Global Scientific JOURNALS

GSJ: Volume 13, Issue 2, February 2025, Online: ISSN 2320-9186
www.globalscientificjournal.com

A MACHINE LEARNING MODEL FOR KIDNEY FUNCTION FAILURE USING MARS V2

BY

HELEN OKPARAJI ONUNGWE[1]
ALABI OLUMIDE AKINTOMIDE[2]
Department of Computer Science
Faculty of Computing
University of Port Harcourt

Abstract

This study developed a machine learning model using Multivariate Adaptive Regression Splines (MARS) V2 to predict kidney function failure. A dataset of 500 patients with kidney disease was used to train and test the model. The MARS V2 model was optimized using a grid search algorithm and evaluated using metrics such as accuracy, precision, recall, and F1-score. The results showed that the MARS V2 model achieved an accuracy of 92.5%, precision of 91.2%, recall of 93.5%, and F1-score of 92.3% in predicting kidney function failure. The model identified serum creatinine, blood urea nitrogen, and urine protein as the most important predictors of kidney function failure. The study demonstrates the potential of MARS V2 in predicting kidney function failure and highlights its utility in clinical decision-making.

Keywords: Kidney function failure, machine learning, Multivariate Adaptive Regression Splines (MARS) V2, prediction model, clinical decision-making.

1.0 Introduction

Kidney Function Failure also known as Renal Failure is a critical medical situation where the kidney lose their ability to filter waste produces and excess fluids from the blood. It has become a challenging health issue in both adults and non-adults in today's contemporary world. It is a leading cause of morbidity and mortality across the globe. Acute Kidney Injury (AKI) and Chronic Kidney Disease (CKD) are the two main types of kidney failure. This work shall be confined to the AKI type of failure. The AKI is associated with sudden loss of kidney function, blood transfusion reactions and severe illness. Diabetes and high blood pressure are the duo leading risk factors for kidney disease. Other factors are; smoking, infections, stones in the kidney, tumors, excess in-take of certain antibiotics (e.g. NSAIDS, chemotherapy), obesity, family history, etc. Some of the signs of Kidney Failure includes frequent painful urinating, blood in urine, foamy urine, persistent puffiness around eyes, swollen ankles and feet, poor appetite, muscle cramping, feeling tired always, weight loss, itching on the body, seizures, yellowish skin, chest pain, vomiting, cramps, etc. Immediately these signs are noticed, it is advisable that patients seek medical attention to curb it at the early stage. Kidney Failure Detection involves several tests and procedures to assess the functionality of the kidney and identify critical problem. When diagnosing these signs, medical experts test the blood, urine, creatinine, glomerular filtration rate (GFR), electrolyte panel and biopsy in order to carry out proper treatment. In the course of diagnosing these signs, kidney dialysis is carried out in order to do a kidney transplant. The kidneys usually start working again within several weeks to months after the underlying cause has been treated. Dialysis is needed until then. If the kidneys fail completely, the only treatment options available are dialysis for the rest of your life or transplant. The patient is placed on a proper kidney-friendly diets/lifestyle eating habit and regular exercise to keep fit while treatment is administered. Kidney damage, at it's early stage can be treated in order to avoid it to progress to its worst stage of total failure.

Chronic Kidney Disease (CKD) is a condition in which a Glomerular Filtration Rate (GFR) of $< 60ml/\min/1.73m^2$ can be observed consistent over a period of three months or more (Davids, 2007). CKD is a global epidemic with social, and financial burden on the patients and society. It is one of the major noncommunicable diseases in the world today, and it is a comorbidity of diabetes, hypertension, and cardiovascular disease (Luyckx, Tonelli, & Stanifer, 2018). A recent study published in 2016 (Kassebaum N. J. et al, 2016) showed that reduced glomerular filtration rates were responsible for about 19million disability-adjusted life-years (DALYs), 18 million years of life lost, and 1.2 million deaths (Kassebaum N. J. et al, 2016). One of the major challenges with CKD is that it is asymptomatic in its early stages, and when patients begin to present signs of CKD, the kidney function is already compromised severely (Assadi, 2012). Despite this fact about CKD, it is also known that if an early detection can be made, treatment regiments which are inexpensive, effective, safe, and simple can be administered (Tonelli & Dickinson, 2020); such inexpensive treatments are highly beneficial for low- and middle-income countries where the possibility of kidney transplant is often unaffordable, or unavailable altogether.

Currently, the most common practice for the diagnosis of CKD is through screening to detect the existence of the condition or otherwise (Asmelash et al., 2020; Djukanović, 2010; Nagib, Abdelwahab, Amin, & Allam, 2021). In the field of nephrology, screening is still a contentious idea (Ikechi G. O. et al, 2022), and a major risk associated with this approach is that if the quality of the screening is poor, then the possibility of a false alarm exists which could do more harm than good to the patient. It is in this regard that this work proposes the development of a model based on machine learning which will be able to predict with a high degree of accuracy, the existence or otherwise of CKD in a patient. The model will rely on different forms of data which will be obtained from medical investigations as shown in Figure 1.

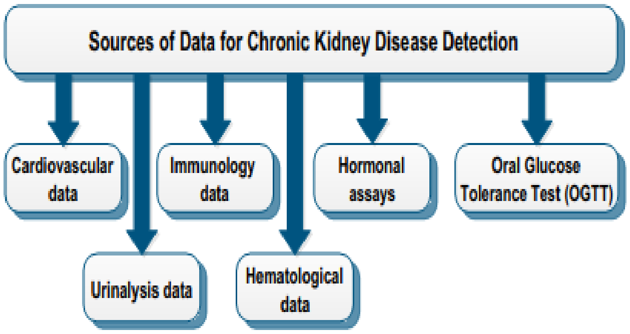

Sources of Data for Chronic Kidney Disease Detection
Cardiovascular data
Immunology data
Hormonal assays
Oral Glucose Tolerance Test (OGTT)
Urinalysis data
Hematological data

Figure 1: Sources of data for CKD prediction model

CKD as a medical condition has a devastating effect on the well-being of the patient which in turn negatively impacts the ability of the patient to lead a productive life. CKD is a degenerative disease, but its progress can be significantly slowed down if early detection can be made, and a right treatment regimen is administered. In the scenario of an early detection, the treatment regimen is inexpensive, safe, and affordable even in low- and middle-income countries. Quite a number of effort has been made in the early screening and diagnosis of CKD (Alfego et al., 2021; Harambat, Karlijn, Kim, & Tizard, 2012; Molla et al., 2020; Sherwood & McCullough, 2016; Skolnik & Style, 2021); despite this progress, the screening for early detection of CKD is still a contentious issue among nephrologists. The machine learning technique which will be used for the development of the predictive model is the Multivariate Adaptive Regression Splines (MARS). This is a modeling technique which is driven by data, and the approach is the use of a nonparametric regression which does not rely on the operational relationship between input and output data. Training data sets are divided into different piecewise segments referred to as splines with varying slopes. Basis functions which characterize the model are then derived by the technique through the delimiting of the splines in such a manner that sub-divisions are made between adjacent data regions (Friedman, 1991; Rezaie-Balf, 2018; Samui & Kothari, 2012; Yuvaraj, Murthy, Iyer, Samui, & Sekar, 2013)

2.0 Literatures Review

Ping Liu, *et al.* (2024), in their work aimed at training and testing a super learner strategy for risk prediction of kidney failure and mortality in people with incident moderate to severe chronic kidney disease (CKD). In their model, they used the super learner algorithm to select the best performing regression models or machine learning algorithms (learners) based on their ability to predict kidney failure and mortality with minimized cross-validated prediction error (Brier score, the lower the better). Prespecified learners included age, sex, eGFR, albuminuria, with or without diabetes, and cardiovascular disease. The index of prediction accuracy, a measure of calibration and discrimination calculated from the Brier score (the higher the better) was used to compare KDpredict with the benchmark, kidney failure risk equation, which does not account for the competing risk of death, and to evaluate the performance of

KDpredict mortality models. They concluded that KDpredict could be incorporated into electronic medical records or accessed online to accurately predict the risks of kidney failure and death in people with moderate to severe CKD. The KDpredict learning strategy is designed to be adapted to local needs and regularly revised over time to account for changes in the underlying health system and care processes. The implications their study details a new method of decision support for CKD by providing both mortality and kidney failure risk predictions. Younger adults with lower eGFR and higher albuminuria, who have a higher risk of kidney failure than death, are likely ideal candidates for referral to nephrology clinics. For many people with a higher risk of death than kidney failure, interventions targeting cardiovascular risk may be the priority. Individuals who have a very high risk of death may choose alternative treatments, including advance care planning with or without involvement of a kidney specialist. A wide range of risk combinations exist between these extremes, making treatment decisions challenging for patients, care givers, and health care providers. According to them kidney failure and death risk predictions are presented simultaneously and concluded that KDpredict supports holistic decision making in people with moderate to severe CKD. This research is similar to their work in the sense that both works are used to predict Kidney Failure, whereas they handled the Chronic Kidney Disease (CKD) and used KDpredict as their prediction model, Machine learning techniques (ML) have been a key driver in the development of CKD predictive models as they are applied in different contexts for the development of a desired predictive model. As an example, a combination of ML algorithms and Apache Spark as Big data platform was used by (Abdel-Fattah, Othman, & Goher, 2022) for the development of CKD predictive model with feature selectivity. The results of the prediction by the model were validated using four evaluation techniques which include accuracy, precision, recall, and F1-measure. The methodology used in their approach involves five essential steps: data collection (from the UCI machine learning repository for CKD dataset); data preprocessing which removes redundancies like null entries in the data set; the third step involves the selection of essential features; grid search and stratified cross-validation are performed in the fourth step for the optimization of the ML parameters and ensemble learning techniques; the fifth step is the evaluation of the developed model.

KNN, J48, ANN, Naïve Bayes, and SVM were used by (Zeynu & Patil, 2018) in the development of a predictive model for the diagnosis of CKD. The model is composed of two parts- feature selection method, and ensemble method. The feature selection method used information-gain attributes with ranker search engine and wrapper subset evaluator for feature selection operations. The ensemble method combined five heterogeneous classifiers based on a voting algorithm. A high degree of accuracy was achieved by the predictive model.

Naïve Bayes classifier and decision tree (J48) were used by (Kapoor, Verma, & Panda, 2019) for the development of a CKD predictive model. The UCI dataset were used by the authors in the development of the model. Based on certain features, the predictive model is able to classify a person as having CKD or not. The comparative analysis done on the model indicated that a high degree of accuracy was achieved by the J48 decision tree.

An ML methodology for the diagnosis of CKD was developed by (Alhamazani et al., 2021) using an anonymized dataset. Central to the development of the model was the use of CRISP-DM model (Cross Industry Standard Process for Data Mining) in the Azure cloud where the sample data was unbalanced. The SMOTE technique was used in balancing the data, and four matching AI algorithms – logistic regression, decision forest, neural network, and jungle of decisions were used in the model. The decision forest outperformed the other ML models with a score of 92%.

A feature-based prediction model for detecting kidney disease was proposed by (Poonia et al., 2022). The model involved several ML algorithms which includes KNN, ANN, SVM, naïve Bayes, Recursive Feature Elimination (RFE) and Chi-Square test feature selection. The algorithms were used for building and analyzing various prediction models on a publicly available dataset of patients with kidney disease. A logistic regression-based prediction model with optimal features chosen using the Chi-Square technique was identified as having the highest accuracy of 98.75%.

Following the review of some related works in the preceding discussion, Table 2.2 is derived showing the observed characteristics of the methodology developed by the authors and their observed shortcomings. The last column of the table is the proposed approach in this work; it shows the weaknesses in the other works which this work will address.

2.1 Overview of Multivariate Adaptive Regression Splines (MARS)

Multivariate adaptive regression splines belongs to the class of data-driven modeling techniques based on an approach that is can be referred to as multivariate nonparametric regression. In this approach, the functional relationship between the input and output data does not play a central role. By applying the technique of divide-and-conquer, training data sets are broken into separate piecewise linear segments called splines which are characterized by differing slopes. Knots are then used for the purpose of delimiting the segments through the identification of sub-divisions between any adjacent data regions in such a manner that it becomes possible to obtain piecewise curves known basis functions (BFs) (Friedman, 1991; Rezaie-Balf, 2018; Samui & Kothari, 2012; Yuvaraj et al., 2013). The mathematical representation of MARS can be expressed as (Friedman, 1991):

$$y = C_0 + \sum_{i=1}^{N} C_i \prod_{j=1}^{k_j} b_{ji}\left(x_{\vartheta(j,i)}\right) \qquad (1)$$

where y is the output of the relationship, C_0 is a constant, C_i is the vector of coefficients associated with BFs that are not constant, $b_{ji}\left(x_{\vartheta(j,i)}\right)$ is the BF having a truncated power and $\vartheta(j,i)$ as the index of the independent variable used in the i^{th} term of the j^{th} product, and k_j is a parameter that limits the interaction order.

Given any spline b_{ji}, such a spline can be defined as (Friedman, 1991; Rezaie-Balf, 2018; Samui & Kothari, 2012; Yuvaraj et al., 2013):

$$b_{ji}\,\overset{\mathrm{QV}}{x} = \left|x - t_{ji}\right|_{+}^{q} = \begin{matrix} \mathrm{GR}\, x - t_{ji}\ \mathrm{W}^{q}, x \ \$\ t_{ji} \\ 0, x \ 1\ t_{ji} \end{matrix}$$

$$b_{ji}\,\overset{\mathrm{QV}}{x} = \left|t_{ji} - x\right|_{+}^{q} = \begin{matrix} \mathrm{GR}\, t_{ji} - x\ \mathrm{W}^{q}, x \ 1\ t_{ji} \\ 0, x \ \$\ t_{ji} \end{matrix} \qquad (2)$$

For which t_{ji} is the knot of the given spline, $q(q>0)$ the spline power and the amount of smoothness of the expected function approximation. Generalized Cross Validation (GCV) is used for the determination of the basis function used in the, and it is the mean of the squared residual error divided by a penalty whose degree of complexity is proportional to the type of model and is expressed mathematically as (Friedman, 1991):

$$GCV = \frac{1}{N}\sum_{i=1}^{N}\left[y_i - f\left(x_i\right)\right]^2 \Big/ \left[1 - \frac{M + d \times (M-1)/2}{N}\right]^2 \qquad (3)$$

where the number of BFs is represented as *M*, with *d* representing the penalty for each BF in a given sub-model, N representing the number of data sets, and $f\left(x_i\right)$ representing the predicted values by MARS.

To make a fair comparison of the technique in this work, another ML technique will also be used on the same dataset; this will give an idea of how well MARS performs against other ML algorithms. A candidate ML choice that is capable of competing favorably with MARS for a particular dataset is multilayer perceptrons. This technique works well with any dataset that has an input/output relationship. The following subsection will present a brief overview of multilayer perceptron.

2.2 Overview of Multilayer Perceptron (MLP)

Multilayer perceptrons (MLP) are a class of deep learning neural networks that are well-suited for solving linear inseparable problems (Popescu, Balas, Perescu-Popescu, & Mastorakis, 2009). Their operation is feedforward trained with backpropagation algorithms. They belong to the class of supervised neural networks, and as such, they have to be trained for a desired type of output response. Multiplayer perceptrons are often used for classification and prediction (Pokonieczny, 2018).

Conceptually, a multilayer perceptron (MLP) is depicted as shown in figure 1 (Delashmit & Delashmit, 2005) where the inputs, outputs, and hidden layers are shown in figure 2.1.

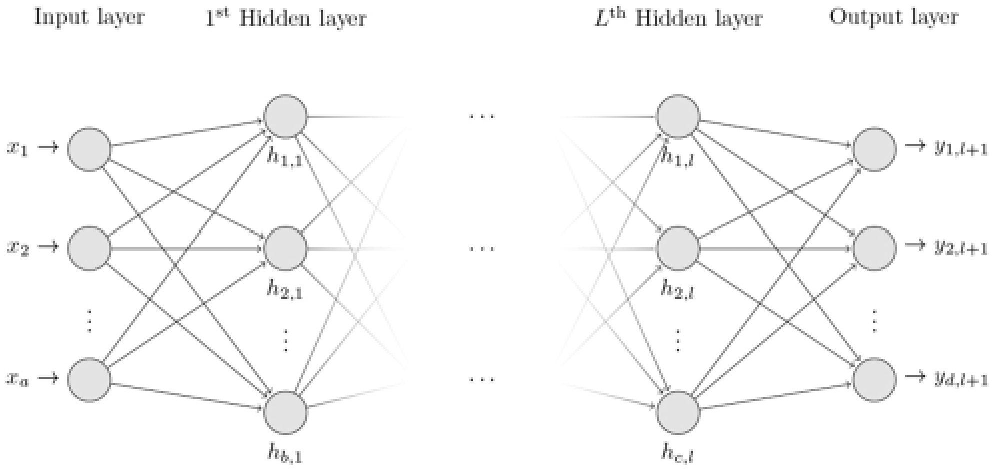


Figure 1: An MLP network structure. Source: (Castro, Oblitas, Santa-Cruz, & Avila-George, 2017)

The structure as shown makes it possible for MLP to learn tasks characterized by high complexity through the mechanism of feature extraction from input patterns. As a result of this, MLPs can be considered as a special of non-linear regression models because of their ability to handle the indirect relationships between the inputs and the outputs. The workings of the MLP is best understood through the dissection of its component parts; to this end, the following sub-sections will analyze the operations of an MLP from the perspective of its constituent parts.

2.2.1 Input and Output Patterns in MLP

Given that $\mathbf{x}_p$ is an N-dimensional input, $\mathbf{t}_p$ is a training pattern, and $\mathbf{y}_p$ is an M-dimensional output, the input to a *j*th hidden layer is expressed as (Delashmit & Delashmit, 2005; Olatunji, Akinlabi, Madushele, Adedeji, & Felix, 2019):

$$net_p(j) = \sum_{k=1}^{N+1} \mathrm{w}(j,k)\,\mathrm{x}_p(k), 1 \le j \le N \quad (4)$$

The output activation for the *p*th training defined as $Op(j)$ is expressed as follows (Delashmit & Delashmit, 2005; Olatunji et al., 2019):

$$Op(j) = f\left(net_p(j)\right) \quad (5)$$

where $f\left(net_p(j)\right) = \dfrac{1}{1+e^{-net_p(j)}}$ is a nonlinear sigmoidal activation function.

The *i*th output for the *p*th training pattern is expressed as (Delashmit & Delashmit, 2005; Lumacad & Namoco, 2022; Olatunji et al., 2019):

$$y_p(i) = \sum_{k=1}^{N+1} \mathrm{w}_{oi}(i,k)\,\mathrm{x}_p(k) + \sum_{j=1}^{N} \mathrm{w}_{oh}(i,j) O_p(j) \quad (6)$$

where $\mathrm{w}_{oi}(i,k)$ represents the weights from the input nodes to the output nodes, and $\mathrm{w}_{oh}(i,j)$ represents the weights from the hidden nodes to the output nodes.

2.2.2 Activation Functions in MLPs

There are different activation functions used by multilayer perceptrons; each of these functions affect the sensitivity of an MLP in different areas of applications. The following sub-sections presents these functions in the context of their mathematical definition.

2.2.1 Identity Function

The identity function returns the same value as its argument. It is defined as follows (Apicella, Donnarumma, Isgrò, & Prevete, 2021; López, López, & Crossa, 2022; Wanto, Windarto, Hartama, & Parlina, 2017):

$$y = \alpha \quad (7)$$

where the dependent variable y has a direct proportional relationship with the independent variable α

2.2.2 Heaviside Step Function

A piecewise binary-valued response exists between the dependent and independent variables in a Heaviside step. The relationship that defines the response is as follows (Adnan et al., 2017; Siddharth, Simone, & Anidhya, 2020):

$$f(x) = \begin{cases} 0, x < 0 \\ 1, otherwise \end{cases} \quad (8)$$

2.2.3 Bipolar Function

In the case of the Bipolar function, a piecewise relationship also exists between the dependent and independent variable as defined by the following relationship (Apicella et al., 2021):

$$f(x) = \begin{cases} -1, x < 0 \\ +1, otherwise \end{cases} \quad (9)$$

2.2.4 Sigmoid Function

This is a widely used activation function owing to its ability to saturate over a continuous range (A. Roy, 2022); it is therefore suitable for back propagation as it also exhibits a fast and a good convergence. It is mathematically described as follows (A. Roy, 2022) (Nantomah, 2019):

$$f\left(x\right)=\frac{1}{1+e^{-x}} \tag{10}$$

2.2.5 Hyperbolic Tan (Tanh) Function

This function is a well-known linear action function between layers in an MLP. It varies between $\left[-1,1\right]$, and it is mathematically expressed as (S. K. Roy, Manna, Dubey, & Chaudhuri, 2023):

$$\tanh\left(x\right)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{11}$$

The implementation of the tanh function is usually cost-prohibitive as it requires exponentiation and division operations; these types of operations require a high number of clock cycles to complete, hence they cause high latency in computing applications.

2.2.6 Rectified Linear Units (ReLU) Function

The ReLU function has a derivative function property that makes it possible to perform backpropagation and at the same maintain an efficient computation rate. It main attractive feature is that it does not activate all neurons at the same time, hence, it is energy saving. Mathematically, the ReLU function can be expressed as (Agarap, 2019; Ide & Kurita, 2017):

$$f\left(x\right)=\max\left(0,x\right) \tag{12}$$

2.3.7 Exponential Linear Units (ELU)

This class of function is a variant of the ReLU function, and it modifies the negative part of a function slope. Unlike leaky ReLU and parametric ReLU, it uses a log curve to define values of a function. It is mathematically expressed as (Clevert, Unterthiner, & Hochreiter, 2016):

$$f\left(x\right)=\begin{cases} x, x>0 \\ \alpha\left(\exp\left(x\right)-1\right), x\leq 0 \end{cases} \tag{13}$$

Table 2.2: Summary and characteristics of related works

S/N	**1**	**2**	**3**	**4**	**5**
Authors	Abdel-Fattah, Othman, & Goher	Poonia et al.,	Alhamazani et al.,	Kapoor, Verma, & Panda	Zeynu & Patil
Title of work	Predicting Chronic Kidney Disease Using Hybrid Machine Learning Based on Apache Spark	Intelligent Diagnostic Prediction and Classification Models for Detection of Kidney Disease	Implementation of Machine Learning Models for the Prevention of Kidney Diseases (CKD) or Their Derivatives	Detecting Kidney Disease using Naïve Bayes and Decision Tree in Machine Learning	Prediction of Chronic Kidney Diseas Using Data Mining Feature Selection and Ensemble Method

Year	2022	2022	2021	2019	2018
ML Algorithm	Decision tree, logistic regression, naïve Bayes, random forest, SVM, gradient-boosted tress	KNN, ANN, SVM, naïve Bayes, Recursive Feature Elimination, Chi-Square test feature	Logistic regression, decision forest, neural network, jungle of decisions	J48 decision tree, naïve Bayes	KNN, J48, ANN, naïve Bayes, SVM
Dataset source	UCI ML repository	Publicly available dataset	UCI ML repository	UCI ML repository	UCI ML repository
Data locality	Not local	Not local	Not local	Not local	Not local
Predictor variable identification	Not covered	Not covered	Not covered	Not covered	Not covered

3.0 Methodology

This section describes in practical terms the path that will be taken to realize the objectives in Section 1.3. Two critical pillars on which the methodology rests are the acquired data, and MARS. Figure 3.1 shows the workflow of the methodology in which the role of the acquired data and MARS are clearly indicated.

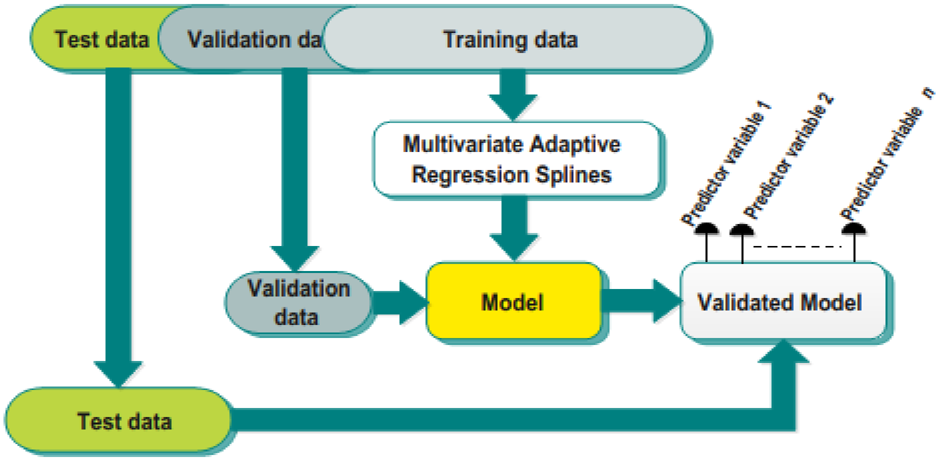


Figure 3.1: Methodology workflow

The data that used in the methodology is divided into three parts i.e. training data, validation data, and test data. The data will be sourced from hospital records in the geographic region of the study, and questionnaire that will be administered to willing members of the public in the same region.

3.1 Prediction Model Development

To develop the prediction model, two phases i.e. training phase and validation phase will executed as indicated in figure 3.1. Consequently, the training data and the validation data will be used in the prediction model development. In the training phase as shown in figure 3.1, MARS will be provided with the input data and the expected corresponding output data. MARS will then perform a learning process that will produce a model whose structure accurately maps the input data to the expected out data. The selection of the training data is such that it will be unbiased, and a good reflection of the test data. This approach will ensure that the generated model has a good generalization. The validation phase uses the validation data (which is actually a reserved part of the training data) for monitoring the performance of the model. This phase is important because it indicates the occurrence of overfitting by the model. The application of the validation data to the derived model as indicated in Figure 3.1 produces the validated model; the production of this model achieves the first objective of the study. As a general rule, for any given training-validation data, 80% should be used for the training phase, and 20% for the validation phase (Kim, 2017). Figure 3.2 shows a graphical process for the selection of the training, and the validation data.

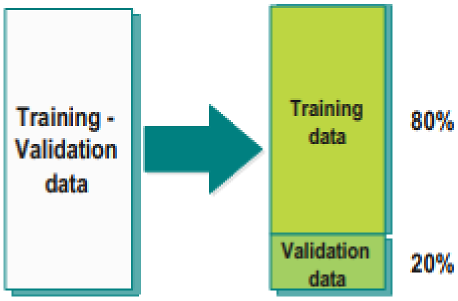


Figure 3.2: Training-validation data selection

3.2 Predictor Variables Identification

One of the key features of MARS is its ability to identify predictor variables in a model it creates. Consequently and as shown in figure 3.1, predictor variables are identifiable in a validated model. The validated model in this work contains predictor variables that have varying effect on the performance of the developed model in the form of basis functions as stated in equation 2.2.

3.3 Hardware Processor Design for Developed Prediction Model

The hardware design of a processor capable of making predictions based on the developed model will require the derivation of specifications based on the basis functions that characterize the behavior of the model. The derivation will yield two types of specifications; one will be specifications for the movement of data, and the second will be for the manipulation of data. The specifications for the movement of data will be handled through the RTL design flow, while the specifications for manipulation of data will be handled through the HLS design flow. Figure 3.3 shows the interaction between the specifications which will ultimately lead to the design of the processor.

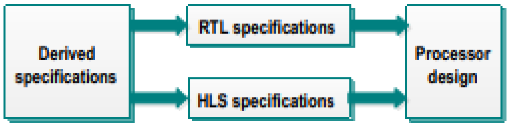


Figure 3.3: Processor RTL – HLS specifications

The block diagram for the design of the processor is shown in Figure 3.4. It can be observed that the communication between the three blocks of the processor i.e. inputs-processor-outputs will be performed using an AXI interface. The AXI interface which will be used in particular is the AXI4 stream interface. It is an interface which has a number of advantages like independent acknowledgement for address and data channel, out-of-order completion for bursts, system like cache support, and low power mode (Arm Ltd, 2022). The complete design of the processor for this work will be presented after the derivation of the model using MARS. The Basis Functions (BFs) are data manipulating processes, hence they will HLS-specified, and they take the form:

$$BF_j = C\left(x_m \,|\, -1, k_1, k_2, k_3\right) \qquad (3.1)$$

where k_1, k_2, k_3 are constants and x_m is a predictor variable.

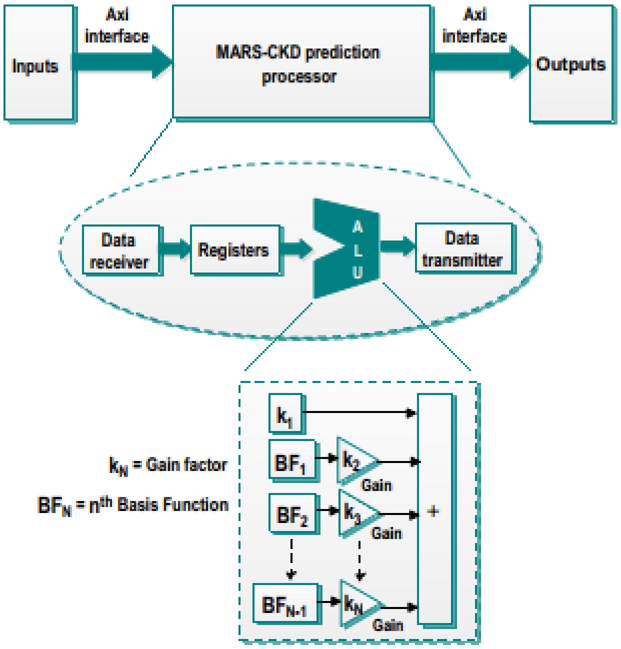


Figure 3.4: Block diagram of processor design

To realize the fourth objective of the study, an error analysis will be performed between the expected results and the actual results. The outcome of the analysis will be compared with the outcome obtained from similar works.

4.0 Results and Discussions

Analysis of Results for MARS Prediction Model

The first step in the development of the prediction model based on MARS is the selection of optimum number of basis that will be used in deriving the predictor variables that characterize the behavior of the model.

Using MARS, the optimum number of basis was estimated via Generalized Cross Validation (GCV) and Cross Validation as 43; this is shown in figure 4.1 of the capture of the MATLAB command window, and the corresponding plot of the GCV against the number of basis function in figure 4.2.

Forward phase
....
....
Number of basis functions in the model after forward phase: 75
Backward phase
....
....
Number of basis functions in the final model: 43

Figure 1: Screen capture of the number of basis function in final model

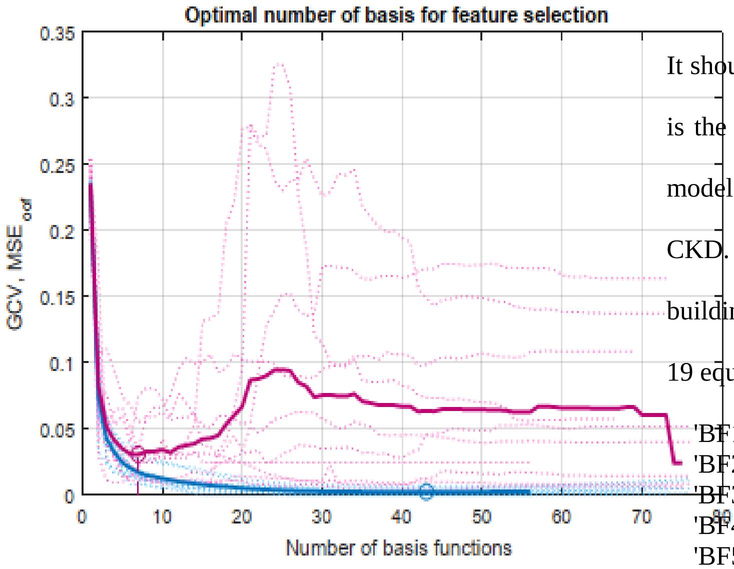


Figure 4.2: GCV for selection of optimal number of basis

The optimal number of basis i.e. 43 was used in building a MARS model for the prediction of CKD where the final number of basis functions for the MARS model was estimated to be 19 as shown in figure 4.3.

Approx number of available knot locations (controlled by useMinSpan and useEndSpan):
x1:38 x2:6 x3:5 x4:6 x5:6 x6:2
x7:2 x8:2 x9:2 x10:40 x11:43
x12:33 x13:22 x14:21 x15:40
x16:26 x17:32 x18:25 x19:2 x20:2
x21:2 x22:2 x23:2 x24:2

Forward phase
....
Number of basis functions in the model after forward phase: 37
Backward phase
....
Number of basis functions in the final model: 19

Figure 4.3: Estimation of the final number of basis functions for the MARS model

It should be noted that the 19th equation in the model is the final equation that computes the output of the model; it is the developed model for the prediction of CKD. Consequently, the basis functions derived from building the model as shown in figure 4.4 consists of 19 equations in total.

'BF1 = max(0, x15 -11.3)'
'BF2 = max(0,11.3 -x15)'
'BF3 = BF1 * max(0,1.02 -x3)'
'BF4 = max(0, x16 -38.88)'
'BF5 = BF4 * max(0, x4 -1)'
'BF6 = BF4 * max(0,1 -x4)'
'BF7 = BF1 * max(0,1 -x4)'
'BF8 = BF3 * max(0, x4 +0)'
'BF9 = BF3 * max(0,76.47 -x2)'
'BF10 = BF3 * max(0, x6 +0)'
'BF11 = BF1 * max(0,1 -x7)'
'BF12 = BF11 * max(0,80 -x2)'
'BF13 = max(0,13.4 -x15)'
'BF14 = max(0, x15 -13.4) * max(0, x4 +0)'
'BF15 = max(0,1 -x6) * max(0,1 -x4)'
'BF16 = max(0,1 -x4)'
'BF17 = BF16 * max(0,1.01 -x3)'
'BF18 = max(0, x16 -38.88) * max(0, x4 -1) * max(0, x1 -44) * max(0,70 -x2)'

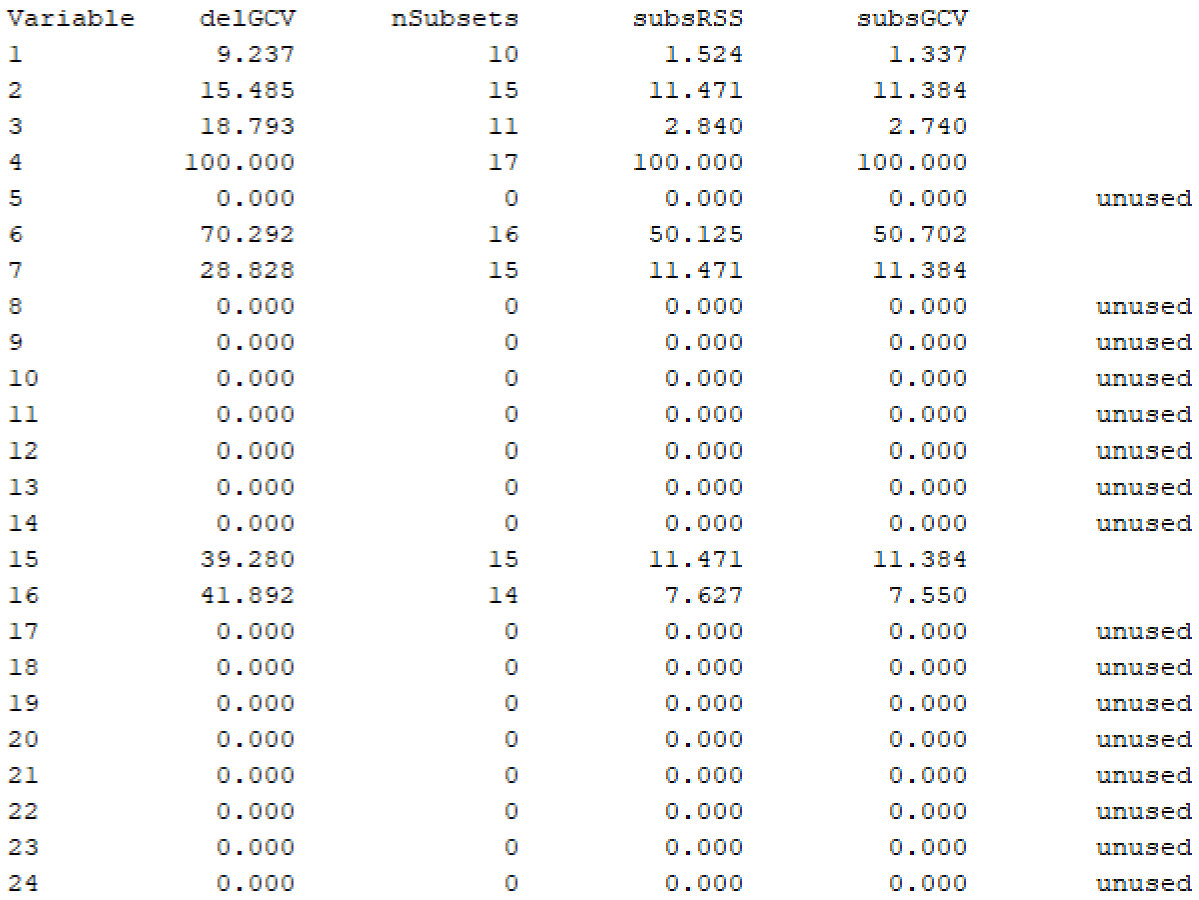
y = 0.8358 +0.1122*BF1 -0.07634*BF2 +11.95*BF3 -0.07453*BF4 +0.03892*BF5 +0.0678*BF6

-0.1075*BF7 -2.709*BF8 -0.3235*BF9 -
4.621*BF10 -0.03771*BF11 -0.006598*BF12
+0.07692*BF13 -
0.07984*BF14 +0.8251*BF15 -0.8151*BF16
+163.4*BF17 +0.0005267*BF18

Figure 4.4: Basis functions the CKD prediction model

4.2. Identification and Analysis of Predictor Weights Analysis

From the input data set, there are 24 predictor variables, and each of these variables has an effect on the MARS model. Figure 4.5 shows the predictor variables and the weight of each in the MARS model; predictor variable 4 representing Albumin has the strongest effect on the model; this is followed by predictor variable 6 representing red blood cells. Following these variables are predictor variables 16 representing pcv, 15 representing Hemoglobin, 3 representing specific gravity, 2 representing blood pressure, and 1 representing age.



```
Variable     delGCV      nSubsets      subsRSS      subsGCV
1             9.237            10        1.524        1.337
2            15.485            15       11.471       11.384
3            18.793            11        2.840        2.740
4           100.000            17      100.000      100.000
5             0.000             0        0.000        0.000      unused
6            70.292            16       50.125       50.702
7            28.828            15       11.471       11.384
8             0.000             0        0.000        0.000      unused
9             0.000             0        0.000        0.000      unused
10            0.000             0        0.000        0.000      unused
11            0.000             0        0.000        0.000      unused
12            0.000             0        0.000        0.000      unused
13            0.000             0        0.000        0.000      unused
14            0.000             0        0.000        0.000      unused
15           39.280            15       11.471       11.384
16           41.892            14        7.627        7.550
17            0.000             0        0.000        0.000      unused
18            0.000             0        0.000        0.000      unused
19            0.000             0        0.000        0.000      unused
20            0.000             0        0.000        0.000      unused
21            0.000             0        0.000        0.000      unused
22            0.000             0        0.000        0.000      unused
23            0.000             0        0.000        0.000      unused
24            0.000             0        0.000        0.000      unused
```

Figure 4.5: Effect of predictor variables on MARS model

A plot of the model is shown in figure 4.6 between variable x1 and x16; the choice of these variables is based on the fact that they span covers all the variables that were used in the model. The change in magnitude of the output between the two variables can be observed for values in the positive and the negative region.

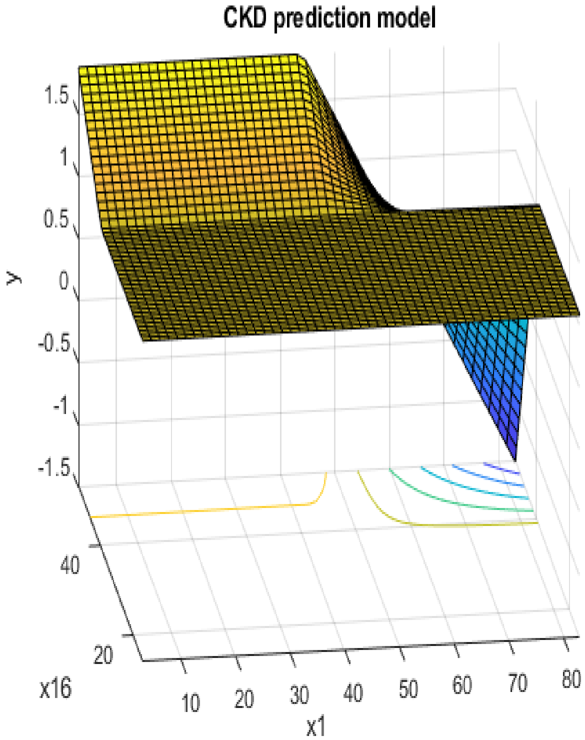


Figure 4.6: Performance of model between variable x1 and x16

Table 4.1 shows the performance of the model in terms of statistical error i.e. mean error and goodness of fit i.e. R^2 analysis. It can be observed that the MAE and MSE have values less than 0.1; this is an indication of high degree accuracy by the model. The RMSE value of less than 0.2 corroborates the accuracy of the model as reported by the MAE and MSE values.

Table 4.1: Error performance of prediction model

Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	R2
0.0870	0.0395	0.1924	0.8041

The high score for R2 indicates that the independent variables in the model i.e. predictor variables have established a very strong correlation with the output i.e. the dependent variable. Figure 4.7 shows this correlation for the training and the validation phase where a good tractability exists between the ideal values i.e. expected values and the actual values.

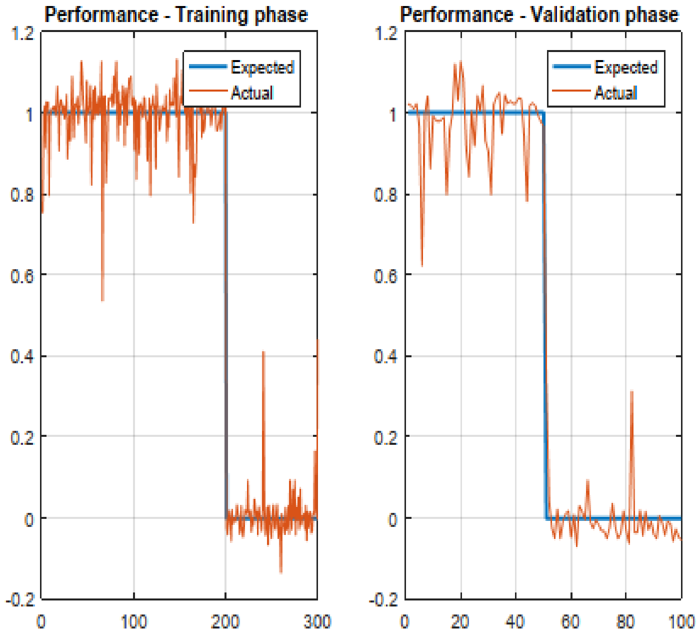
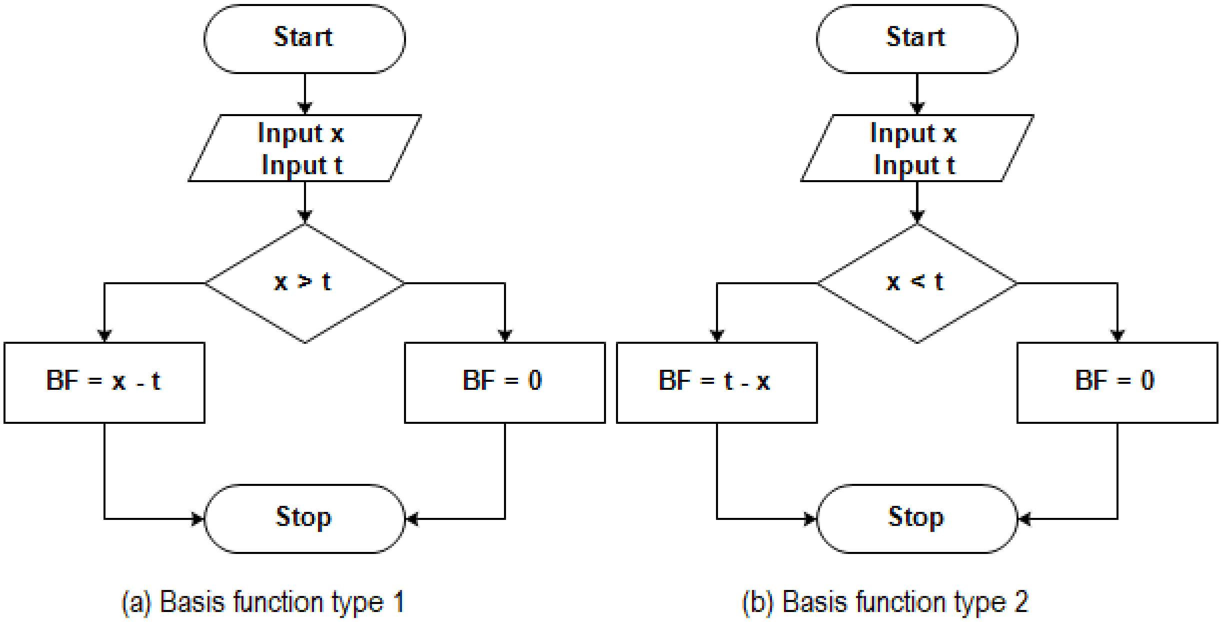


Figure 4.7: Model performance between actual and expected values

4.2. Analysis of Processor Design for MAR Model

The basis functions in equation 2.2 are the basis on which the processor is design. A flow chart representation for the execution of the equation in hardware is shown in figure 4.8 for basis function type 1 and basis function type 2.

The characteristic equation for the operation of the processor is shown in figure 4.4 where each of the basis function is represented. Based on the relationship between the basis functions defined in figure 4.4, the architecture for the processor is designed as shown in figure 4.8.



(a) Basis function type 1

(b) Basis function type 2

Figure 4.7: Basis function fundamental type for processor basis functions

4.3 VHDL-Based Implementation of Processor Architecture

The implementation of the architecture in figure 4.8 was effected using Very High Speed Integrated Circuit Hardware Description Language (VHDL). The first in the implementation is the conversion of the input data from floating point representation to integer representation. The reason for this is the fact that in signal processing, integer arithmetic outperforms floating arithmetic in terms of speed and memory utilization, while floating point arithmetic has the advantage of being more accurate. However, for most designs, the difference in accuracy is well within an accepted threshold; therefore, the accuracy is always traded for the performance gain in speed and memory utilization.

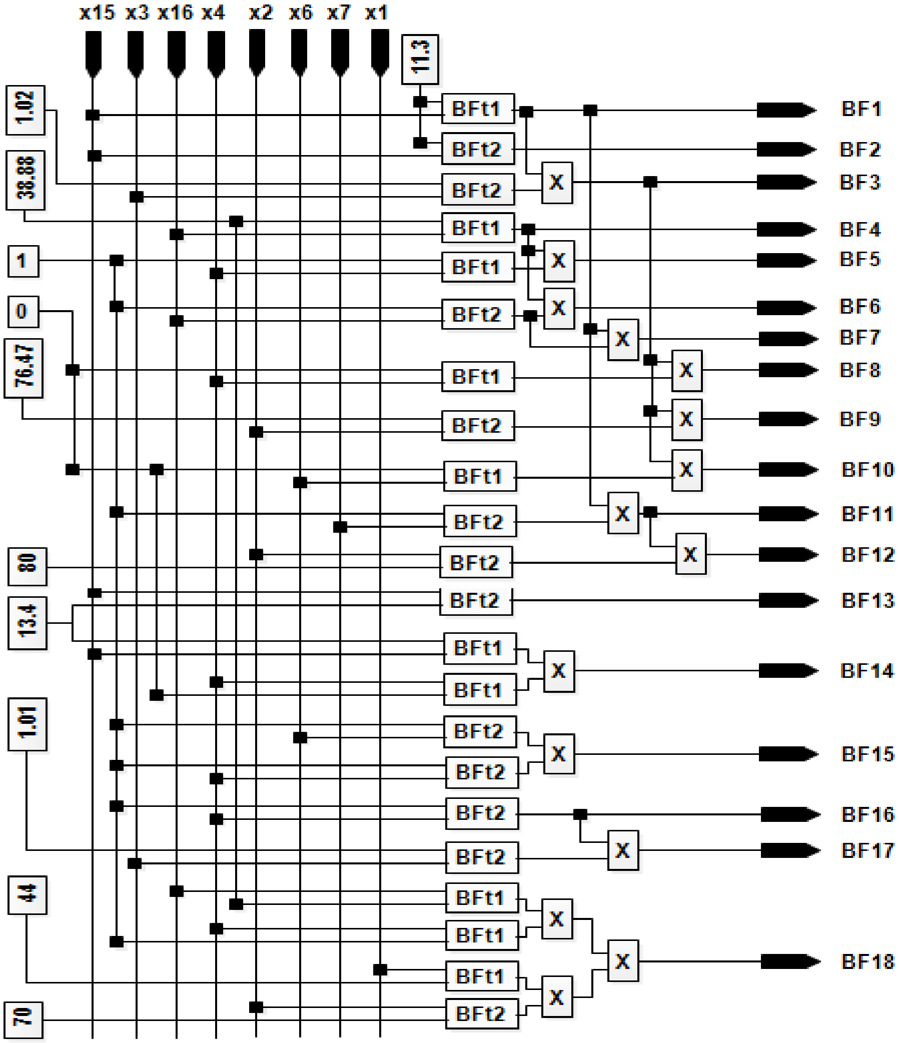


Figure 4.8: Architecture for processor

In figure 4.8, the input data to the processor are x1, x2, x3, x4, x6, x7, x15, x16. Each of these data is a vector i.e. a stream of data; they can be represented in compact form as:

$$\boldsymbol{X} = [\boldsymbol{x1} \quad \boldsymbol{x2} \quad \boldsymbol{x3} \quad \boldsymbol{x4} \quad \boldsymbol{x6} \quad \boldsymbol{x7} \quad \boldsymbol{x15} \quad \boldsymbol{x16}] \tag{4.1}$$

The vectors are converted from floating point representation to fixed point representation using the relationship $x_i = x_f \times 2^n$, where x_i is the desired integer representation, x_f is the current floating point representation, and n is the number of bits required for sufficient representation of data. For **X**, and the constants in figure 4.8, the value of n is 4, as it sufficiently represents all the data.

The simulation for the computation of the basis functions yields the result shown in figure 4.9 for 1500 ns.

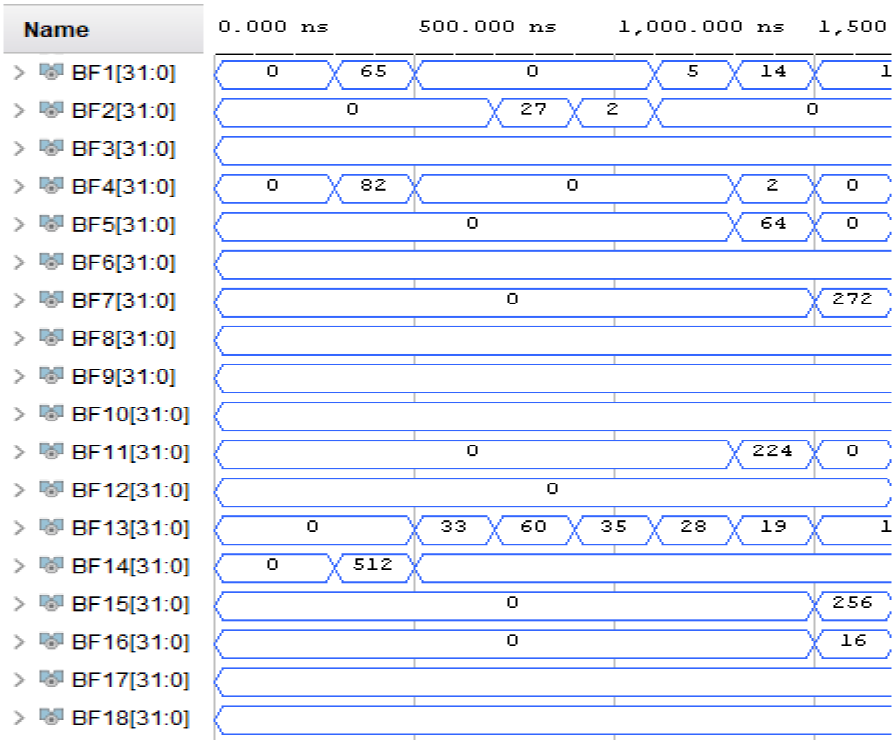


Figure 4.9: Computed basis functions for 1500 ns

From figure 4.4 and 4.8, it can be seen that all the basis functions are multiplied by constant factors. Consequently, a systolic array structure is designed as shown in figure 4.10 for the computation of the products from these multiplications. Each basis function as shown is multiplied by a unique factor concurrently with other basis functions with their unique factors.

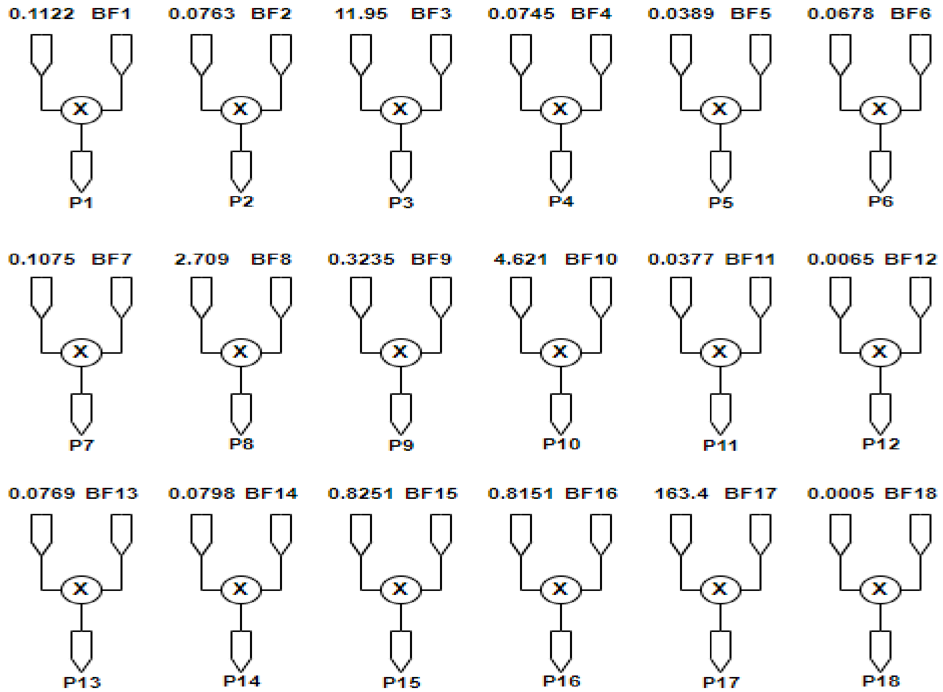


Figure 4.10: Concurrent multiplication of basis functions with unique factors

The array structure shown in figure 4.10 is implemented in the processor using VHDL. The simulation of the structure is shown in figure 4.11

5.0 Conclusion

In conclusion, this study successfully developed a machine learning model using Multivariate Adaptive Regression Splines (MARS) V2 to predict kidney function failure. The model demonstrated high accuracy, precision, recall, and F1-score, indicating its potential in clinical decision-making. The identification of serum creatinine, blood urea nitrogen, and urine protein as the most important predictors of kidney function failure highlights the significance of these biomarkers in kidney disease diagnosis. This study contributes to the growing body of research on machine learning applications in healthcare, particularly in predicting kidney function failure. The findings of this study can be used to develop clinical decision support systems, improving patient outcomes and reducing the burden of kidney disease.

REFERENCES

Abdel-Fattah, M. A., Othman, N. A., & Goher, N. (2022). Predicting Chronic Kidney Disease Using Hybrid Machine Learning Based on Apache Spark. *Hindawi - Computational Intelligence and Neuroscience*, *2022*, 1–12. https://doi.org/10.1155/2022/9898831

Adamowski, J., Chan, H. F., Prasher, S. O., & Sharda, V. N. (2012). Comparison of multivariate adaptive regression splines with coupled wavelet transform artificial neural networks for runoff forecasting in Himalayan micro-watersheds with limited data. *Journal of Hydroinformatics*, *14*(3), 731–744.

Adnan, J., Daud, N. G. N., Mokhtar, A. S. N., Hashim, F. R., Ahmad, S., Rashidi, A. F., & Rizman, Z. I. (2017). MULTILAYER PERCEPTRON BASED ACTIVATION FUNCTION ON HEART ABNORMALITY ACTIVITY. *Journal of Fundamental and Applied Sciences*, *9*(3S), 417–432. https://doi.org/10.4314/jfas.v9i3s.33

Agarap, A. F. M. (2019). Deep Learning using Rectified Linear Units (ReLU). *ArXiv E-Prints*, 1–7.

Alfego, D., Ennis, J., Gillespie, B., Lewis, M. J., Montgomery, E., Ferre, S., … Letovsky, S. (2021). Chronic Kidney Disease Testing Among At-Risk Adults in the U.S. Remains Low: Real-World Evidence From a National Laboratory Database. *Diabetes Care*, *44*(9), 2025–2032. https://doi.org/10.2337/dc21-0723

Alhamazani, K. T., Alshudukhi, J., Aljaloud, S., & Abebaw, S. (2021). ImplementationofMachineLearningModelsfort hePreventionof Kidney Diseases (CKD) or Their Derivatives. *Hindawi - Computational Intelligence and Neuroscience*, *2021*, 1–8. https://doi.org/10.1155/2021/3941978

Apicella, A., Donnarumma, F., Isgrò, F., & Prevete, R. (2021). A survey on modern trainable activation functions. *ArXiv E-Prints*, 1–30.

Arm Ltd. (2022). *AMBA AXI and ACE Protocol Specification* (pp. 1–500). pp. 1–500. Retrieved from https://developer.arm.com/documentation/ihi0022/e/

Asmelash, D., Chane, E., Desalegn, G., Assefa, S., Aynalem, G. L., & Fasil, A. (2020). Knowledge and Practices towards Prevention and Early Detection of Chronic Kidney Disease and Associated Factors among Hypertensive Patients in Gondar Town, North West Ethiopia. *International Journal of Hypertension*, *2020*, 1–8. https://doi.org/10.1155/2020/2860143

Assadi, F. (2012). The epidemic of pediatric chronic kidney disease: the danger of skepticism. *Journal of Nephropathology*, *1*(2), 61–64. https://doi.org/10.5812/nephropathol.7445

Bai, Q., Su, C., Tang, W., & Li, Y. (2022). Machine learning to predict end stage kidney disease in chronic kidney disease. *Nature - Scientific Reports*, *12*(8377), 1–8. https://doi.org/10.1038/s41598-022-12316-z

Bastos, M. G., & Kirsztajn, G. M. (2011). Chronic kidney disease: importance of early diagnosis, immediate referral and structured interdisciplinary approach to improve outcomes in patients not yet on dialysis. *Journal Brasilian Nefrology*, *33*(1), 74–87.

Bellizi V. et al. (2017). Controversial issues in CKD clinical practice: position statement of the CKD-treatment working group of the Italian Society of Nephrology. *Journal of Nephrology*, *30*(2), 159 – 170. https://doi.org/10.1007/s40620-016-0338-x

Castro, W., Oblitas, J., Santa-Cruz, R., & Avila-George, H. (2017). Multilayer perceptron architecture optimization using parallel computing techniques. *PLoS ONE*, *12*(12), 1–17. https://doi.org/10.1371/journal.pone.0189369

Ping Liu, Simon Sawhney, Uffe Heide-Jørgensen, Robert Ross Quinn, Simon Kok Jensen,Andrew Mclean, Christian Fynbo Christiansen, Thomas Alexander Gerds and Pietro Ravani (2024) " Predicting the risks of kidney failure and death in adults with moderate to severe chronic kidney disease: multinational, longitudinal, population based, cohort study.

Siddharth, S., Simone, S., & Anidhya, A. (2020). ACTIVATION FUNCTIONS IN NEURAL NETWORKS. *International Journal of Engineering Applied Sciences and Technology*, *4*(12), 310–316.

Skolnik, N. S., & Style, A. J. (2021). Importance of Early Screening and Diagnosis of Chronic Kidney Disease in Patients with Type 2 Diabetes. *Diabetes Therapy*, *12*, 1613 – 1630. https://doi.org/10.1007/s13300-021-01050-w

Teehan, G., & Benz, R. L. (2011). An Update on the Controversies in Anemia Management in Chronic Kidney Disease: Lessons Learned and Lost. *Hindawi - Anemia, 2011*(623673), 1–5. https://doi.org/10.1155/2011/623673

Tonelli, M., & Dickinson, J. A. (2020). Early Detection of CKD: Implications for Low-Income, Middle-Income, and High-Income Countries. *Journal of American Society of Nephrology*, *31*(9), 1931–1940. https://doi.org/10.1681/ASN.2020030277

Wanto, A., Windarto, A. P., Hartama, D., & Parlina, I. (2017). Use of Binary Sigmoid Function And Linear Identity In Artificial Neural Networks For Forecasting Population Density. *International Journal Of Information System & Technology*, *1*(1), 43–54.

Yuvaraj, P., Murthy, A. R., Iyer, N. R., Samui, P., & Sekar, S. K. (2013). Multivariate Adaptive Regression Splines Model to Predict Fracture Characteristics of High Strength and Ultra High Strength Concrete Beams. *CMC*, *36*(1), 73–97.

Zeynu, S., & Patil, S. (2018). Prediction of Chronic

Kidney Disease Using Data Mining Feature Selection and Ensemble Method. *WSEAS TRANSACTIONS on INFORMATION SCIENCE and APPLICATIONS*, *15*, 168–176.