THE SCIENTIFIC TEMPER (2026) Vol. 17 (4): 6056-6067

E-ISSN: 2231-6396, ISSN: 0976-8653

Doi: 10.58414/SCIENTIFICTEMPER.2026.17.4.16

https://scientifictemper.com/

RESEARCH ARTICLE

Chronic Kidney Disease Detection using Imputation-Aware Deep Neural Network

I. Siddik[1*], K. N. Abdul Kader Nihal[2]

Abstract

Chronic renal disease damages kidney function. Hypertension, diabetes, and cardiovascular disease are associated with chronic kidney disease. Risk factors for chronic kidney disease encompass age, genetic predisposition, hypertension, diabetes, obesity, proteinuria, and dyslipidemia. Tests, blood pressure measurements, and medical imaging can assist deep neural networks in diagnosing chronic renal disease. These models can identify minuscule patterns that are imperceptible to individuals with an accuracy of 97–100%. Artificial intelligence has progressed through deep neural networks, which are capable of processing intricate data and exhibit enhanced accuracy. Deep neural networks for tabular data reconstruct multi-layered connections among data points. Data, statistics, and machine learning experts employ this strategy to evaluate datasets and analyses containing missing data. Imputation substitutes missing data with alternative values. It infers each absent item, analyzes each completed dataset individually, and subsequently amalgamates the results after inputting various values. Imputation-aware deep neural networks effectively manage absent values. Managing real-world datasets with absent data is challenging. This is relevant to numerous enterprises, including the healthcare sector. To safeguard valuable data, these networks employ fundamental imputation to prevent users from removing missing rows or columns. This strategy preserves the sample size of the dataset to ensure the validity of statistical tests. It enables models to assimilate comprehensive data, reducing bias and improving precision.

Keywords: Chronic kidney disease, Deep neural networks, Multi-Layer Perceptron, Normalization, Handling Missing Values and Train–Test Split.

Introduction

Kidney failure is a critical issue which contributes an essential role in the growth and spread of many diseases.Hypotheses about dysfunction of kidney organ signify the significant and complicated impact on the state of health in general, which promotes the development of different chronic illnesses and associated problems. Chronic kidney disease is recognized as the condition that affects performance and prompts the further influence on other systems in the body (Larsen et al., 2024). Considerable research has proved an effective correlation between chronic kidney disease and cognitive impairment. It is highly vital to be aware of factors that will predispose you to the complications associated with chronic kidney disease. Number of factors has been identified to promote the progression including hypertension, diabetes, obesity and lifestyle choices (e.g., smoking and use of medications). Two areas that are increasingly getting intertwined are chronic kidney illness and medical data analysis. They adopt sophisticated approaches in order to learn more, prevent, diagnose, and treat chronic kidney disease. Data analytics is relevant to chronic kidney disease because it may enhance health outcomes, enhance treatment plans, and manage healthcare expenditures. These technologies can scan enormous amounts of data in a shorter amount of time, discover complex interactions between data and provide predictive insights in a scale never witnessed (Taha et al., 2022). Figure 1 represents the missing data which is a major issue when it comes to

[1]Research Scholar, PG and Research Department of Computer Science, Jamal Mohamed College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli – 620020, India

[2]Assistant Professor, PG and Research Department of Computer Science, Jamal Mohamed College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli – 620020, India

***Corresponding Author:** I. Siddik, Research Scholar, PG and Research Department of Computer Science, Jamal Mohamed College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli – 620020, India, E-Mail: isk@jmc.edu

How to cite this article: Siddik, I., Nihal, K.N.A.K. (2026). Chronic Kidney Disease Detection using Imputation-Aware Deep Neural Network. The Scientific Temper, **17**(4):6056-6067.

Doi: 10.58414/SCIENTIFICTEMPER.2026.17.4.16

Source of support: Nil

Conflict of interest: None.

Received: 15/03/2026 **Accepted:** 01/04/2026 **Published:** 25/04/2026

Original Dataset					Imputed Dataset			
A	B	C	D		A	B	C	D
1		3	4	Applying Data Imputation	1	2	3	4
5	6		8		5	6	7	8
9	10	11			9	10	11	12

Figure 1: Data imputation

data analysis and it may significantly influence the type of insights that can be made out of the information. Enter the missing numbers using the mean or median of the numbers available. This is a simple approach, yet it may reduce variability, and therefore, one must be cautious when he or she applies it particularly in cases where there are no values (Keerthana & Sherly Puspha Annabel, 2025).

Data imputation replaces missing values in datasets with substitute estimates to maintain data integrity and enable accurate analysis. This process is crucial in machine learning and statistics, as discarding incomplete data can introduce bias and reduce dataset size. User should be aware of the reasons of missing the data and apply the appropriate way to handle it. Ensure that user checks and enhances the way can collect data. Organizations can improve their data and their as well as analytic results by employing an imputation strategy that allows the use of imputation strategies with a high degree of control over data. Imputation is a statistical technique to estimate and insert missing data elements in a collection of data (Pereira et al., 2024). To be able to maintain a fair and bias-free dataset, it is significant to maintain as much data as possible. Complete-case analysis is usually applied when analysts lack imputation. This may reduce the sample size, and thus less reliable statistical results will be produced. It is applying the data that is present to give

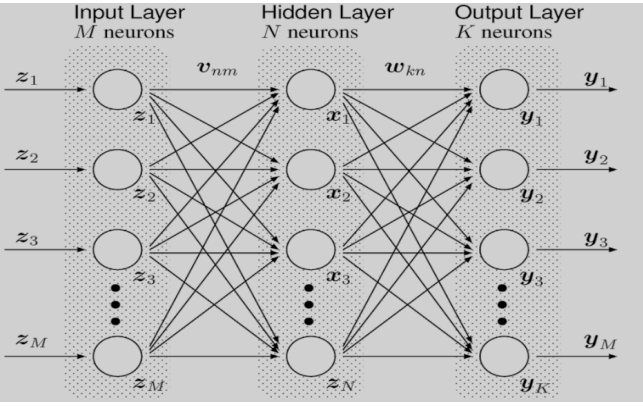


Figure 2: Neural Networks for Dual Input Processing

educated approximations over the missing figures. This ensures that the data is proper and the findings of statistical research is more precise.

To address the challenges that come along with the lack of data in datasets, and, as a result, enhance the performance of the models, there is a viable approach involving the utilization of sophisticated data modeling techniques that incorporate the aspect of missingness in the feature representation. Prior to developing matrices that are cognizant of Missingness, one should determine what they are. This indicates that a study of imputation and other follow-up studies can yield a greater effectiveness (Sullivan et al., 2021). Integrating the most useful concepts of contemporary machine learning and statistical models into a single solution would address the issues that arise when the data is not available. Advanced algorithms are able to discover the missing numbers by examining the relationships between the numbers. This is the advanced data imputation. Such a representation of the feature to be used in future research is more than a fill-in procedure.

It also looks into what exactly makes something to be absent and the manner in which it occurs in greater detail. Neural networks to handle duo-input data are capable of holding the conveyed values and the corresponding arrangement of missing values (Wan et al., 2025). By looking at the raw and imputed attributes simultaneously, models can better comprehend how the data is structured. Machine learning that involve artificial neural networks with more than two layers to extract the insights using big data. Deep learning is highly efficient since they are able to discover intricate trends in information. Certain individuals categorize deep learning as an independent subfield within the vast area which analyze and draw conclusions out of large volumes of data. Deep learning models are great tools for tasks like identifying images because they are based on how the brain works and are good at recognizing more complicated patterns in data.

Deep learning uses multi-layered artificial neural networks to look at and work with enormous amounts of data in order to do hard tasks. This is possible because the networks are designed to work like the brain (Al-Rasheed et al., 2025). It is precisely what makes them so helpful. The advancement in the neural network structures has enabled the deep learning algorithms to address challenging tasks in a variety of disciplines. Such strategies facilitate easier comprehension of complex concepts that are derived out of data by computers. The neurons are considered to be the nodes of a deep neural network (DNN). These neurons have so many layers. Once the data is received in the input layer, it passes through one or more hidden layers which handle it and then the prediction is made by the output layer. The network consists of weighted connections with all the cells being linked. The network manipulates and transforms the data via weights, biases, and activation functions.

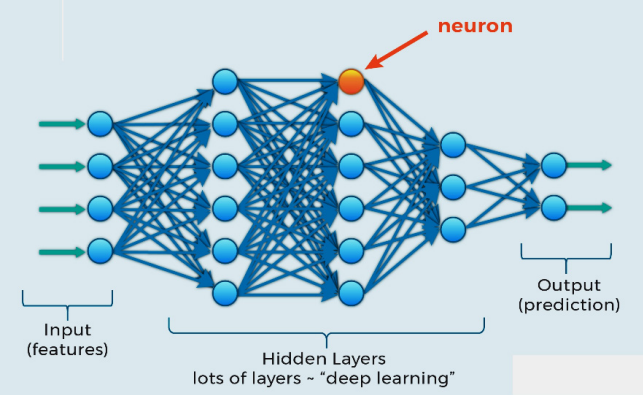


Figure 3: Deep Neural Networks

DNNs have the ability to automatically extract valuable information in raw data. The number of layers in these neurons is few. Information is received on input layer and subjected to hidden layers. Lastly, a prediction is made by output layer. Each cell in the network is connected with each other with varying weights. Weights, biases and activation functions are used in the network to change and process data in general. Deep Neural Networks are capable of extracting useful information of raw data in a short time having no assistance of an operator (Dhal et al., 2025). Through this, user can do more difficult things. They are also much slower and require more data, time and computer power to train as compared to other methods of learning. The Imputation-based deep neural networks (DNNs) are trendy deep learning models, which address the issue of missing data by identifying trustworthy values to compensate the differences. This ensures they are very precise and dependable particularly in those tough datasets that may identify intricate patterns and associations amid time and data. Multi-Layer Perceptron (MLP) is one examples of artificial neural network that contains a number of layers of interconnected neurons that are arranged into a hierarchy (Chu et al., 2021). This is done by ensuring that imputation process is also included in its design or training plan. Unlike regular MLPs, imputation-aware models do not require a complete dataset to operate. They can make better and less biased predictions when they apply context and they are excellent with incomplete data.

Literature Survey

There has been a change in the lifestyles and dieting patterns of the populations in the past years. They have changed their eating habits to consume more caloric foods and are less active because they do not engage in physical activities. Metabolic issues and other health complications such as heart disease and chronic kidney disease have been caused by the increasing rates of overweight and obesity. Premature aging phenotype is an indicator of chronic kidney disease (CKD). Patients with CKD have impaired biochemical and functional states associated with old age earlier than expected given their chronological age. Chronic kidney disease is an illustrative example of a clinical model of hastened biological aging, which is triggered by oxidative stress, ongoing inflammation, and cellular senescence, resulting in the multisystem failure, which mimics the normal aging process at accelerated pace

Its proposed (Braga et al., 2025) that chronic kidney disease (CKD) to be an abnormal renal structure and/or abnormal renal functional abnormalities lasting more than three months, which lead to progressive renal failure and impaired glomerular filtration rate (GFR).

It's proposed that (Larsen, D., Varanasi, L., & Estrella, M. M. (2024) chronic kidney disease poses significant challenges to the health systems and has adverse effects on the global health outcomes. These markers could help to understand the evolution of the disease and their impact on the cardiovascular condition, but they have not been studied in depth in the patients with chronic kidney disease. This literary review explores the complex connection between old age and chronic kidney disease. It discusses the significant biomarkers of aging, their significance in diagnosis and the fact that they can be diagnosed using serotherapies in order to address the issues associated with premature aging in this cohort of individuals.

According to (Bansal & Choncol, 2025), metabolic dysfunction related to kidney disease (MDAKD) is highly significant, and it needs to focus more on its clinical features and pathophysiological progression. MDAKD is kidney disease, which develops as a result of metabolic malfunction and the type of people who are pocked with metabolic syndrome. MDAKD has diabetes and obesity-related kidney disease and is being rapidly expanded to encompass a proliferation of rare kidney diseases. MDAKD is currently being grouped under a category of diseases brought about by metabolism dysfunctions. Recently, authors have been urged to stop using the old nomenclature (e.g., diabetic kidney disease) to use words that clarify that the pathophysiology of the diseases is based on metabolic dysfunction.

Some interesting ideas were discussed (Zobair et al., 2023), concerning the use of IoMT-powered wearable biosensors to serve rural indigenous populations with heart issues. The authors also discussed the impacts of these sensors on clinical care and telehealth in the treatment of heart disease. This paper makes us realize the functioning of telehealth models based on wearable IoMT sensors. It demonstrates the value of obtaining patient data in the remote location.

On the Internet of Medical Things (IoMT), (Ma et al., 2020) which introduced HMANN method which uses ultrasound image as a preprocessing phase and uses the kidney image as a segmentation of regions of interest. HMANN method

is extremely precise in the kidney contour segmentation problem and it is very fast compared to the former methods.

Its proposed (Nazari and Abdelrasoul, 2025) that the machine learning adopts methods in which such methods expound the relationships among the parameters in chronic kidney disease (CKD). Particularly the author study the outcomes on platelet counts and, as well, the prothrombin time (PT) levels. In this case author uses neural networks to assist him in assessing the impact of variables of the patients such as biological sex and age. It is significant that they are blood related.

Real-time model was introduced (Manoj Kumar et al., 2025) with efficient medical disease prediction and recommendation, it initially pre-processes provided medical data set. It then clusters provided data set to more than one level, and all this is in line with the disease stage. It extracts trace features and trains them with the help of neural networks once it has our groups. They approximate this value to various features. This computation is done on each of the classes of diseases and it reaches the output layer. Disease weight is then calculated using the method. It is interesting that a higher disease weight class is selected and this becomes the result.

Model offers (Sawhney et al., 2023) that provides perfect testing accuracy in a classification task through Deep Learning methods that directly gain knowledge of the reports of dataset attributes and effectively diagnose chronic kidney disease. The main contribution of this work is that it developed deep neural network architecture to diagnose chronic kidney disease with its perfect accuracy. The concise description of the type of multilayer perceptron that the model is a "deep" MLP classifier, in this instance, which is constructed with the help of the PyTorch library, is one of my favorite aspects of the paper.

Problem Statement

Chronic kidney disease complicates the elimination of waste and additional fluids of the blood by the kidneys and this may lead to the accumulation of harmful elements in the body. More and more, deep neural networks (DNNs) are used to diagnose chronic kidney disease with a high amount of accuracy. The high volume of the research results and theoretical models suggests that the models are excellent in CKD diagnosis and show impressive results. In statistics and machine learning, analysis of missing or incomplete data is the most important thing to do when dealing with data. It makes datasets handy and aids in making studies deliver the appropriate results. Imputation is an easy method of addressing this challenge by addressing the gaps. It is essential to secure databases and enhance data analysis. This method can ensure the protection of datasets, eliminate the risk of losing data and enhance the effectiveness of statistical and predictive models. Deep neural networks (DNNs) that are imputation-sensitive are a major advance in the problem of missing data in datasets. They are effective in most of the areas, yet in healthcare data in particular. The designs of these models have incorporated imputation techniques that have enabled them to learn strong representations and at the same time they do so with the incomplete data.

Methodology

The methodology for applying the Imputation-Aware Multi-Layer Perceptron (IAMLP) for the detection of chronic kidney disease (CKD) is structured comprehensively to address the challenges posed by missing and incomplete data in medical datasets. The dataset used primarily comes from public repositories, which contains various medical records of patients, including both CKD diagnosed and non-diagnosed individuals. This dataset typically includes

attributes such as age, blood pressure, specific gravity, and other relevant medical features necessary for CKD analysis. Data preprocessing is critical due to the prevalence of missing values in medical records. Categorical variables must be converted into numerical formats. To enhance model performance and reduce dimensionality, feature selection techniques are employed. The IAMLP integrates a multi-layer perceptron architecture that includes several hidden layers to capture complex data relationships. A unique component handles the missingness of data by adopting a dual approach feature that exhibit data with missing values are flagged and processed differently within the neural network. A robust validation process is implemented. The dataset is split into training and testing sets commonly utilized to ensure that every instance of the dataset is used for both training and testing, reducing the risk of overfitting. Important model parameters are fine-tuned using grid search or random search techniques to optimize the performance of the IAMLP. The effectiveness of the IAMLP model is evaluated using standard metrics like Accuracy, Precision, Recall and F1 Score. Finally, the methodology concludes with real-world implementation strategies, focusing on the application of IAMLP in clinical settings.

Dataset Description

Dataset used for the experimental analysis of CKD data set as shown in Table 1, which is the one that is used to identify kidney diseases and is known and commonly utilized in medical data analysis and predictive analytics. It has been narrowly tailored to accommodate studies on the early detection and categorization of CKD with regard to clinical and laboratory parameters. The dataset presents a harmonized set of patient records which is a reflection of the medical and biochemical picture of a particular patient. It is a perfect marker of intelligent diagnostic systems and machine learning algorithms, including the IA-MLP, that would manage missing data and lead to accurate disease prediction.

Table 1: Chronic Kidney Disease Dataset Description

S. No	*Attribute*	*Type*	*Explanation*
1	Patient Age	Numerical (Years)	Indicates the individual's age in years
2	Arterial Pressure	Numerical (mmHg)	Systolic/diastolic blood pressure reading
3	Specific Gravity Index	Categorical	Represents urine density for concentration analysis
4	Protein Albumin Level	Categorical	Measures albumin presence noted during urinalysis
5	Urine Glucose	Categorical	Glucose level detected in urine samples
6	RBC Condition	Categorical (Normal/Abnormal)	Status of red blood cells in urine
7	Pus Cell Status	Categorical (Normal/Abnormal)	Indicates white blood cell presence in urine
8	Pus Cell Aggregation	Categorical (Present/Absent)	Notes clustering of pus cells
9	Bacterial Trace	Categorical (Present/Absent)	Marks presence of infection-causing bacteria
10	Random Blood Glucose Level	Numerical (mg/dL)	Measures glucose level at a random time
11	Urea Level in Blood	Numerical (mg/dL)	Amount of urea circulating in the bloodstream
12	Serum Creatinine Value	Numerical (mg/dL)	Level of creatinine in serum for kidney efficiency
13	Sodium Concentration	Numerical (mEq/L)	Sodium electrolyte level
14	Potassium Concentration	Numerical (mEq/L)	Potassium electrolyte level
15	Hemoglobin Level	Numerical (g/dL)	Oxygen-carrying protein quantity in blood
16	Packed Cell Fraction	Numerical (%)	Proportion of packed red cells
17	WBC Count	Numerical (cells/cmm)	Number of white blood cells
18	RBC Count	Numerical (millions/cmm)	Count of red blood cells in circulation
19	Hypertension Indicator	Categorical (Yes/No)	Shows whether the patient has high blood pressure
20	Diabetes Condition	Categorical (Yes/No)	Identifies diabetic status
21	Coronary Artery Condition	Categorical (Yes/No)	Indicates history of cardiovascular disease
22	Appetite Status	Categorical (Good/Poor)	Describes appetite level
23	Pedal Swelling	Categorical (Yes/No)	Presence of edema in legs/feet
24	Anemia Status	Categorical (Yes/No)	Detects shortage of healthy red blood cells
25	Diagnostic Category	Target Class	CKD or Non-CKD classification

Data Preprocessing

Preprocessing of raw medical records is an essential stage before applying any machine learning or deep learning model thereby improving model reliability and predictive performance. Because medical data, including the one that is used in the detection of CKD, is usually not devoid of inconsistency, missing values, and variations in the scale of variables, adequate preprocessing is used to achieve data quality, consistency, and reliability of the predictive model. The dataset was subjected to the following systematic processes before the application of the IAMLP model. Assume that the initial CKD data is:

$$D = \left\{ \left(X_i, Y_I \right) \right\}_{i=1}^{N} \tag{1}$$

Where, $X_i = \left[X_{i1}, X_{i2}, \ldots..X_{id} \right]$ represents the feature vector of patient i with d attributes, $Y_i \in \{0,1\}$ represents the class label (0 = Non-CKD, 1 = CKD).

Data Cleaning

The initial phase involved data cleaning to remove unwanted noise and inconsistencies from the dataset. Duplicate records were identified and removed to maintain data integrity for the following:-

$$D' = D / \left\{ X_i | X_i = X_j \; for \; some \; i \neq j \right\} \tag{2}$$

Categorical inconsistencies such as "yes," "Yes," and "YES" were standardized to a uniform form:

$$X_{ij} = f_{std} \left(X_{ij} \right) \tag{3}$$

Where f_{std} is a text normalization function ensuring consistent categorical representation. Outlier detection for each continuous feature X_j was performed using the Z-score method as:

$$Z_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j} \quad (4)$$

Data points with $|Z_{ij}|>3$ were treated as outliers and replaced using median imputation as:

$$X_{ij}^{*} = \begin{cases} median(X_j), \ if\ |Zij| > 3 \\ X_{ij}, \ otherwise \end{cases} \quad (5)$$

This ensures robustness by reducing distortion from extreme values.

Handling Missing Values

One of the common issues with healthcare datasets is that they may contain missing values as a result of incomplete diagnostic records or absent lab results. A structured imputation strategy was used to deal with this problem. In case of continuous (numerical) variables, the imputation method was dependent on the distribution of the feature. Characteristics that have a near-normal distribution were imputed using the mean whereas characteristics with skewed distribution were imputed using the median to minimize bias due to non-symmetric characteristics. In the case of categorical variables, which include hypertension, diabetes mellitus, or appetite, the lack of a certain variable was substituted with the mode, which is the most common category. This hybrid method of imputation retained the statistical characteristics of every attribute and reduced the chance of maxims of bringing in artificial distortions in the data.

Let $\Omega \subset \{1,\ldots,N\} \times \{1,\ldots,m\}$ denote the set of missing entries.

For continuous attributes as:

$$X_{ij}^{*} = \begin{cases} \overline{X_j}, \ if X_j \ follows\ a\ normal\ distribution \\ median\ (X_j), \ if X_j \ is\ Skewed \end{cases} \quad (6)$$

For categorical attributes as:

$$X_{ij}^{*} = mode\left(X_j\right) \quad (7)$$

The hybrid imputation scheme preserves statistical distribution and reduces bias.

Missingness Flags

A distinctive and intelligent feature of the preprocessing pipeline was the incorporation of missingness flags. For each feature X_j containing missing entries, a binary indicator variable was introduced as:

$$m_{ij} = \begin{cases} 1, \ if\ (i,j) \in \Omega \\ 0, \ otherwise. \end{cases} \quad (8)$$

The augmented dataset thus becomes as:

$$X_i^{'} = \left[X_j, m_i \right] \epsilon \mathcal{R}^{2m}) \quad (9)$$

This enhancement allows the model to learn patterns related to missingness, providing imputation-awareness in the MLP architecture.

Normalization

Once the missing values and patterns of missingness had been addressed, the pattern of missingness (patterns) can be normalized using features to get them into the range of consistent values. Because various attributes in the CKD dataset have different ranges (e.g. blood pressure can be between 60180 mmHg, whereas serum creatinine can have values less than 10mg/dl), normalization makes sure that no single data point has influence on the learning process as a result of its size. In order to have similar feature scales, Min-Max normalization was used on each feature as:

$$X_{ij}^{norm} = \frac{X_{ij} - \min\left(X_j\right)}{\max\left(X_j\right) - \min\left(X_j\right)} \quad (10)$$

This transformation maps all feature values into the range [0,1], ensures where no single attribute dominates learning process and improving gradient stability during optimization.

Train–Test Split

The dataset was split into training and testing in order to determine how the model would perform on new data. The methodology involved a ratio of 80:20 with 80% of data being used to training IA-MLP model and 20% of data for testingIA-MLP model. The training set was used to optimize the parameters of the model, and the test set was used as independent data in an objective evaluation of the performance of the model. This split had the advantage of being the test predictive results of the trained model on data that was unseen previously, eliminating overfitting and giving a realistic estimate of its diagnostics ability. The end processed data set was divided in the following way:-

$$D_{train}\, U\, D_{test} = D',\ D_{train} \cap D_{test} = \varnothing \quad (11)$$

Where, $|D_{train}|$ = 0.8N, $|D_{test}|$ = 0.2N

The training subset was used for learning model parameters $\varnothing$, and testing subset for generalization capability. Accordingly, the resulting structured dataset is formulated into the final input matrix, as represented as:

$$X^{*} = \left[X_{clean}, M_{flag} \right]_{norm} \quad (12)$$

Where X_{clean} represents imputed data, and M_{flag} represents missingness indicators. This matrix forms the input to the Imputation-Aware MLP(IA-MLP) for CKD prediction.

The pretreatment of data ensured that the CKD dataset was free of garbage, complete, uniform, and scalable enabling the commencement of analysing it using deep

learning. Every cleaning, imputation, normalization and dataset partitioning procedure helped to strengthen the strength of the suggested IAMLP model. With the inclusion of missingness flags and intelligent imputation, the model did not only learn using the available medical data, but also using the underlying structure of missing data, such that, it provided a more accurate and reliable CKD prediction framework.

Imputation-Aware Multi-Layer Perceptron (IA-MLP)

Figure 4 elaborates on architecture of proposed IA-MLP model which had to be developed to be able to absorb nonlinear patterns and its relationship even where data is missing. The model was preprocessed, augmented with features, and trained based on the following methodology. The first stage of the IA-MLP network has a fully connected feed-forward issue. The input layer takes 2D features allowing the network to distinguish real and the imputed data points.

Mathematically, the computation in each of the hidden layers l can be written as:

$$h^{(l)} = \sigma\left(W^{(l)}h^{(l1)} + b^{(l)}\right) \tag{13}$$

Where $W^{(l)}$ represents the weight matrix of the l[th] layer, $b^{(l)}$ denotes the bias vector, and $\sigma(.)$ is the ReLU activation function defined as:

$$\sigma(x) = max(0, x) \tag{14}$$

This brings in non-linearity and avoids the vanishing gradient problems. Two hidden layers (64 and 32 neurons) are used in the network. This order is optimal in complexity, and computation speed of the model. The rate of dropout is 0.2, which is implemented after each hidden layer to randomly kill one fifth of neurons during training. It is a regularization method that reduces overfitting and improves model generalization by avoiding memorizing

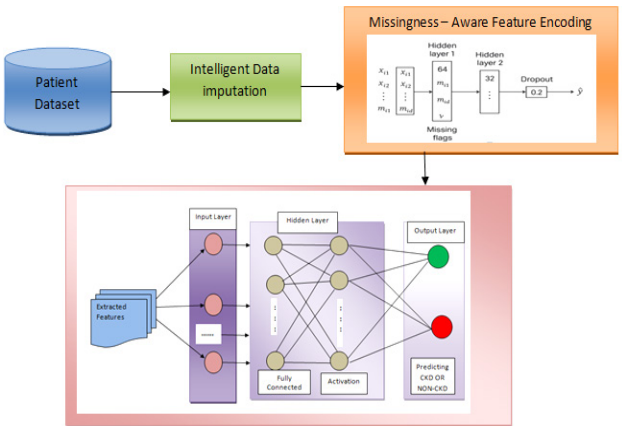


Figure 4: Architecture of the proposed Imputation-Aware Multi-Layer Perceptron (IA-MLP)

particular data patterns by the network. The output layer is made up of one neuron, which has a sigmoid activation function, which is good in the case of binary classification. In Equation (15), the network's final activation layer applies the sigmoid function, which converts the computed value into a probability ranging from 0 to 1. This probability indicates the likelihood that a given instance belongs to the CKD class.

$$\hat{y}_i = \sigma\left(W^{(l)}h^{(l1)} + b^{(l)}\right) = \frac{1}{1 + e^{\left(W^{(l)}h^{(l-1)} + b^{(l)}\right)}} \tag{15}$$

The second step model model training is the iterative optimization over 100 epochs which is sufficient to ensure a suitable learning convergence. The sub-steps of each epoch are: The pass of the input data through the network layers is sequential and the activations are calculated with the aid of the ReLU and sigmoid functions as indicated above. The Binary Cross-Entropy (BCE) loss function, is employed to quantify the deviation between the true class label y_i and the predicted probability $\hat{y}_i$ as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i log\left(\hat{y}_i\right) + \left(1 - y_i\right) log\left(1 - \hat{y}_i\right)\right] \tag{16}$$

This function penalizes incorrect predictions more heavily, ensuring balanced optimization for both CKD and non-CKD cases. The IA-MLP architecture is a feedforward neural network with the following configuration as Table 2.

Network is optimized to reduce binary cross entropy loss with Adam optimizer and an initial learning rate of 0.001. It was trained with 100 epochs and early stopping was done on the basis of validation loss. The Adam optimizer based on the calculated gradients as:

$$\theta \leftarrow \theta - n.\nabla_\theta L \tag{17}$$

Here, η represents the step size or learning rate applied during parameter optimization, and $\nabla_\theta L$ refers to the gradient loss function with respect to model parameters θ. Training willstopped prematurely in case of over-fitting. This criterion guarantees that the model has high generalization capacity and is not over adapting to the training data. All of these measurements determine the diagnostic reliability and discriminative capacity of the suggested framework.

Table 2: IA-MLP Architecture Configuration

Layer	*Type*	*Units*	*Activation*
Input	Combined features + missingness flags	25	Linear
Hidden Layer 1	Dense	64	ReLU
Hidden Layer 2	Dense	32	ReLU
Dropout Layer	Regularization	0.2	Linear
Output Layer	Dense	1	Sigmoid

Algorithm 1: Imputation-Aware Multi-Layer Perceptron (IA-MLP)

Input:

- $D = \{(X_i, Y_i)\}_{i=1}^{N}$ *N instances and d features.*
- f_{imp}*: Imputation function (mean/median for continuous, mode for categorical).*
- *Hyperparameters: learning rate η, number of epochs E=100, dropout rate p=0.2.*

Output:

- *Trained IA-MLP model M.*
- *Performance metrics*

1. Load dataset D with N instances and d features

2. For each feature x_j in D:

Compute mean or median (for continuous)

Compute mode (for categorical)

3. For each instance x_i in D:

For each feature x_ij:

If x_ij is missing:

Replace x_ij ← f_imp(x_j)

Set m_ij ← 1

Else:

m_ij ← 0

Concatenate: $\tilde{x}$_i ← [x_i ,M_i]

4. Normalize features: $\tilde{x}$_i ← ($\tilde{x}$_i - min) / (max - min)

5. Split dataset into training (80%) and testing (20%)

6. Initialize MLP architecture:

Input layer size = 2d

Hidden layers: [64, 32] neurons, ReLU activation

Dropout rate = 0.2

Output layer: 1 neuron, Sigmoid activation

7. For each epoch in range(1,100):

Forward pass:

Compute $h^{(l)} = \sigma\left(W^{(l)}h^{(l1)} + b^{(l)}\right)$

Compute loss L = BCE(y, ŷ)

Backpropagate and update weights using Adam optimizer

Apply early stopping if validation loss increases

8. Evaluate model on test set using: Accuracy, Precision, Recall, F1-score, ROC–AUC

9. Return trained IA-MLP model and performance metrics

Table 3: Performance Evaluation of the Proposed IA-MLP model

Metrics	*Training set*	*Testing set*
Accuracy	98.20%	96.50%
Precision	95.80%	94.70%
Recall(Sensitivity)	97.30%	95.60%
F1 Score	96.50%	95.10%

Results and Discussion

IAMLP model was built with the help of Python and TensorFlow and Scikit-learn applications to analyse the Kaggle CKD dataset, including both numerical and categorical health variables. The main aim of the study was to evaluate the effectiveness of the model in the process of reliably diagnosing CKD regardless of the occurrence of missing or partial medical records. The IA-MLP adopted an imputation-conscious method that effectively handles incomplete variables without impacting on the prediction efficiency. The model incorporated pre-processing techniques such as normalization, missing value imputation, and feature scaling, followed by deep learning training to explain the complex data associations. The results were compared with traditional machine learning to determine strength and correctness. Evaluation criteria like Accuracy, Precision, Recall (Sensitivity), F1-Score, and ROC–AUC Score were employed to evaluate the model's efficacy in accurately identifying CKD cases and providing dependable diagnostic performance in real-world data scenarios.

Experimental Setup

The proposed experimental setup implemented using Python 3.10 and appropriate libraries TensorFlow, Scikit-learn, NumPy, Pandas, and Matplotlib tools that are used to develop a model, preprocess the data, compute the numbers, and visualize the results. The main objectives during establishment of the system were to ensure that the system was easy to operate and accelerate business. The CKD dataset obtained from UCI Machine Learning Repository was partitioned into two subsets for experimentation. Approximately 80% of the records were assigned for model training, while the remaining 20% were reserved for performance evaluation on unseen test data with Adam optimizer to optimize the model weights.

Model Performance

The Table 3 demonstrated performance evaluation of proposed model which gives highly effective results both training and testing datasets.

As indicated in Figure 5, the proposed model achieved 98.20% and 96.50% on training and testing set respectively. This implies that it has a good generalization capability with minimal overfitting. There is not many false-positive predictions that are made by the model, with its training and testing precision ratings of 95.80% and 94.70% respectively.

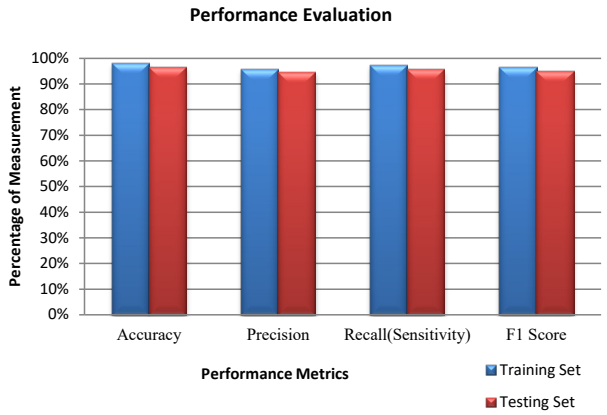


Figure 5: Performance Evaluation of the Proposed IA-MLP Model

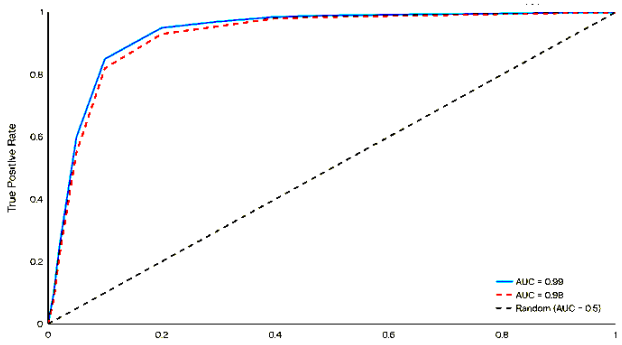


Figure 6: ROC-AUC Performance Evaluation Chart for IA-MLP Model

It implies that CKD patients can be detected with reliability. The recall (sensitivity) scores of 97.30%, and 95.60 percent, respectively, on training and testing data indicate that this model has ability to identify real CKD patients appropriately. The score of F1-score, a balance of precision and recall, was 96.50% in training and 95.10% in testing, indicating that the performance was stable and good.

Table 4: ROC-AUC Performance Evaluation Chart for the Proposed IA-MLP Model

Metric	Training Set Range [0.0 to 1.0]	Testing Set Range [0.0 to 1.0]
ROC-AUC	0.99	0.98

Table 5: Comparative Analysis of Various Machine Learning Models

Model	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*
Logistic Regression	91.8	90.2	89.6	89.9
Decision Tree Classifier	90.4	88.7	90.5	89.6
Random Forest	94.3	93.5	94.1	93.8
Support Vector Machine	92.6	91.4	90.8	91.1
Proposed IA-MLP	96.5	94.7	95.6	95.1

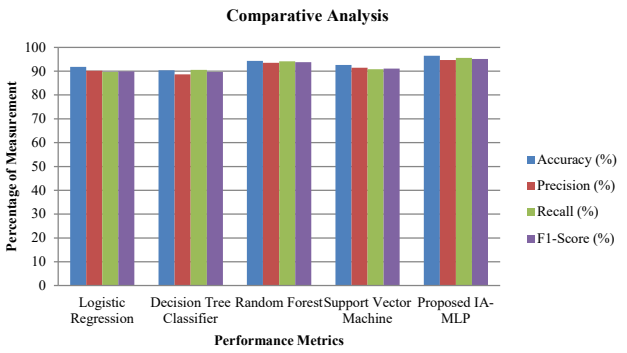


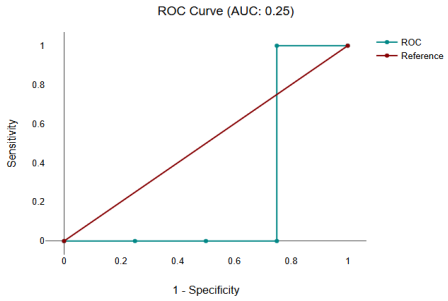
Figure 7: Comparison of Evaluation Metrics across Multiple Machine Learning Techniques

ROC AUC (Receiver Operating Characteristic - Area Under the Curve) is key machine learning metrics evaluating binary classifiers, measuring how well a model distinguishes between positive and negative classes by plotting True Positive Rate (Sensitivity) vs. False Positive Rate (1-Specificity) across various classification thresholds. The area under that ROC curve, providing a single score for model performance

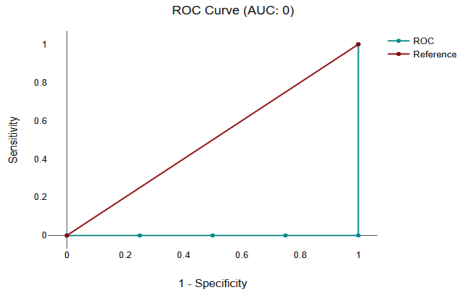
The fact that the ROC -AUC score is 0.98 for Testing Set and 0.99 for Training Set which is another indicator that the model is very good in distinguishing between CKD and no-CKD patients. This demonstrates that it is generally valid and applicable in medical diagnosis.

Table 6: ROC-AUC Performance Evaluation Chart for the Proposed IA-MLP Model

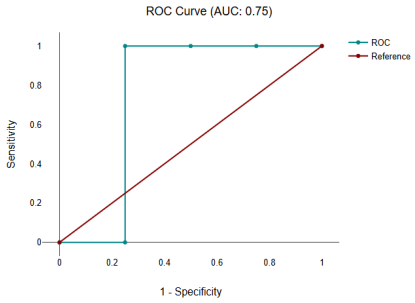
Model	*ROC–AUC Range [0.0 to 1.0]*
Logistic Regression	0.92
Decision Tree Classifier	0.91
Random Forest	0.95
Support Vector Machine	0.93
Proposed IA-MLP	0.98



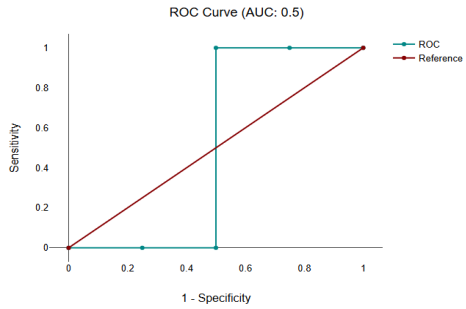
8a) Logistic Regression



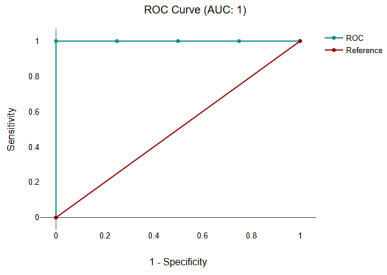
8b) Decision Tree



8c) Random Forest



8d) SVM



8e) Proposed IA-MLP

Figure 8a-e):ROC-AUC Performance Evaluation Chart for the Proposed IA-MLP Model

Comparative Analysis of Various Machine Learning Models

As the comparative analysis of the various machine learning models used to predict CKD as presented in Table 5 shows, proposed IA-MLP outperformed all the standard classifiers in all evaluation criteria. The IA-MLP was the most accurate (96.5%), the most precise (94.7%), the most recalled (95.6%), and most F1-score (95.1) thus indicating that it was the most competent in identifying CKD patients correctly without including false positives and false negatives.

Conversely, traditional models showed somehow lower results with an accuracy of 90.4% to 94.3% percent. The best performance of conventional methods was achieved by the Random Forest, however, the proposed IA-MLP in Figure 7 was significantly better because of its deep learning framework and the mechanism of imputation-awareness.

Conversely, traditional models such as the Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine showed somehow lower results of ROC -AUC ranges from 0.92 to 0.98. The best performance of conventional methods was achieved by the Random Forest, however, the proposed IA-MLP in Table 6was significantly better because of its deep learning framework and the mechanism of imputation-awareness. The ROCAUC measure of 0.98 also indicates that the IA-MLP model is more useful in data classification and generalization as compared to traditional methods.

Conclusion

The IA-MLP model was particularly successful in identifying CKDand it's justified. It does not just learn how input relates to output, but also within the inputs themselves, between the biochemical variables, the missing values and a myriad of possible disease-relevant patterns present in the entire dataset. The over-fitting-reducing dropout regularization by the team also ensures that the impressive performance of model is achieved by genuine abilities of model, not just as a consequence of false correlations.

The ROC -AUC score of 0.98 is a clear indication that the IA-MLP is excellent at distinguishing between CKD and non-CKD cases. That is, it is highly accurate. The other thing it did is to rank the features which seemed to be considered the most important to the IA-MLP decisions. It has already known through clinical experience that these are highly informative measures of kidney health and, therefore, the salience of these measures in the model makes clinical sense and further confirms the model.

Acknowledgement

We acknowledge the support of the PG & Research Department of Computer Science, Jamal Mohamed College (Autonomous), Trichy for providing Compu-tational resources. This research was not supported by any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Al-Rasheed, A., Saqib, S. M., Asghar, M. Z., Mazhar, T., Osman, A. S. A., Shahid, M., Iqbal, M., & Khan, M. A. (2025). Classifying kidney disease using a dense layers deep learning model. *SLAS Technology, 33*. DOI: 10.1016/j.slast.2025.100324

Bansal, A., & Chonchol, M. (2025). Metabolic dysfunction–associated kidney disease: Pathogenesis and clinical manifestations. *Kidney International, 108*(2), 194–200. DOI: 10.1016/j.kint.2025.01.044

Braga, C. E. H. V. P. F., de Brito, J. S., Ribeiro, M., Coutinho-Wolino, K. S., Regis, B., Calixto, B., Rodrigues, R. C. B., Wang, A. Y. M., Stenvinkel, P., & Mafra, D. (2025). Premature aging in chronic kidney disease: Decoding senescence biomarkers and therapeutic opportunities. *Biochimie*. DOI: 10.3390/ijerph18158044

Chu, J., Chen, J., Chen, X., Dong, W., Shi, J., & Huang, Z. (2021). Knowledge-aware multi-center clinical dataset adaptation: Problem, method, and application. *Journal of Biomedical Informatics, 115*. https:doi.org/10.1016/j.jbi.2021.103710

Dhal, P., Pradhan, B., Fiore, U., Francis, S. A. J., & Roy, D. S. (2025). A clinical diabetes prediction based support system based on the multi-objective metaheuristic inspired fine tuning deep network. *Information Fusion, 122*. DOI:10.1016/j.inffus.2025.103188

Keerthana, G., & Sherly Puspha Annabel, L. (2025). A survey on big data classification. *Data & Knowledge Engineering, 156*. https://doi.org/10.1016/j.datak.2025.102408

Larsen, D., Varanasi, L., & Estrella, M. M. (2024). Chronic kidney disease—Part 1: Evaluation & risk assessment in CKD, methods to delay CKD progression. *Advances in Kidney Disease and Health, 31*(6), 538–545. DOI: 10.1053/j.akdh.2024.07.004

Ma, F., Sun, T., Liu, L., & Jing, H. (2020). Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network. *Future Generation Computer Systems, 111*, 17–26. DOI:10.1016/j.future.2020.04.036

Manoj Kumar, M., Siva, R., & Baskar, M. (2025). Real-time multi level chronic disease prediction and recommendation model using deep learning. *Results in Engineering, 28*. https://doi.org/10.1016/j.rineng.2025.107478

Nazari, S., & Abdelrasoul, A. (2025). Neural network-based analysis of clinical and demographic variables for predicting platelet counts and prothrombin time in chronic kidney disease patients. *Engineering Applications of Artificial Intelligence, 159*(Part C). https://doi.org/10.1016/j.engappai.2025.111741

Pereira, R. C., Abreu, P. H., Figueiredo, M. A. T., & Rodrigues, P. P. (2024). Imputation of data missing not at random: Artificial generation and benchmark analysis. *Expert Systems with Applications, 249*(Part B). https://doi.org/10.1016/j.eswa.2024.123654

Sawhney, R., Malik, A., Sharma, S., & Narayan, V. (2023). A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease. *Decision Analytics Journal, 6*. https://doi.org/10.1016/j.dajour.2023.100169

Sullivan, T. R., Lee, K. J., Ryan, P., & Salter, A. B. (2021). Missing data in clinical research: A tutorial on multiple imputation. *Canadian Journal of Cardiology, 37*(9), 1322–1331. DOI: 10.1016/j.

cjca.2020.11.010

Taha, A., Iman, Y., Hingwala, J., Askin, N., Mysore, P., Rigatto, C., Bohm, C., Komenda, P., Tangri, N., & Collister, D. (2022). Patient navigators for CKD and kidney failure: A systematic review. *Kidney Medicine, 4*(10). DOI: 10.1016/j.xkme.2022.100540

Wan, Q., Liu, J., Liu, T., Zhou, R., & Qin, P. (2025). Dual channel and dual feedback loop self-learning memristive neural network circuit and its application. *Neural Networks, 192*. DOI:10.1016/j.neunet.2025.107929

Zobair, K. M., Houghton, L., Tjondronegoro, D., Sanzogni, L., Islam, M. Z., Saker, T., & Islam, M. J. (2023). Systematic review of Internet of Medical Things for cardiovascular disease prevention among Australian First Nations. *Heliyon, 9*(11). https://doi.org/10.1016/j.heliyon.2023.e22420