Model for blood cholesterol concentration prediction using Mesoscopic simulation techniques with Dissipative Particle Dynamics (DPD) and Linear Regression

Reyna Nohemy Soriano-Machorro
División de Estudios de Posgrado e Investigación
Tecnológico Nacional de México/I.T. Orizaba
Orizaba, Ver. México
reynasorimach@gmail.com

Luis Rolando Guarneros-Nolasco
División de Estudios de Posgrado e Investigación
Tecnológico Nacional de México/I.T. Orizaba
Orizaba, Ver. México
ORCID: 0000-0001-6379-4969

María Antonieta Abud Figueroa
División de Estudios de Posgrado e Investigación
Tecnológico Nacional de México/I.T.
Orizaba, Ver. México
ORCID: 0000-0001-9166-3413

Beatriz Alejandra Olivares Zepahua
División de Estudios de Posgrado e Investigación
Tecnológico Nacional de México/I.T. Orizaba
Orizaba, Ver. México
ORCID: 0000-0003-2799-0887

José Luis Sánchez-Cervantes
División de Estudios de Posgrado e Investigación
CONACYT-Instituto Tecnológico de Orizaba
Orizaba, Ver. México
ORCID: 0000-0001-5194-1263

Nancy Aracely Cruz-Ramos
División de Estudios de Posgrado e Investigación
Tecnológico Nacional de México/I.T. Orizaba
Orizaba, Ver. México
nancy.cramos5@gmail.com

***Abstract*—Four out of ten Mexican adults have high cholesterol levels. According to the National Institute of Cardiology. There are different types of cholesterol, which depend on what the lipoprotein transports. They are: low-density lipoprotein (LDL), or "bad" cholesterol, which transports cholesterol particles throughout the body, and HDL cholesterol (high-density lipoprotein), which transports cholesterol particles throughout the body. Cholesterol particles throughout the body, and HDL (high-density lipoprotein), or "good" cholesterol, which carries (High Density Lipoproteins), or "good" cholesterol, collects the excess cholesterol and carries it back to the liver. LDL cholesterol causes hardening and narrowing in the walls of the arteries and can form a clot that causes a heart attack or stroke. Considering this problem, our paper proposes the use of mesoscopic simulation techniques with a Dissipative Particle Dynamics method a model for the prediction of cholesterol concentrated in the blood, to raise public awareness about the care and prevention of cardiovascular diseases, also the use of parallel computing with the CUDA language (Compute Unified Device Architecture).**

Keywords—Web applications, Mesoscopic simulation, Cholesterol, Linear Regression

I. INTRODUCTION

Cholesterol is a waxy, fat-like substance found in every cell of the human body. Cholesterol is important for good health and is needed to make cell walls, tissues, hormones, vitamin D and bile acid, it also comes from the consumption of foods of animal origin. There are two types of cholesterol in the human body, HDL (high density lipoprotein composed of lipids and proteins), commonly known as good cholesterol and LDL (low density lipoprotein) known as bad cholesterol, which when there is too much cholesterol in the blood, can accumulate in the walls of blood vessels, blocking blood flow to tissues and organs, and increasing the risk of heart disease and stroke, and according to the National Institute of Cardiology, four out of ten Mexican adults have high cholesterol.

In the medical domain, the mesoscopic simulation can support to specialists to maintain a communicative, intelligent and well-targeted interrelationship to meet specific objectives by eliminating irrelevant information for the treatment of diseases using 3D models. Due to, that mesoscopic simulation, is located between the atomistic and macroscopic scales, is the macroscopic scale, is Dissipative Particle Dynamics (DPD), which is able to capture processes that take place at the microns. Processes that take place in microseconds, thus allowing the analysis of molecule currents and their motions based on the and their motions based on predefined velocity-density functions and capacities. and density functions. Individual molecules can be tracked to identify agglomerations or clusters. Clusters or clusters of molecules that cause an effect depending on the application domain. Depending on the application domain. For this reason in this paper, it is proposed the development of a model for the prediction of cholesterol concentration in the blood, to raise awareness among the population about the care and prevention of cardiovascular diseases with an approach with Big Data and mesoscopic simulation techniques with a DPD Particle Dynamics method. The paper is organized in follow sections: Section II presents work related to fluid simulation and data mining for cardiovascular disease prevention, the III section includes the architecture design of the proposed solution, section IV describes the Case of Study and, finally section V presents the conclusion and future work.

II. RELATED WORK

A. Fluid simulation models

High density lipoproteins (HDL, high density lipoproteins commonly called good cholesterol) were studied in [1]. The authors analyzed 47 subjects with different HDL levels and studied their analyses with the R language for statistical computing, using the GROMACS package (v. 4.0) to perform the simulations. The result obtained was the molecular profile of HDL particles which was combined with dynamic structural modeling. The interfacial and structural properties

of fluids confined by different walls at mesoscopic scale using DPD simulations in the Grand Canonical ensemble were studied in [2]. The whole methodology was implemented in simulation code called SIMES, the authors used a graphical processing unit (GPU), for confined fluids they used DPD and they made a hybrid with the Grand Canonical Monte Carlo (MC) methodology to simulate wall-confined fluids. A red blood cell (RBC) flow simulation package with GPU-accelerated chemical transport property based on an adaptation of the transport dissipative particle dynamics (tDPD) method was presented in [3]. The C/C++, CUDA (Compute Unified Device Architecture) C/C++ and MPI (Message Passing Interface) languages were used for programming. The simulation package processes all computational workloads in parallel per GPU and incorporates multi-stream scheduling and non-blocking MPI communications to improve scalability between nodes.

A method for the simulation of red blood cell flows in different tubes was developed by Ye et [4], the methods used were smoothed dissipative particle dynamics (SDPD) and dissipative particle dynamics (DPD), with parameters having specific physical significance, combined with thermal fluctuations in a mesoscale simulation and the immersed boundary method (IBM). The simulations were demonstrated with CFD (Computational Fluid Dynamics) by simulating red blood cell flows in rectangular, cylindrical, curved, bifurcated and constricted tubes. Likewise, a GPU-accelerated package for flow simulation in nano- to micropore networks with a mesoscale model of many-body dissipative particle dynamics (mDPD) was presented by Xia et al. [5]. The authors used the many-body dissipative particle dynamics (mDPD) method, they also used the Velocity-Verlet algorithm, the programming languages used were CUDA C / C++ with MPI and OpenMP.

B. Data mining application for cardiovascular disease prevention

Work using data mining techniques for disease prevention continues.

The association between cholesterol efflux capacity and mortality risk in patients with cardiovascular disease (coronary artery disease) was studied in [6]. The authors conclude by suggesting that cholesterol outflow capacity is a predictor of mortality from all cardiovascular problems in both acute coronary syndrome (ACS) and diabetic ketoacidosis (DKA). Similarly a meta-analysis of data from patients with no history of cardiovascular disease, using two methods of increasing complexity to model repeated measurements was performed by Paige et al. [7]. The authors conclude that if a mean of a post-baseline repeat measurement of systolic blood pressure is incorporated, it is possible that total cholesterol and HDL cholesterol in cardiovascular disease risk prediction models will result in slight improvements in risk discrimination and reclassification. Moreover, an advanced data mining technique to develop an effective cardiac disease prediction system using a neural network was presented by Singh et al. [8]. An efficient system for predicting heart disease was presented ed by Prushottam et al. [9], they used databases with clinical information of cardiac patients and applied data mining to process the obtained data set with input values. When testing the system, it obtained 86.3% for prediction in the test phase, and in the training phase obtained a prediction of 87.3%, this tool predicts early cardiac diseases.

C. Table 1 shows a comparison of the related papers in this section.

TABLE I. TABLE 1. COMPARATIVE TABLE BETWEEN THE PROPOSED WORK AND RELATED WORK.

Article	**Simulation initiatives**			
	Model	***Simulation***	***DPD***	***GPU***
Yetukuri et al. [1]	X	X		
Terrón-Mejía et al. [2]	X	X	X	X
Blumers et al. [3]	X	X	X	X
Ye et al. [4]	X	X	X	
Xia et al.[5]	X	X	X	X
	Prediction initiatives			
	Data Mining		***Linear Regression***	
Liu et al. [6]	X			
Paige et al. [7]	X		X	
Singh et al.[8]	X			
Purushottam et al.[9]	X			
DPD = Dissipative Particle Dynamics or derivatives ***GPU= Graphical Processing Unit*** ***App CD= Cholesterol detection mobile-app***				

Fig. 1. Comparative table between the proposed work and related work

Therefore, after performing an analysis of the articles discussed in this chapter, it was found that works [2], [3] and [5] use GPU architecture, which has advantages such as reducing data processing time, since it is 15 times faster than using a CPU architecture, so for this work GPU accelerated simulation was chosen, where it allows parallel processing between one or more GPUs and CPUs. In [2] the DPD method was used, also variant methods of DPD were used in [3], [4], and [5] for the simulation of simple and complex fluids, these approaches allow justifying the use of DPD in this work where different levels of cholesterol in blood will be simulated, because it is the method that best describes complex fluids, since it has a better control of the transport properties. In [1] cholesterol was simulated with different methods and simulation software, approaches about simulation of red blood cells were found as in [3]and [4], unlike the previous works, different levels of cholesterol in blood were simulated.

Works [7]-[9] used data mining techniques to detect and predict cardiovascular disease, also in [6]-[9]cholesterol was used as a predictor of risk for heart ailments.

In contrast, this work will only predict cholesterol with user-supplied data. In this work, data mining techniques will also be applied to predict cholesterol levels and heart disease risk.

III. Architectural Design

This section presents a description of the architecture of the proposed solution for this work. The structure of the architecture is shown in Figure 2.

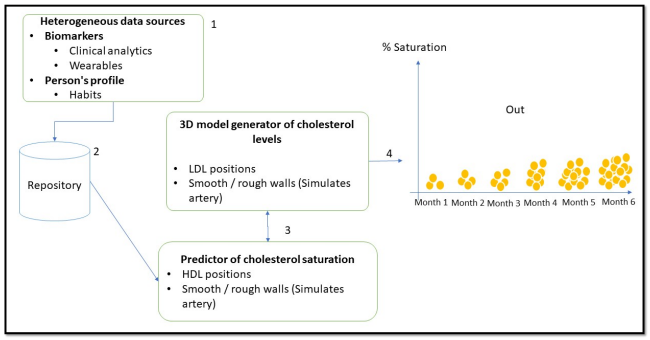

Fig. 2. Architectural Design

A. *Description of the architecture*

- **Heterogeneous Data Sources Module:** This module is responsible for reading the biomarkers and habits of the person's profile.
- **Repository**: The repository is responsible for storing the data from the heterogeneous data sources module in a database management system.
- **Cholesterol saturation predictor module:** This module receives data from each user's profile store in the repository to predict the LDL positions that the user currently has and will have in time lapses, also the module will predict the thickness of the walls in different time lapses.
- **3D model generator of cholesterol levels:** In this module, 3D models of the output of LDL molecules at different time lapses are generated with the data from the cholesterol saturation predictor module.
- **Output Module:** This module shows the user the 3D models of cholesterol saturation in the different time lapses simulated in the 3D model generator module of cholesterol levels.
- **DPD Methodology:** For the simulation of cholesterol concentration in blood, the DPD dissipative particle methodology will be used, since DPD allows capturing processes that take place in the microseconds, thus allowing the analysis of molecule currents and their movements based on predefined capacities and velocity-density functions.

For the creation of the system, CUDA [10] will be used as a language for parallel computing with anNvidia graphics card, in addition to the Python language for data mining [11], this alternative also suggests the NoSQL database MongoDB [12], since among the supported languages is Python[11], in addition to its structure is oriented to objects and documents, likewise the simulation tool Gromacs will be used [13], since it has compatibility with CUDA and supports modern molecular dynamics.

IV. Concentration Prediction Cholesterol Model

For this work we used the methodology for the evaluation of the cardiovascular disease dataset proposed by Guarneros-Nolasco et al. [14] to create the model, the data loading process was carried out, to later choose the variables, which were age versus cholesterol obtained from the Cleveland Dataset. The following Listing shows a listing of the code for the creation of the model and its training.

```
#Creating model
set.seed(1234)
modeloLR <- lm(Colesterol ~ Edad, data = nltrain)
summary(modeloLR)
#Prediction for dataset training
y_predict <- predict(modeloLR, nltrain)
ggplot() + geom_point(data = nltrain, aes(x = Edad, y =
        Colesterol), size = 0.9) +
 geom_line(aes( x = nltrain$Edad, y = y_predict), 9
        color = "red") +
xlab("Edad") +
 ylab("Colesterol") +
 ggtitle("Curva de Ajuste sobre Conjunto de
Entrenamiento (nltrain)")
##Predicting cholesterol between two ages
rango.edades <- data.frame(Edad = seq(20, 60))
predict_value <- predict(modeloLR, rango.edades)
glimpse(predict_value)
df_predic=as.data.frame(predict_value)
write.csv(df_predic,'prediccion20_60.csv')
```

List. 1. Listing of code from model

We continued with the selection of the population and took as a sample a group of data of 260 people to run the regression and logistic classifier. Figure 3 shows the graph of cholesterol by age.

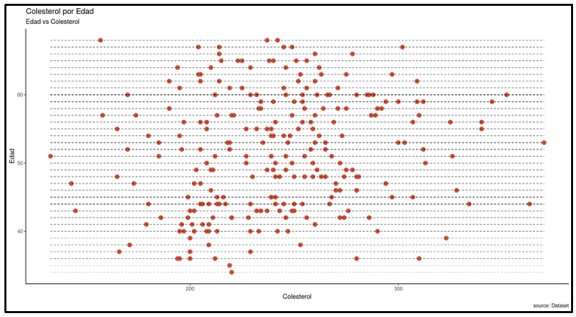

Fig 3. Graph of colesterol by age

The graph in Figure 4 shows the prediction obtained by the model in an age range from 20 to 60 years, in which the mg in one deciliter of cholesterol in the blood according to age is represented.

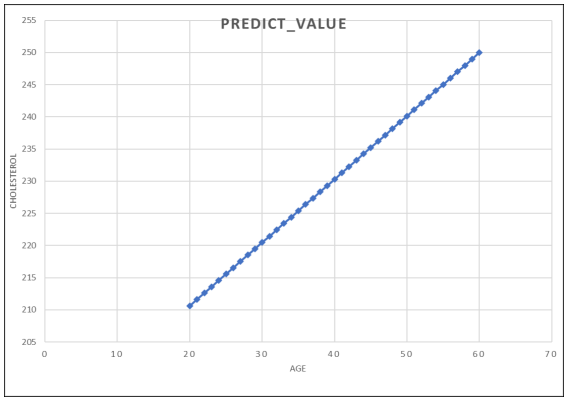

Fig 4. Prediction result in a range of 20 to 60 years of age.

Below are the figures obtained by the simulation module with the data provided by the model at ages 20 with 210.64 mg/dl (Figure 5), 40 with 230.30 mg/dl (Figure 6) and 60 age with 250.95 mg/d (Figure 7), in the figures the yellow color

represents cholesterol in mg/dl and the red color the rest of the molecules in the blood.

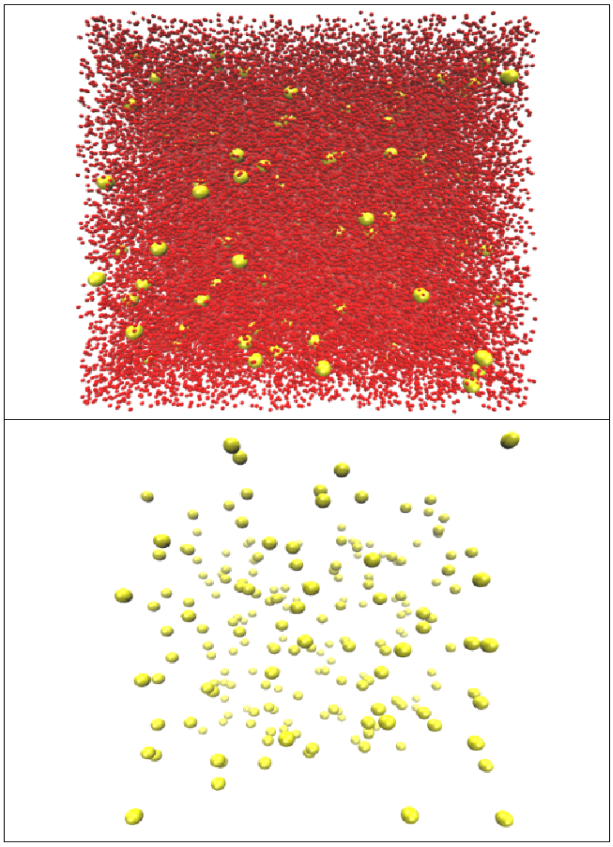

Fig. 5. Simulated blood 20 year olds with cholesterol of 210.64 mg/dl.

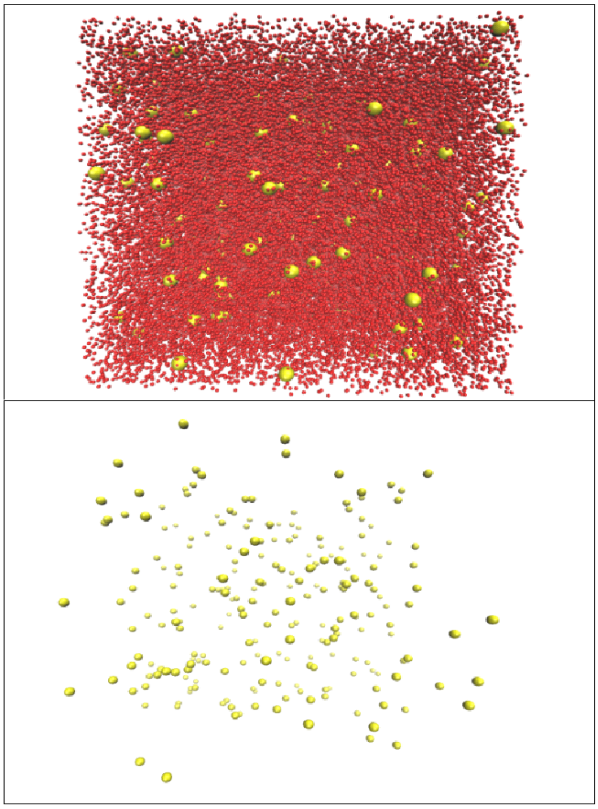

Fig. 6. Simulated blood 40 year olds with cholesterol of 230.30 mg/dl.

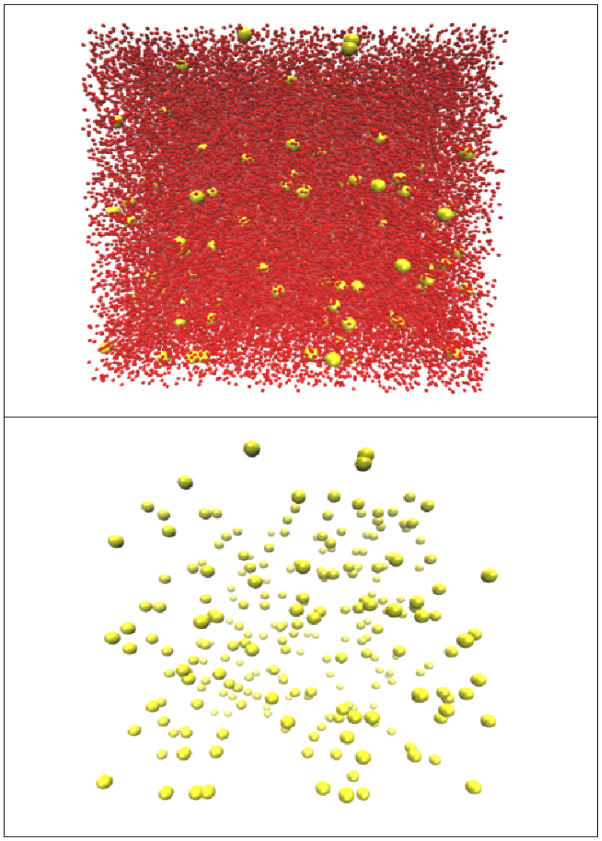

Fig. 7. Simulated blood 60 year olds with cholesterol of 250.49 mg/dl.

The simulation module calculates random positions within a cube representing 2 types of molecules, resulting in the number of cholesterol particles contained in a deciliter according to age. A Listing of the code of the simulation module is shown in Listing 2 (A, B) below.

```
print("\n  Creando configuracion inicial \n\n")
   Natoms1 = Natoms.natoms
   Natomsf = Natoms1
   num = Natoms1/Rho
   Lx= num**(1/3)
   Ly = Lx
   Lz = Lx
  print("  Longitud de la caja de simulacion
        \n\n", Lx)
   Rx[1]= (frand())*Lx
    Ry[1]= (frand())*Ly
    Rz[1]= (frand())*Lz
    i=2
    while i <= Natoms1:
      rxp = (frand())*Lx
      ryp = (frand())*Ly
      rzp = (frand())*Lz
      #print(rxp,ryp,rzp)
      addparticle = 0
      j=1
      i2=i-1
      while j <= i2 :
        rxj = Rx[j]
        ryj = Ry[j]
        rzj=Rz[j]
        #print(rxj,ryj,rzj)
        rxij = rxj - rxp
        ryij = ryj - ryp
        rzij = rzj - rzp
        n3=((rxij*rxij) + (ryij*ryij) +
       (rzij*rzij))
```

List. 2. (A) Listing fragment of the simulation module.

```
30        rij  = math.sqrt(n3)
31        if  rij <= 0.25:
32          addparticle = 1
33          j=i
34        else:
35          j+=1
36      if addparticle == 0 :
37        Rx[i] = rxp
38        Ry[i] = ryp
39        Rz[i] = rzp
40        i+= 1
41        if i%100000 == 0 :
42          print(" No. part %d \n",i)
```

List. 2. (B) Listing fragment of the simulation module.

V. Conclusions and Future Work

Our initiative seeks to help raise public awareness about the care of cholesterol levels in the blood, since a model of the representation in the third dimension supported by prediction models, helps the perspective of the health of people in the future, as it is visually attractive to show cholesterol levels over time facilitating the possibility of generating predictions of possible cardiovascular complications or the early detection of them, in order to people take the careful in they healthcare.

As future work, the model will include help to predict the 2 types of cholesterol in the blood (HDL and LDL) in addition to the total cholesterol that a user could have in a few years, as well as to present a simulation with DPD of the concentration of cholesterol. This informations, will present to the users as a friendly interface from a progressive Web application.

VI. Acknowledgments

This work was sponsored by the Council for Scientific Research and Technological Development in Veracruz (COVEICYDET). We thank COVEICYDET for supporting this work through project name Prevention and Early Detection of Cardiovascular Diseases (Arrhythmias & Tachycardias) using Machine Learning Techniques, Big Data, and the Internet of Things (ID 12 1806). The authors are very grateful to National Technological of Mexico for supporting this work. Also, this research paper was sponsored by the National Council of Science and Technology (CONACYT), as well as by the Public Education Secretary (SEP) through PRODEP.

[1] L. Yetukuri *et al.*, “Composition and lipid spatial distribution of HDL particles in subjects with low and high HDL-cholesterol,” *J. Lipid Res.*, vol. 51, no. 8, pp. 2341–2351, 2010.

[2] K. A. Terrón-Mejíia, I. Higuera-Ciapara, E. Mart\'\inez-Benavidez, J. Hernández, and R. López-Rendón, “Dissipative Particle Dynamics Simulations of Self-Assemblies of Liposomes for Drug Delivery Applications,” in *Liposomes-Advances and Perspectives*, IntechOpen, 2019.

[3] A. L. Blumers, Y.-H. Tang, Z. Li, X. Li, and G. E. Karniadakis, “GPU-accelerated red blood cells simulations with transport dissipative particle dynamics,” *Comput. Phys. Commun.*, vol. 217, pp. 171–179, 2017, doi: https://doi.org/10.1016/j.cpc.2017.03.016.

[4] T. Ye, N. Phan-Thien, C. T. Lim, L. Peng, and H. Shi, “Hybrid smoothed dissipative particle dynamics and immersed boundary method for simulation of red blood cells in flows,” *Phys. Rev. E*, vol. 95, no. 6, p. 63314, Jun. 2017, doi: 10.1103/PhysRevE.95.063314.

[5] Y. Xia *et al.*, “A GPU-accelerated package for simulation of flow in nanoporous source rocks with many-body dissipative particle dynamics,” *Comput. Phys. Commun.*, vol. 247, p. 106874, 2020, doi: https://doi.org/10.1016/j.cpc.2019.106874.

[6] C. Liu *et al.*, “Cholesterol efflux capacity is an independent predictor of all-cause and cardiovascular mortality in patients with coronary artery disease: A prospective cohort study,” *Atherosclerosis*, vol. 249, pp. 116–124, 2016, doi: https://doi.org/10.1016/j.atherosclerosis.2015.10.111.

[7] E. Paige *et al.*, “Use of Repeated Blood Pressure and Cholesterol Measurements to Improve Cardiovascular Disease Risk Prediction: An Individual-Participant-Data Meta-Analysis,” *Am. J. Epidemiol.*, vol. 186, no. 8, pp. 899–907, 2017, doi: 10.1093/aje/kwx149.

[8] P. Singh, S. Singh, and G. S. Pandi-Jain, “Effective heart disease prediction system using data mining techniques,” *Int. J. Nanomedicine*, vol. Volume 13, no. T-NANO 2014 Abstracts, pp. 121–124, Mar. 2018, doi: 10.2147/IJN.S124998.

[9] Purushottam, K. Saxena, and R. Sharma, “Efficient Heart Disease Prediction System,” *Procedia Comput. Sci.*, vol. 85, pp. 962–969, 2016, doi: https://doi.org/10.1016/j.procs.2016.05.288.

[10] N. Corporation, “CUDA,” 2021. https://developer.nvidia.com/cuda-zone.

[11] P. S. Foundation, “Python,” 2021. https://www.python.org/.

[12] I. © 2021 MongoDB, “La base de datos líder para aplicaciones modernas.” https://www.mongodb.com/es.

[13] GROMACS, “No Title,” *About GROMACS*, 2018. http://www.gromacs.org/About_Gromacs.

[14] L. R. Guarneros-Nolasco, N. A. Cruz-Ramos, G. Alor-Hernández, L. Rodríguez-Mazahua, en J. L. Sánchez-Cervantes, “Identifying the Main Risk Factors for Cardiovascular Diseases Prediction Using Machine Learning Algorithms”, Mathematics, vol 9, no 20, 2021.