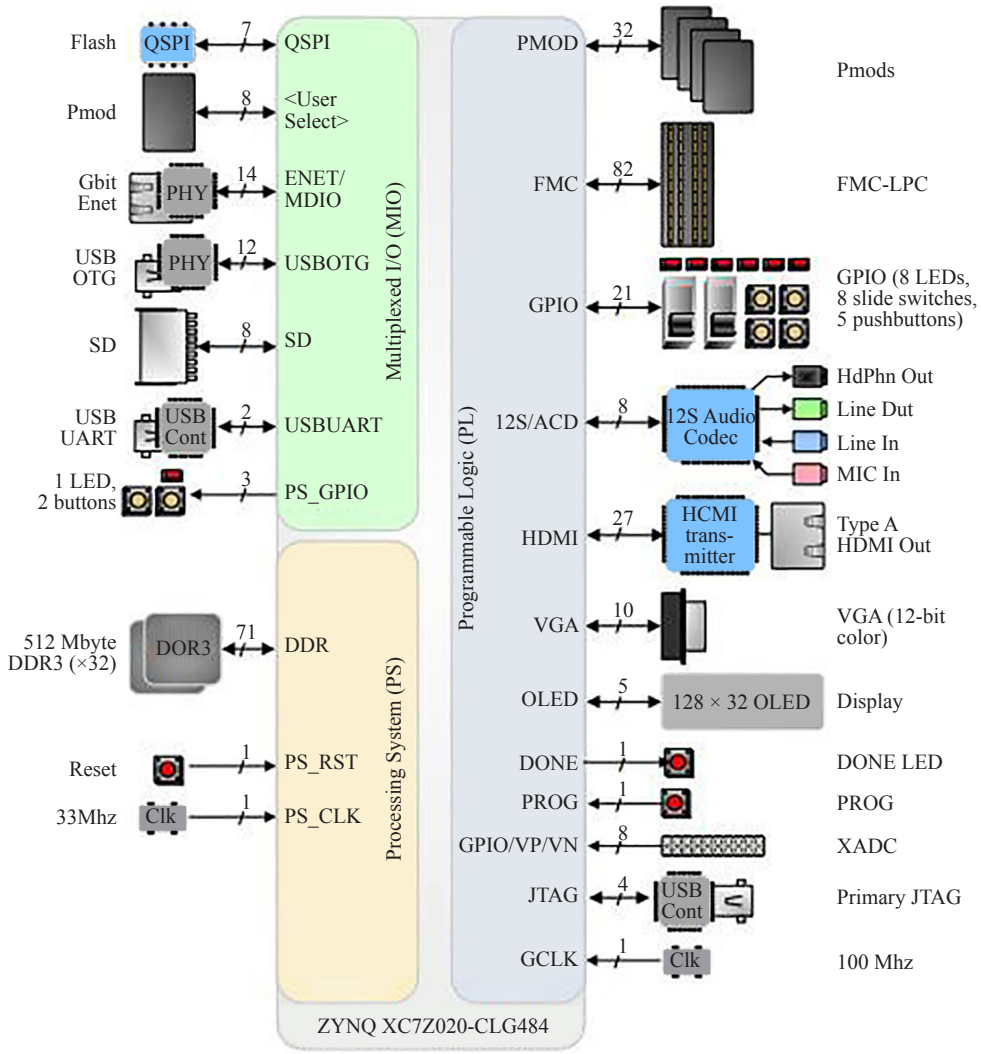

Figure 5. ZedBoard block diagram [24]

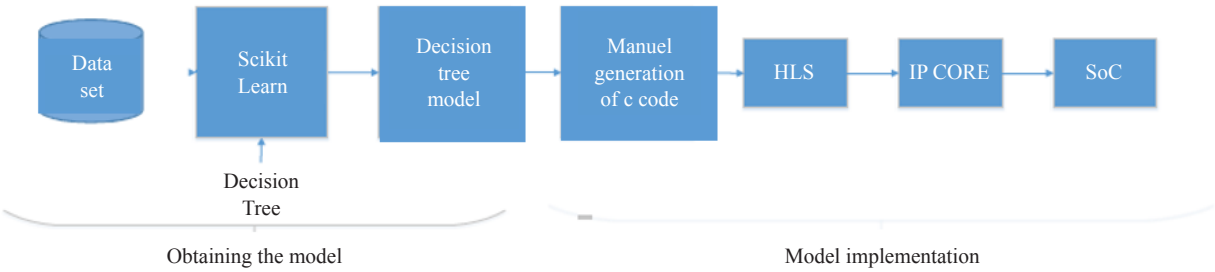

Figure 6. Explanatory diagram of the design flow of our solution

5. Conclusions and perspectives

In this present study, we looked at the implementation of a system on a chip dedicated to the prevention of diabetes.

We had thought of this implementation through a development methodology facilitating rapid prototyping. Thus, a comparative study of various platforms implementing ML algorithms has led us to determine both the best platform and the best algorithm to be applied for the prevention of diabetes.

As it had been demonstrated, the couple Scikit Learn and the DT was the relevant choice when it came to obtaining the ML model. The best precision (88.31%) was obtained for the DT algorithm (depth of 5) using CART as a division criterion, compared to other DTs with different depths and to other ML algorithms such as Bagged Tree, SVM, and KNN.

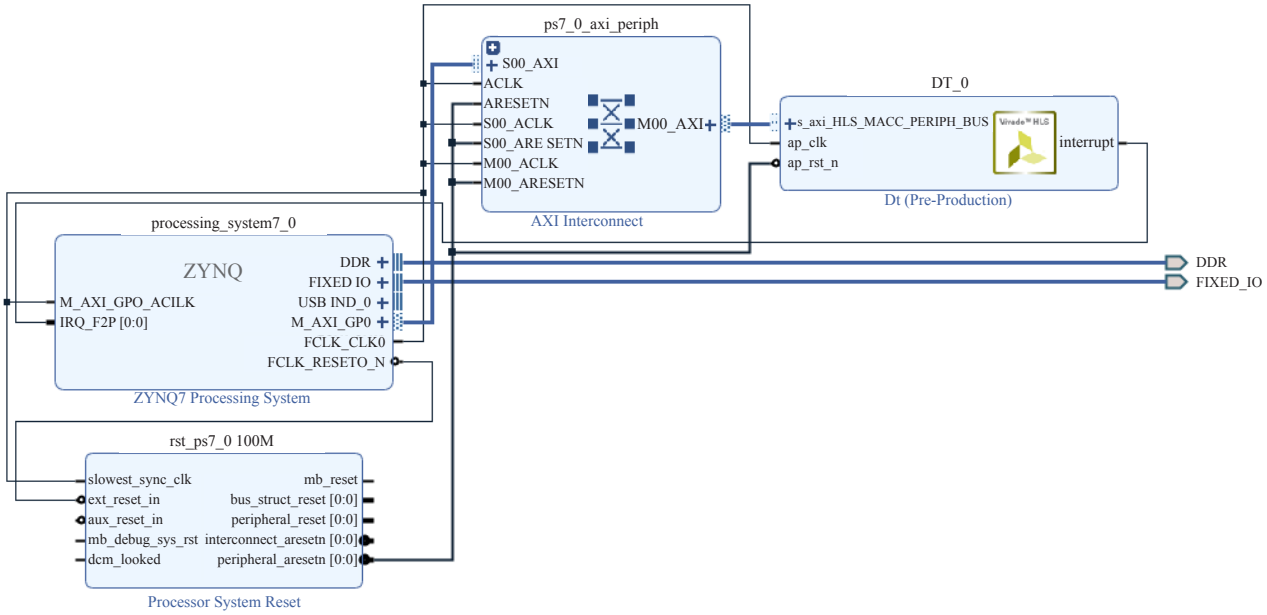

Figure 7. Block diagram of the implementation on a SoC target

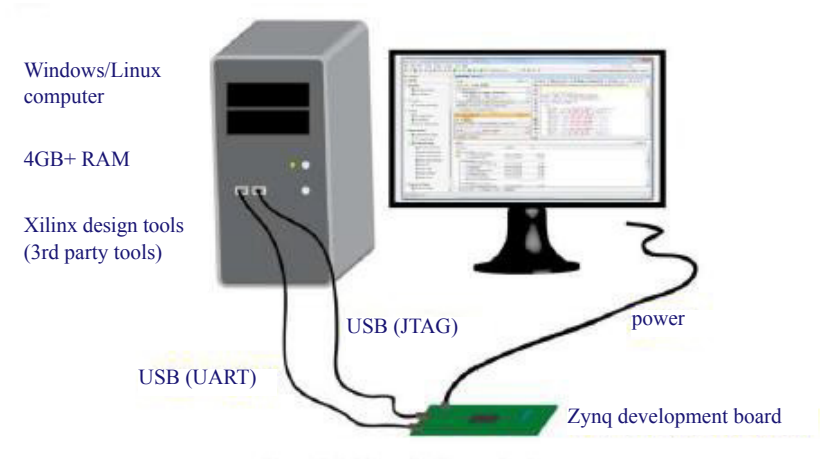

Figure 8. Test device

For the implementation of our model, we opted for the HLS approach. This choice was deduced from comparative work on implementation approaches; the criterion was first the prototyping time. Other studies have shown that Vivado HLS is the best high-level synthesis tool considering several criteria, such as the ease of implementation, the possibility

of model optimization, and the availability of documentary resources. The DT obtained was implemented in hardware using a ZedBoard-type FPGA platform. We verified the correct functioning of the IP obtained by comparing the results obtained with those obtained by a purely soft implementation. In our development approach, the transition from the learning model to its implementation is subject to manual translation. This limitation is quite natural because ML platforms are dedicated to "Data Scientists". We propose, as a perspective, the development of an automatic translator into description languages or (and) into imperative languages.

As today, many types of embedded systems can interface without reasonable effort with Cloud systems for data exchange, data storage, and processing, we propose as another perspective for our project, a connection of the SoC that we have developed with the Cloud. The use of Cloud-level accelerators is the potential that can exceed the technical limits of the Cloud, such as the transfer of functions from a real-time system to the Cloud which imposes a number of constraints in terms of synchronization, such as real-time communication, closed-loop runtime, and gigue, which might be difficult to achieve today.

Conflict of interest

The authors declare that they have no conflicts of interest.

References

[1] World Health Organization. Global leprosy update, 2013; reducing disease burden. *Releve Epidemiologique Hebdomadaire.* 2014; 89(36): 389-400.

[2] Ba-Alwi FM, Hintaya HM. Comparative study for analysis the prognostic in hepatitis data: Data mining approach. *International Journal of Scientific & Engineering Research*. 2013; 4(8): 680-685.

[3] Ibrahim F, Taib MN, Abas WABW, Guan CC, Sulaiman S. A novel dengue fever (DF) and dengue haemorrhagic fever (DHF) analysis using artificial neural network (ANN). *Computer Methods and Programs in Biomedicine*. 2005; 79(3): 273-281.

[4] Sharma P, Saxena K, Sharma R. Efficient heart disease prediction system. *Procedia Computer Science*. 2016; 85: 962-969.

[5] Deepika K, Seema S. Predictive analytics to prevent and control chronic diseases. In: *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*. IEEE; 2016. p. 381-386.

[6] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. *Procedia Computer Science*. 2018; 132: 1578-1585.

[7] Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: Literature review. *Journal of Medical Internet Research*. 2018; 20(5): e10775.

[8] Pomey MP, Flora L, Karazivan P, Dumez V, Lebel P, Vanier MC, et al. The Montreal model: Challenges of the relational partnership between patients and health professionals. *Public Health*. 2015; 1(HS): 41-50.

[9] Halimi S. Diabète de type 1 de l'enfant en France: des données plus précises, mais un risque d'acidocétose inaugurale inchangé malgré une campagne active [Type 1 diabetes in children: More precise French data but a risk of inaugural ketoacidosis unchanged despite active campaign]. *Metabolic Disease Medicine*. 2017; 11(8): 675-674. Available from: https://doi.org/10.1016/S1957-2557(17)30161-X.

[10] Wu Y, Ding Y, Tanaka Y, Zhang W. Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention. *International Journal of Medical Sciences*. 2014; 11(11): 1185.

[11] Willi C, Bodenmann P, Ghali WA, Faris PD, Cornuz J. Active smoking and the risk of type 2 diabetes: A systematic review and meta-analysis. *JAMA*. 2007; 298(22): 2654-2664. Available from: https://doi.org/10.1001/jama.298.22.2654.

[12] Song Y, Liang J, Lu J, Zhao X. An efficient instance selection algorithm for k nearest neighbor regression. *Neurocomputing*. 2017; 251: 26-34.

[13] Pradeep K, Naveen N. Predictive analysis of diabetes using J48 algorithm of classification techniques. In: *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*. IEEE; 2016. p. 347-52.

[14] Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes

or prediabetes by risk factors. *The Kaohsiung Journal of Medical Sciences*. 2013; 29(2): 93-99.
[15] Roy A, Qureshi S, Pande K, Nair D, Gairola K, Jain P, et al. Performance comparison of machine learning platforms. *INFORMS Journal on Computing*. 2019; 31(2): 207-25.
[16] Saravananathan K, Velmurugan T. Analyzing diabetic data using classification algorithms in data mining. *Indian Journal of Science and Technology*. 2016; 9(43): 1-6.
[17] Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*. 2020; 18(1/2): 90-100. Available from: https://doi.org/10.1016/j.aci.2018.12.004.
[18] Hossain ME, Uddin S, Khan A. Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes. *Expert Systems with Applications*. 2021; 164: 113918.
[19] Zhang Z, Han H, Cui X, Fan Y. Novel application of multi-model ensemble learning for fault diagnosis in refrigeration systems. *Applied Thermal Engineering*. 2020; 164: 114516.
[20] Aher SB, Lobo L. Data mining in educational system using weka. In: *International Conference on Emerging Technology Trends (ICETT)*. 2011; 3: 20-25.
[21] Hao J, Ho TK. Machine learning made easy: A review of Scikit-learn package in python programming language. *Journal of Educational and Behavioral Statistics*. 2019; 44(3): 348-361.
[22] Maxwell AE, Warner TA, Fang F. Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*. 2018; 39(9): 2784-2817.
[23] Singh S, Gupta P. Comparative study ID3, cart and C4. 5 decision tree algorithm: a survey. *International Journal of Advanced Information Science and Technology (IJAIST)*. 2014; 27(27): 97-103.
[24] Rutkowski L, Jaworski M, Pietruczuk L, Duda P. *The CART decision tree for mining data streams. Information Sciences*. 2014; 266: 1-15.
[25] Yeniçeri R, Hüner Y. HW/SW codesign and implementation of an IMU navigation filter on Zynq SoC with Linux. In: *2020 7th International Conference on Electrical and Electronics Engineering (ICEEE)*. IEEE; 2020. p. 351-354.