Research Journal of Humanities and Social Sciences
(An Open Access International Peer Reviewed Journal)
ISSN: 2945-3968

Review Article

Early Detection of Chronic Kidney Disease Using Logistic Regression with SMOTE Balancing and Grid Search Optimisation: Evidence of Benchmark Saturation and the Case for a Reduced Primary Care Panel

Ugwu Akor Peter[1*], Yateghtegh Solomon Kershima[2], Ewaoche Ameh Emmanuel[3]
[1]Department of Computer Science, Benson Idahosa University, Benin City, Nigeria
[2]Department of Computer Science, Tansian University, Umunya, Anambra State, Nigeria
[3]Department of Medicine and Surgery, Benue State University, Makurdi, Nigeria
Corresponding Author : **Ugwu Akor Peter**

Abstract
Chronic kidney disease advances silently until most kidney function is lost, and in low resource systems patients present when dialysis is the only option. This study develops and stress tests an early prediction model using logistic regression, a classification technique widely applied in clinical research. The model was implemented in Python with the Chronic Kidney Disease dataset from the UCI Machine Learning Repository, holding 400 patient records, 24 clinical and demographic predictors and one binary outcome, with 250 cases and 150 non cases. Exploratory analysis identified 1,012 missing entries, led by red blood cells at 38.00 percent and red blood cell count at 32.75 percent of records. Numeric attributes were imputed with the median and categorical attributes with the mode. The data was divided 70 to 30, and the Synthetic Minority Oversampling Technique was applied to the training partition alone, balancing it at 175 records per class. The base model returned 98.33 percent accuracy, 98.40 percent weighted precision, 98.33 percent weighted recall, an F1 score of 98.34 percent and a ROC AUC of 100.00 percent. Grid search over the penalty term and regularisation strength selected an l2 penalty with C equal to 2.8072, yielding no improvement on held out data at 96.67 percent accuracy. Ten fold cross validation returned 99.14 percent. A reduced panel of eight inexpensive predictors averaged 97.11 percent across 30 repeated splits, against 98.53 percent for the full 24 predictor model. The benchmark is saturated, and practical gain lies in cutting input cost rather than chasing accuracy.
Keywords: Chronic Kidney Disease; Grid Search; Logistic Regression; Machine Learning; Screening; SMOTE

How to cite this article: Peter UA, Kershima YS, Emmanuel EA . Early Detection of Chronic Kidney Disease Using Logistic Regression with SMOTE Balancing and Grid Search Optimisation: Evidence of Benchmark Saturation and the Case for a Reduced Primary Care Panel. **Research Journal of Humanities and Social Sciences.2026 Oct 10, 5(4):216-236**
Source of support: Nil.
Conflict of interest: None
DOI: doi.org/10.58924/rjhss.v5.iss4.p15

Received: 10-08-2026
Revised: 12-08-2026
Accepted 25-09-2026
Published: 10-10-2026

Copyright:© 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. INTRODUCTION

Chronic kidney disease is a gradual, irreversible decline in kidney function that takes place over months and years. Kidneys remove waste products from the blood and help regulate fluids and electrolytes, blood pressure, and red blood cell production. As nephron damage progresses, these functions are lost one after another, and the damage often goes unnoticed for a long time. A patient does not lose filtration capacity until it is significantly reduced. It is characterised by abnormalities in kidney structure or function for at least 3 months, as determined by the estimated glomerular filtration rate and albumin-to-creatinine ratio (ACR) [3, 25].

1.1 Stages of Chronic Kidney Disease

Stage 1. Kidney damage with normal or raised filtration, with an estimated glomerular filtration rate of 90 mL/min/1.73 m² or above together with albuminuria. Patients report no symptoms at this point.
Stage 2. Mild reduction in filtration, with a rate between 60 and 89. Detection at this stage

depends almost entirely on laboratory testing rather than clinical presentation.
Stage 3. Moderate reduction, with a rate between 30 and 59. This stage holds the largest share of the global chronic kidney disease population. Anaemia, bone mineral disorder and hypertension begin to appear.
Stage 4. Severe reduction, with a rate between 15 and 29. Preparation for renal replacement therapy begins here.
Stage 5. Kidney failure, with a rate below 15. Dialysis or transplantation is required to sustain life.

1.2 Causes and Risk Factors

In most developed and developing countries, diabetes mellitus and hypertension are the top two causes of chronic kidney disease. A further portion is due to glomerulonephritis, obstructive uropathy, polycystic kidney disease and infection. The pattern differs in sub-Saharan Africa, where glomerulonephritis and causes of unknown origin are more prominent, and exposure to heavy metals, unregulated herbal preparations, analgesic overuse, and sickle cell disease are more common [25]. Age, sex, smoking, obesity, and human immunodeficiency virus infection increase individual risk. Figure 1 summarises the reported contribution of the major causes.

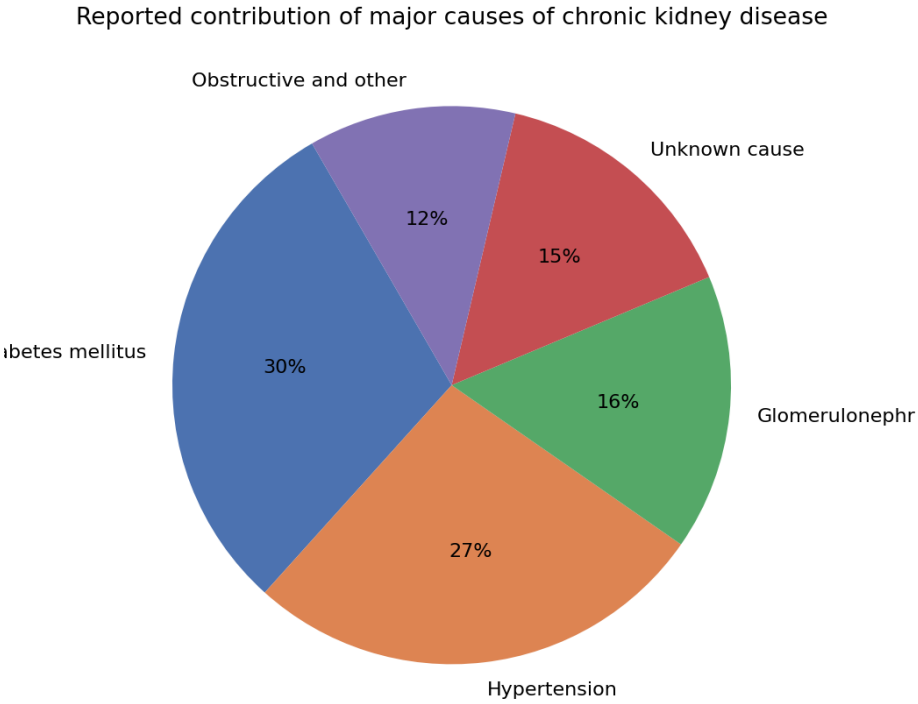


Figure 1: Reported contribution of the major causes of chronic kidney disease

1.3 Clinical Presentation and Systemic Effects

Early signs include fatigue, reduced urination, foamy urine, swelling of the face and ankles, decreased appetite, nausea, itching, and trouble sleeping. These signs resemble those of a common illness, and patients and primary care professionals often attribute them to that. Damage is not limited to the kidney, as filtration decreases. Heart and blood vessel disease, anaemia, bone disorder, metabolic acidosis and hypertension are all consequences of chronic kidney disease and increase the risk of death to many times that of the general population [13]. Pooled prevalence for cognitive impairment is also around 40 percent in all chronic kidney disease populations and is more prevalent in haemodialysis patients [24].

1.4 Statistics of Chronic Kidney Disease

The Global Burden of Disease Study 2021 (GBD 2021) estimates that in 2021, 673.7 million people globally had chronic kidney disease, 1.53 million people died from the disease and 44.45 million disability adjusted life years (DALYs) were lost due to chronic kidney disease at a standardised rate of 8,006 per 100,000 people [3]. During the last 30 years, the number of deaths due to the condition has increased by nearly one-quarter, and it is estimated that the number of deaths will double to 4 million by 2040, when chronic kidney disease will become the fifth biggest cause of death worldwide [13]. Awareness remains the weakest link, with about 10 percent of all at-risk individuals worldwide aware of their status [13]. Low SDI regions have the highest age-standardised death rates and disability burden, and are where screening and renal replacement capacity are the least available.

Africa bears a disproportionate burden. One systematic review and meta-analysis performed throughout the continent estimated the overall prevalence at 15.8 percent in all stages (1 to 5) and 4.6 percent in later stages (3 to 5) in the general population, which rose to 32.3 percent and 13.3 percent respectively among high-risk groups (those with hypertension, diabetes or human immunodeficiency virus infection). Sub-Saharan prevalence rates were approximately three times those of North Africa [16]. A later review estimates sub-Saharan prevalence at close to 18 percent, corresponding to about 174 million people [25]. The pooled prevalence of type 2 diabetes among adults in sub-Saharan Africa is 35.8 percent or nearly one out of three people [23]. Reported prevalence varies widely in Nigeria and across study methods, which is linked to the lack of systematic screening [25]. Figure 2 compares these reported figures.

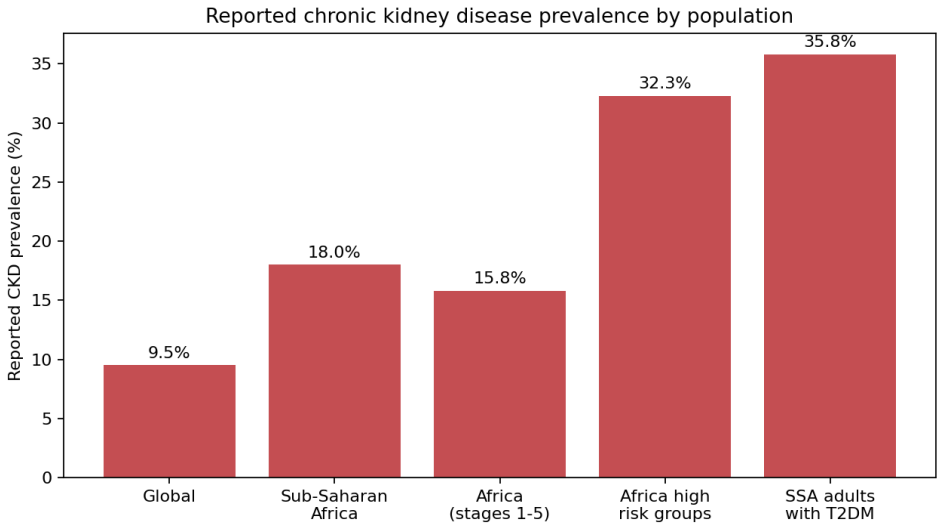


Figure 2: Reported chronic kidney disease prevalence by population

Cost closes off the treatment route for most patients. A systematic review of the economic burden of end-stage renal disease in Africa, standardised to 2025 United States dollars, reported a median annual cost of haemodialysis in Nigeria of 21,052 dollars. The cost of transplantation in the first year plus follow-up was 46,293 dollars, with most of the expenditure paid out of pocket [9]. Compared with average household income, these figures put renal replacement therapy out of reach for most affected families.

1.5 Problem Statement

Chronic kidney disease mortality in resource-limited settings is primarily caused by late presentation and not failure of treatment. Diagnosis typically relies on nephrology

referral and specialist interpretation of laboratory panels, which are lacking outside tertiary centres. Risk markers, blood pressure, blood sugar, serum creatinine, haemoglobin, albumin, and body mass index are already measured during routine clinic visits and remain unused. Machine learning promises to provide a path to obtain that signal at effectively zero marginal cost.

Existing literature on chronic kidney disease prediction has tended to focus on a common benchmark comprising 400 records, with accuracy reported near 1.0 [6] [7]. A result like that indicates a benchmark, not clinical readiness; it also gives a health manager no indication of the cost of running the model or how it will perform if the entire lab panel is not available. Only a handful of studies report minority-class performance after balancing the data, and even fewer assess whether performance holds after the expensive inputs are eliminated.

Logistic regression is well suited for this problem, not only for its accuracy, but for other reasons. The algorithm gives the clinician a coefficient for each predictor, so the clinician can see which variable drove the prediction and by how much. Many ensemble and deep-learning models achieve higher statistics on benchmark data, but their outputs are not amenable to this kind of inspection. Logistic regression also meets the requirement that the model run on modest hardware without a graphics processor, as a model developed for a primary health centre would require.

1.6 Aim and Objectives

This study builds, optimises and stress tests a logistic regression model for early chronic kidney disease detection. The specific objectives are to explore and preprocess the benchmark dataset with explicit handling of missing values and class imbalance, to develop a binary logistic regression classifier, to evaluate it through a confusion matrix and the associated metrics, to optimise it through grid search over the penalty term and the inverse regularisation strength, and to test whether performance survives a reduction of the input panel to eight inexpensive predictors across repeated random splits.

1.7 Machine Learning in Chronic Kidney Disease Prediction

Machine learning concepts come from pattern recognition and enable a system to do a task even if it is not explicitly programmed for it. Mitchell's working definition is: if a computer program's performance on a class of tasks, measured by some performance measure, improves with experience E with respect to that class of tasks, then the program is learning from experience E with respect to that class of tasks. Applied to clinical data, it looks for structure that supports diagnosis, prognosis, and triage in large, heterogeneous records.

Chronic kidney disease prediction has proceeded uniformly. Almasoud and Ward tested several classifiers on a benchmark dataset and found that gradient boosting achieved high accuracy with fewer features [1]. Chittora et al. tested seven classifiers, including artificial neural networks, C5.0, logistic regression, and support vector machines with L1 and L2 penalties, using feature selection before training [5]. Qin et al. addressed heavy missingness in the same dataset and proposed an integrated model combining logistic regression and random forest [26]. A later framework study using random subspace ensembles on the identical 400 records reported an accuracy of 1.00 [6].

This literature has three issues. First, the benchmark is small and very close to linearly separable, meaning the high accuracy reflects the dataset rather than the clinical value. Second, some studies resort to "resampling", which means that synthetic minority instances are added to the test partition, boosting all metrics. Third, virtually no one

describes what happens when expensive lab tests are out of reach, as in settings with the highest disease burden.

In the present study, we take the same algorithm and benchmark as in the literature above and reverse the evaluation. However, it does not ask how much accuracy increases, but rather by how much accuracy decreases when the model is reduced to predictors a district clinic already records, and how robust the answer is to repeated random partitions.

2. MATERIALS AND METHODS

The prediction model intended to establish the probability of chronic kidney disease is presented in this section. Model building was conducted in Python using pandas, scikit-learn, imbalanced-learn, matplotlib and seaborn. Figure 3 illustrates the working of the proposed system.

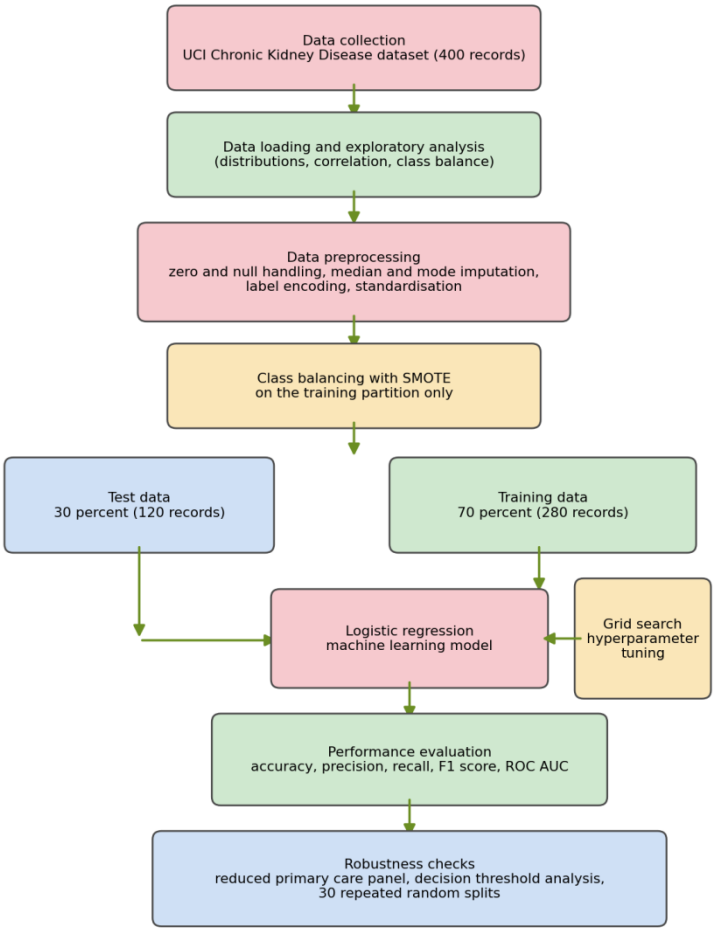


Figure 3: Working of the proposed system

2.1 Dataset Description

The data comes from the Chronic Kidney Disease dataset from the UCI Machine Learning Repository, provided by Rubini, Soundararajan, and Eswaran, and licensed under Creative Commons Attribution (CC BY) [27]. It contains records for 400 patients collected over two months in a hospital setting, 24 independent variables, and one dependent outcome variable. The dataset includes 11 numeric and 13 nominal predictors. Table 1 lists the attributes and meanings.

Table 1: Attributes of the chronic kidney disease dataset

Attribute	Type	Inference
age	Numeric	Age of the patient in years
bp	Numeric	Diastolic blood pressure in mm/Hg
sg	Numeric	Specific gravity of urine
al	Numeric	Albumin level graded 0 to 5
su	Numeric	Sugar level graded 0 to 5
rbc	Nominal	Red blood cells, normal or abnormal
pc	Nominal	Pus cells, normal or abnormal
pcc	Nominal	Pus cell clumps, present or not present
ba	Nominal	Bacteria, present or not present
bgr	Numeric	Random blood glucose in mgs/dl
bu	Numeric	Blood urea in mgs/dl
sc	Numeric	Serum creatinine in mgs/dl
sod	Numeric	Sodium in mEq/L
pot	Numeric	Potassium in mEq/L
hemo	Numeric	Haemoglobin in gms
pcv	Numeric	Packed cell volume
wc	Numeric	White blood cell count in cells/cumm
rc	Numeric	Red blood cell count in millions/cmm
htn	Nominal	Hypertension, yes or no
dm	Nominal	Diabetes mellitus, yes or no
cad	Nominal	Coronary artery disease, yes or no
appet	Nominal	Appetite, good or poor
pe	Nominal	Pedal oedema, yes or no
ane	Nominal	Anaemia, yes or no
classification	Nominal	Outcome variable, ckd or notckd

Research Journal of Humanities and Social Sciences
(An Open Access International Peer Reviewed Journal)
ISSN: 2945-3968

2.2 Data Exploration

Data exploration checks the distributions of each variable, the imbalance of the outcome classes, and relationships among features. The dataset contains 250 records with chronic kidney disease (CKD) and 150 records with non-chronic kidney disease (non-CKD). Figure 4 shows the class counts, and Figure 5 shows the correlation matrix. Positive associations are shown by lighter, warmer cells, and negative associations by darker, cooler cells.

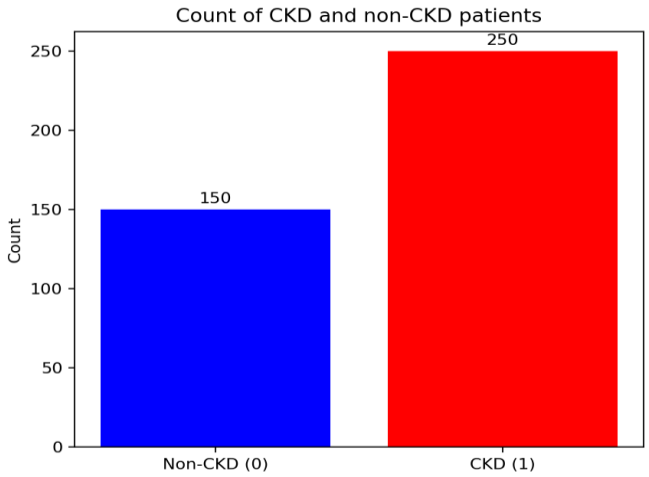


Figure 4: Count of patients with and without chronic kidney disease

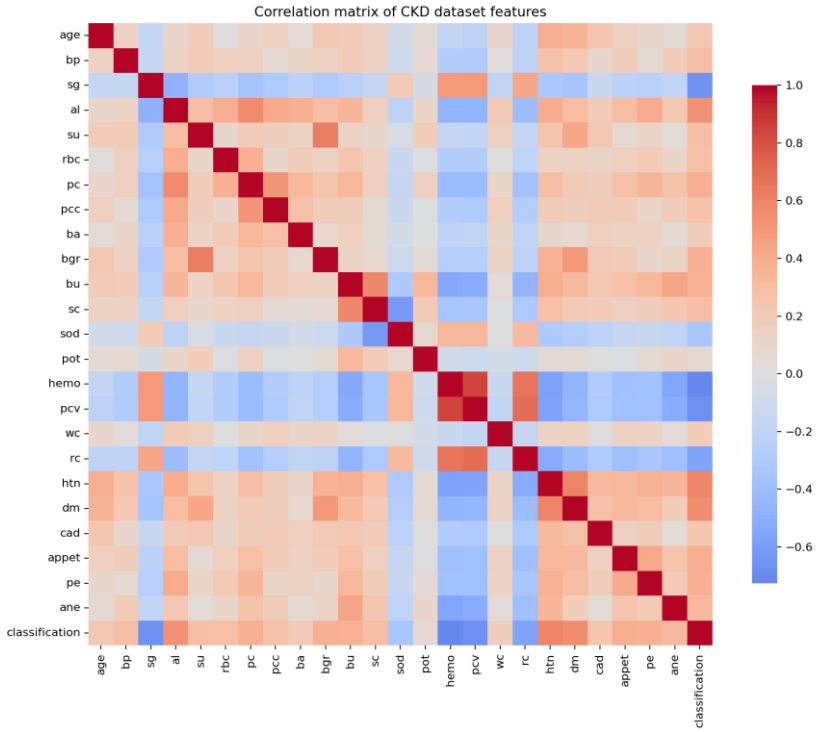


Figure 5: Correlation matrix of the dataset features

Haemoglobin (-0.726), packed cell volume (-0.673), specific gravity (-0.660), hypertension (0.590), red blood cell count (-0.566), diabetes mellitus (0.559), and albumin (0.532) are the most significant predictors of the outcome. Negative values are expected because renal function decreases along with decreases in haemoglobin, packed cell volume, and urine specific gravity. The distribution of nine selected attributes is shown in Figure 6. Before

model fitting, the data for serum creatinine, blood urea, and random blood glucose were standardised because they are strongly right-skewed.

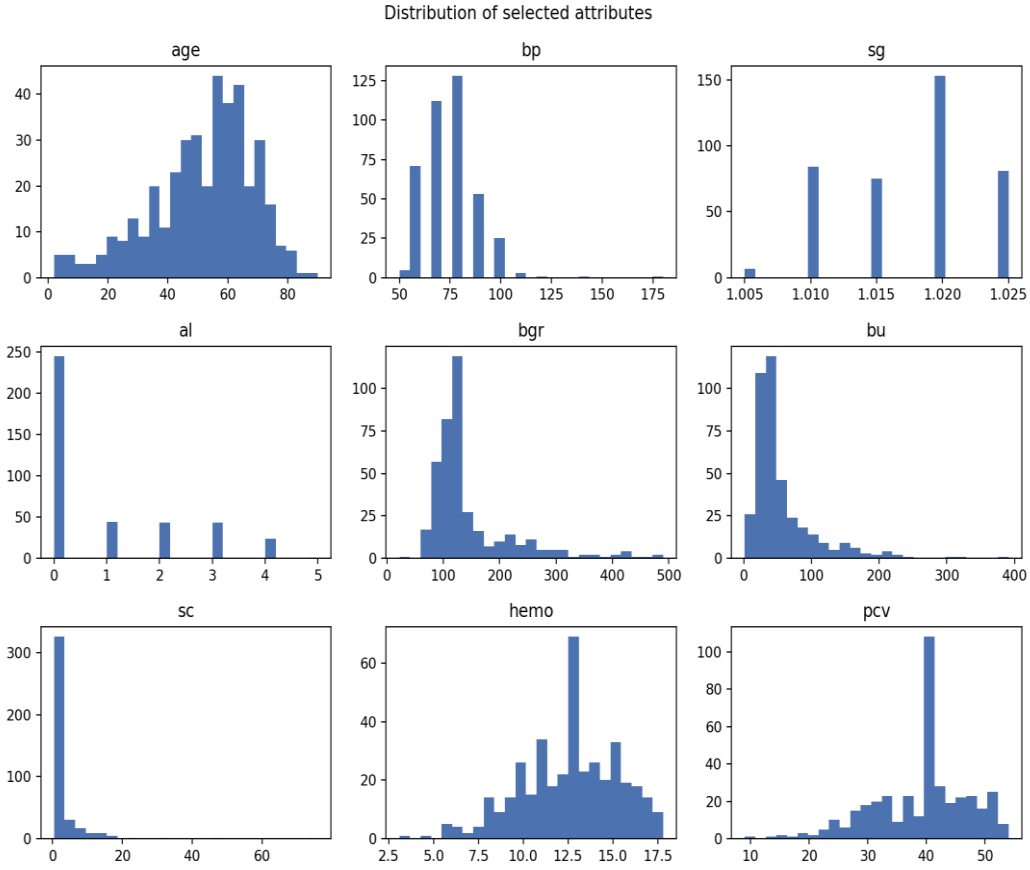


Figure 6: Distribution of selected attributes

2.3 Data Preprocessing

Preprocessing: standardising raw records into a format suitable for training the model and handling missing values, inconsistencies, noise, and irrelevant fields. The following operations were performed here: three.

Cleaning of entry artefacts. The raw file includes tab characters and extra whitespace in a few purported data fields, meaning the same clinical value can appear in multiple strings. We normalised values like yes and diabetes status with a leading tab before encoding. The same defect occurred on the outcome label. Records are silently lost if this is not corrected.

Handling missing values. The dataset has 1,012 NA (missing) values. Table 2 lists the worst-affected attributes. We used the column median (for numeric attributes) and the column mode (for nominal attributes) to impute missing values in attributes affected by the extreme values shown in Figure 6. Figures 7 and 8 show the number of missing counts before imputation and that no attribute has missing counts after imputation.

Table 2: Attributes with the highest proportion of missing entries

Attribute	Missing records	Percentage of dataset
rbc	152	38.00

Attribute	Missing records	Percentage of dataset
rc	131	32.75
wc	106	26.50
pot	88	22.00
sod	87	21.75
pcv	71	17.75
pc	65	16.25
hemo	52	13.00

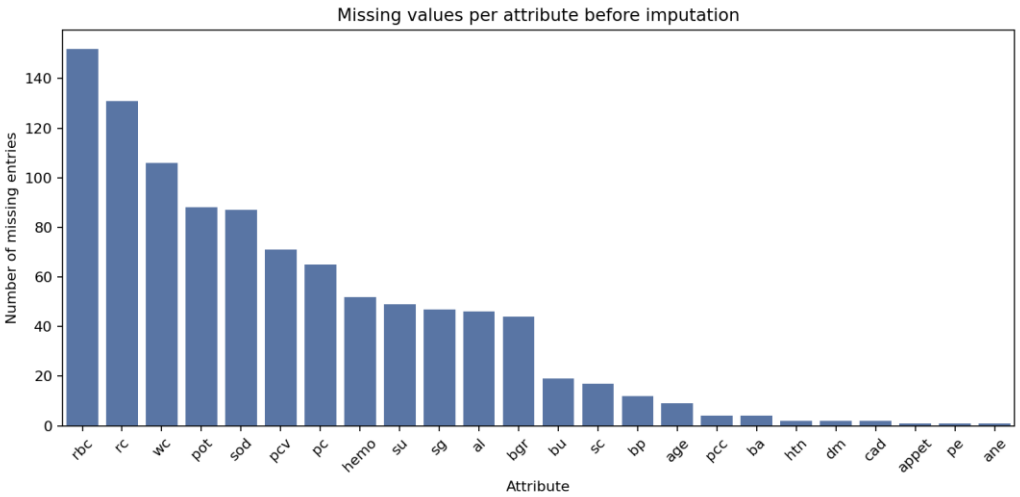


Figure 7: Missing values per attribute before imputation

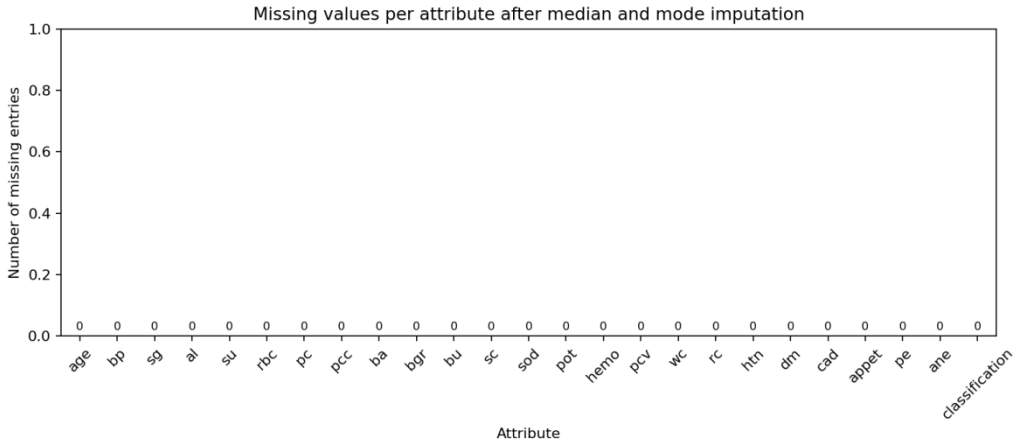


Figure 8: Missing values per attribute after median and mode imputation

Encoding and scaling. Nominal attributes were label-encoded as follows: normal, not present, no, and good were assigned the numeric value of 0, and abnormal, present, yes, and poor were assigned the numeric value of 1. The result was coded as '1' if the disease was chronic kidney disease and as '0' if it was not. Predictors were then scaled to zero mean and unit variance with statistics calculated on the training partition only, meaning that no information from the test partition was used to calculate the scaled predictors.

Class balancing. The result is quite lopsided: 62.50 percent to 37.50 percent. To correct this, we used the Synthetic Minority Oversampling Technique (SMOTE). In contrast, the

technique creates synthetic minority instances along the line segments connecting a minority record to its k nearest minority neighbours; this prevents duplication and overfitting, which can occur with random oversampling. We split the training partition and resampled only that split. This corrects typical issues in several published studies on this data set, where synthetic instances generated from test records are introduced to the training set, and every reported measure is inflated as a result. Table 3 shows the number of classes pre- and post-resampling.

Table 3: Class distribution before and after resampling

Partition	Non-CKD	CKD	Total
Full dataset	150	250	400
Training partition before SMOTE	105	175	280
Training partition after SMOTE	175	175	350
Test partition	45	75	120

2.4 Model Development Using Logistic Regression

Logistic regression can model the relationship between a categorical response variable and a set of covariates. It is a linear model appropriate for classification and is known in the literature as logit regression or maximum entropy classification. In the binary case, the logistic function models the probabilities of the possible outcomes of a single trial. Suppose that the binary responses Y1, ..., Yn are independent, with Yi = 1 if the characteristic is present and Yi = 0 if it is not. Suppose these covariates are represented by x = (x1, x2, ..., xp), where x can be discrete, continuous, or a combination of both. The probability of success is denoted by πi, and the logistic function is given in equations 1 and 2.

$$\text{logit}(\pi i) = \log[\pi i / (1 - \pi i)] = \beta 0 + \beta 1 xi1 + \beta 2 xi2 + ... + \beta pxip \quad (1)$$

$$\pi i = \exp(\beta 0 + \beta 1 xi1 + ... + \beta pxip) / [1 + \exp(\beta 0 + \beta 1 xi1 + ... + \beta pxip)] = \Lambda(\beta' xi) \quad (2)$$

These equations express the probability πi of the record being in a particular category of the dichotomous response variable as lying between 0 and 1, with Λ denoting the logistic cumulative distribution function and the β values being the parameters to be estimated as a vector. The "odds ratio" is $\pi i/(1-\pi i)$. For this study, Binary logistic regression was used since the target takes the values 0,1 (0 means that the person does not have Chronic Kidney Disease and 1 means that the person has Chronic Kidney Disease).

2.5 Training and Testing

The data were split 70:30, with 280 records used for training and 120 for testing. The split was stratified on the outcome, which maintains the class proportion in both parts of the split and avoids an unrepresentative test set from skewing the reported metrics. We repeat the experiment with a fixed random seed to ensure reproducibility.

2.6 Model Evaluation

A confusion matrix summarises a classification model's performance on a test set. The number of true positives, false positives, true negatives, and false negatives are respectively defined as the number of cases correctly predicted as positive, the number of cases predicted as positive when actually they were negative, the number of cases predicted as negative when actually they were negative, and the number of cases

predicted as negative when actually they were positive. In this study, a false negative is a patient with chronic kidney disease sent home without follow-up, which is the costliest error in the clinical setting.

Accuracy is the number of correct predictions divided by the total number of records.
Accuracy = (TP + TN) / (TP + TN + FP + FN) (3)
Precision is the number of correct positive predictions divided by the total number of positive predictions.
Precision = TP / (TP + FP) (4)
Recall, also called sensitivity or the true positive rate, is the number of correct positive predictions divided by the total number of actual positives.
Recall = TP / (TP + FN) (5)
The F1 score is the harmonic mean of precision and recall and balances the two.
F1 = 2 × (Precision × Recall) / (Precision + Recall) (6)

The area under the receiver operating characteristic curve was also computed. It measures the probability that the model assigns a higher score to a randomly chosen positive case than to a randomly chosen negative case, and it is independent of the decision threshold.

2.7 Hyperparameter Tuning

Hyperparameter Tuning is finding the right set of model arguments that are not learned during training. The search method used was grid search, which tries all combinations within a grid and chooses the combination with the highest cross-validation score, searching two hyperparameters. The penalty used either L1 or L2 regularisation, and C, the inverse of the regularisation strength, was searched across thirty logarithmically spaced points between 0.001 and 100. We used five-fold stratified cross-validation during selection, with accuracy as the criterion and a balanced training partition. We then used ten-fold stratified cross-validation to validate the fitted model separately.

2.8 Robustness Checks

We conducted three additional experiments because a single hold-out split on 120 records is not enough to draw a definite conclusion. To separate the contribution of the Synthetic Minority Oversampling Technique, we first fitted a control model without any resampling. Second, we created a reduced model with 8 predictors that can be easily recorded in a district clinic at low cost: age, blood pressure, haemoglobin, random blood glucose, hypertension, diabetes mellitus, albumin, serum creatinine. We used the same grid search procedure to tune the reduced model. Third, we retrained and retested both the full and reduced models using 30 independent train-test splits, and collected the mean and standard deviation of accuracy, F1 score, and recall for the chronic kidney disease class. A reduced model was also used to perform a decision-threshold sweep to quantify the trade-off between missed cases and false alarms.

3. RESULTS

3.1 Loading and Inspecting the Data

The file was loaded into Python as a comma-separated file. After dropping the index field (24 independent variables plus outcome), it returned 400 patient records in 25 columns. Review of the first several rows showed that a few values were blank and that three attributes stored numeric measurements as alphabetic values because of tab stops. Unparseable values in packed cell volume, white blood cell count and red blood cell count were coded as missing. After imputation, all 400 records were kept for the modelling.

3.2 Base Model Performance

The library default regularisation was applied to the base logistic regression model, which was fit on the balanced training partition. Table 4 shows the confusion matrix for the 120 held-out records.

Table 4: Confusion matrix for the base logistic regression model

	Predicted non-CKD	Predicted CKD
Actual non-CKD	45	0
Actual CKD	2	73

The model correctly classified 118 of 120 records. The model identified all non-cases and 73/75 chronic kidney disease cases. The two errors were false negatives (patients who were actually positive were predicted as negative). Table 5 shows the metrics per class.

Table 5: Per class performance of the base model

Class	Precision	Recall	F1 score
Non-CKD	0.957	1.000	0.978
CKD	1.000	0.973	0.986
Weighted average	0.984	0.983	0.983

Weighted accuracy reached 98.33 percent, precision 98.40 percent, recall 98.33 percent and the F1 score 98.34 percent. The area under the receiver operating characteristic curve reached 100.00 percent, which means that no non case received a higher predicted probability than any case at any threshold. Figure 9 presents the confusion matrices and Figure 10 the receiver operating characteristic curves.

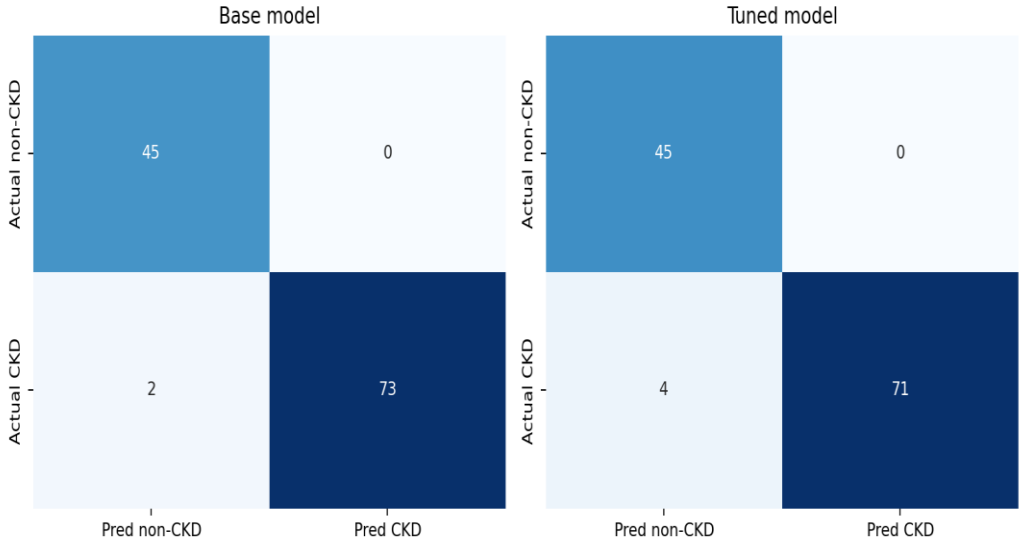


Figure 9: Confusion matrices for the base and tuned models

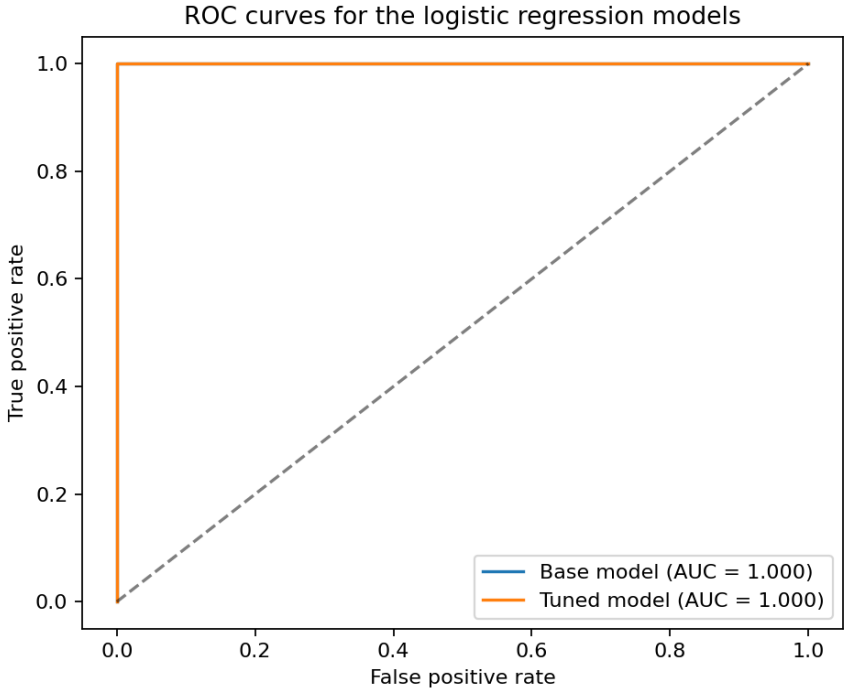


Figure 10: Receiver operating characteristic curves

3.3 Hyperparameter Tuning Results

The grid search checked 60 combinations with 5 folds. It selected an L2 penalty with C = 2.8072 and achieved a mean cross-validation accuracy of 99.14 per cent on the balanced training partition. The selected model achieved 99.14 percent accuracy (SD = 1.83) from ten-fold stratified cross-validation.

The tuned model achieved 96.67 percent accuracy, 96.94 percent precision, 96.67 percent recall and an F1 score of 96.69 percent on the held-out test partition, compared to the untuned base model, which achieved 98.33 percent accuracy. Tuning therefore did not benefit unseen data and cost 1 and two thirds of a percentage point, equivalent to two more false negatives in 120 records. Resampling was done on the fitted control model, and the figures were the same as the original base model, at an accuracy of 98.33 percent.

This is a core conclusion, not a letdown. This cross-validation accuracy of > 99 percent with a standard deviation of 1.83 is near the limit of what this dataset can achieve, and the difference between configurations is just sampling noise from a test set of 120 records, so the grid search had no room to play with. Studies published that report gains of several percentage points by tuning on this benchmark report the same sort of variation. Figure 11 compares the four fitted models.

Research Journal of Humanities and Social Sciences
(An Open Access International Peer Reviewed Journal)
ISSN: 2945-3968

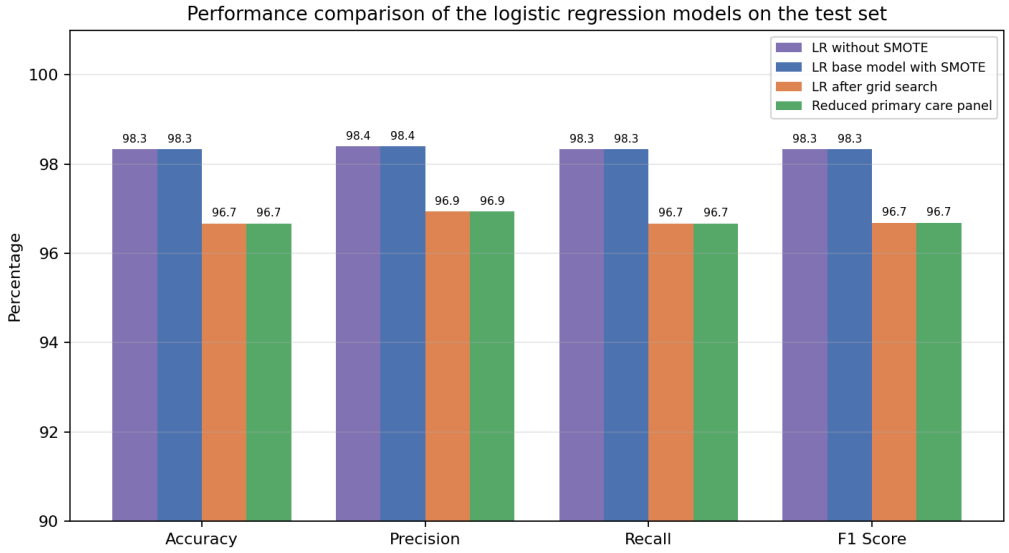


Figure 11: Performance comparison of the fitted logistic regression models

3.4 Feature Importance

The fitted coefficients are directly comparable in magnitude since the predictors were standardised. Table 6 shows the top ten. Haemoglobin has the heaviest weight (- 2.302), followed by specific gravity (-2.143) and packed cell volume (-1.641). Positive weights include hypertension, serum creatinine, appetite, and diabetes mellitus. These signs align with the clinical reasoning, as anaemia & low P.C.V. occur with reduced E.P.O. production, and high serum creatinine with reduced filtration. Figure 12 shows the same information graphically.

Table 6: Ten strongest standardised coefficients in the tuned model

Attribute	Coefficient	Direction of association
hemo	-2.302	Lower haemoglobin raises predicted risk
sg	-2.143	Lower urine specific gravity raises risk
pcv	-1.641	Lower packed cell volume raises risk
htn	1.303	Hypertension raises risk
sc	1.265	Higher serum creatinine raises risk
appet	1.240	Poor appetite raises risk
dm	1.237	Diabetes mellitus raises risk
su	0.882	Higher urine sugar raises risk
al	0.860	Higher albumin grade raises risk
cad	-0.854	Adjusted association in this sample

Research Journal of Humanities and Social Sciences
(An Open Access International Peer Reviewed Journal)
ISSN: 2945-3968

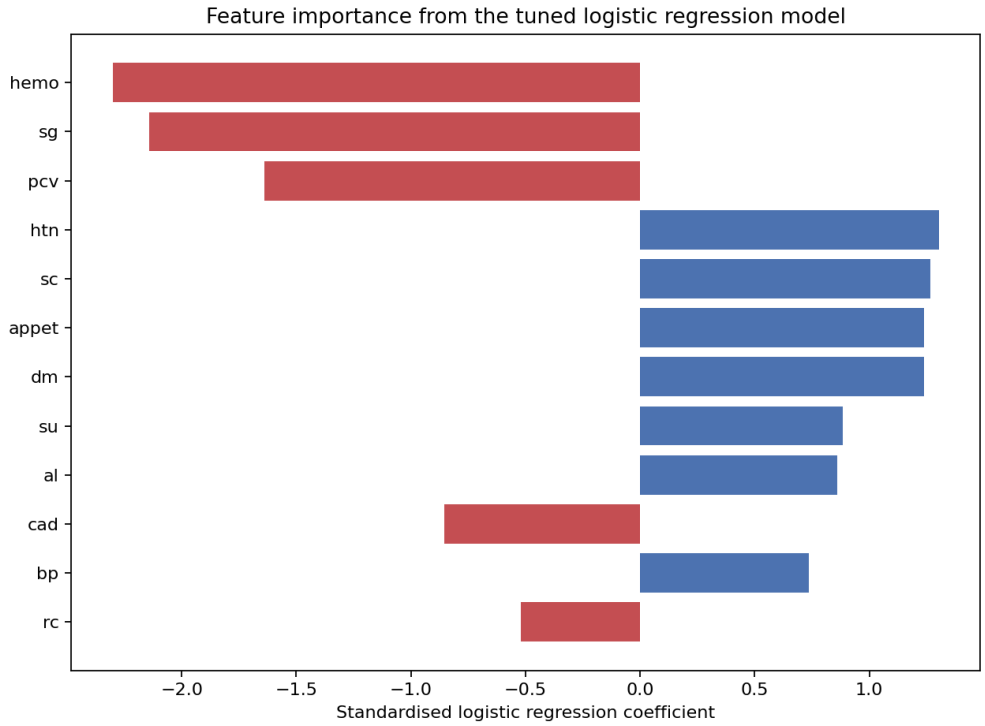


Figure 12: Feature importance from the tuned logistic regression model

3.5 The Reduced Primary Care Panel

The full model includes sodium, potassium, packed cell volume, white blood cell count, red blood cell count and microscopy for pus cells and bacteria - 24 items. Most primary health centres in the highest-burden settings do not have that panel. The reduced model used random blood glucose, hypertension, diabetes mellitus, age, haemoglobin, albumin, serum creatinine, and blood pressure as predictors. We selected the L1 penalty (C = 2.8072) using grid search.

The held-out partition of the reduced model matched the tuned full model exactly, with four false negatives, an F1 score of 96.69 percent, and an accuracy of 96.67 percent. On 30 replications of the random splits, reported in Table 7, the full model achieved 98.53 percent correct with a standard deviation of 1.36, and the reduced model achieved 97.11 percent correct with a standard deviation of 1.23. The cost of a 2/3 reduction in the laboratory panel is thus about a 1.4 percentage-point drop in accuracy and a 1.2 percentage-point drop in recall for disease class. The comparison with error bars is shown in Figure 13.

Table 7: Stability of the two models across 30 random train and test splits

Metric	Full 24 predictor model	Reduced 8 predictor panel	Difference
Accuracy	98.53 ± 1.36	97.11 ± 1.23	1.42
F1 score	98.80 ± 1.12	97.66 ± 0.99	1.14
Recall on CKD class	97.64 ± 2.17	96.40 ± 1.69	1.24

Research Journal of Humanities and Social Sciences
(An Open Access International Peer Reviewed Journal)
ISSN: 2945-3968

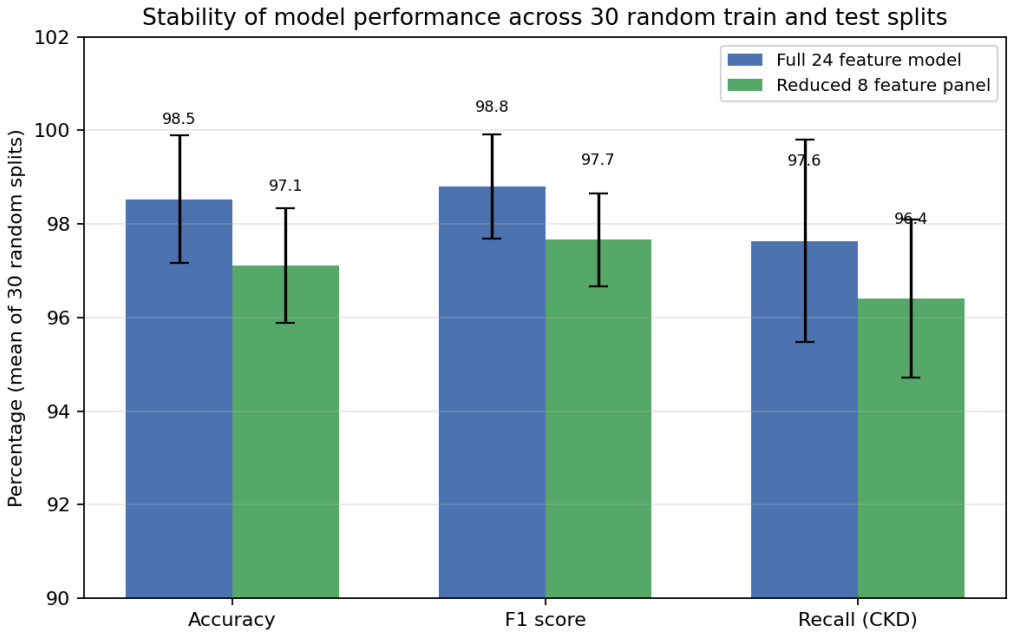


Figure 13: Stability of model performance across 30 random splits

3.6 Decision Threshold Analysis

A classifier gives a probability, and the default cut-off of 0.50 is not clinically significant. A screening application will have a false negative that is more expensive than a false alarm, because a false negative will be undiagnosed and a false alarm will result in one confirmatory test. Table 8 will sweep the threshold for the reduced model.

Table 8: Effect of the decision threshold on the reduced panel model

Threshold	Accuracy	Precision	Recall	False negatives	False positives
0.30	95.83	97.30	96.00	3	2
0.40	95.83	97.30	96.00	3	2
0.50	96.67	100.00	94.67	4	0
0.60	96.67	100.00	94.67	4	0
0.70	95.83	100.00	93.33	5	0

Reducing the threshold from 0.50 to 0.30 recovers one case at the expense of two false alarms and a 0.84 percentage-point loss in headline accuracy. To get a screening that trade is worth taking. This is the real-life choice a health manager has to make, and it is not represented in any paper that reports only accuracy.

4. DISCUSSION

These results yield three findings. First, the benchmark is fully optimised. The default linear model can achieve 98.33 percent accuracy and 100 percent area under the curve, the best possible score. Results from tuning or ensemble architecture on this data set are within the sampling error of a 120-record test partition, and an accuracy of 1.00 on 400 separable records is more a property of this benchmark data set than a sign of clinical readiness.

Second, resampling made no difference here. The control model (with no Synthetic Minority Oversampling Technique) was the same as the balanced model. This class ratio of 62.5:37.5 is not a serious imbalance, and resampling matters more at a much more extreme ratio. It is still good practice to apply it, but applying it before the train-test split, as several published studies on this dataset do, overestimates all reported metrics.

Third, useful leverage does not depend on model complexity; it depends on input costs. The mean accuracy of eight predictors already gathered during routine outpatient care (97.11 percent) is comparable to the mean accuracy of the full 24-item panel (98.53 percent). It's a better trade-off for a district facility to sacrifice 1.4 percentage points and screen all patients than to choose a few for whom it cannot perform sodium, potassium, and full microscopy.

The restrictions must be clearly stated. The dataset consists of 400 data points collected at one centre over 2 months, and the population from which it was sampled is unknown. That performance cannot be automatically translated to a Nigerian or pan-African clinic population, where the aetiological mix differs, and glomerulonephritis and causes of unknown origin play a bigger role. Additionally, there is no monitoring data, as the model predicts prevalent disease. However, before any deployment claim, external validation on locally collected records is still required.

4.1 Conclusion and Future Research

In this study, we applied logistic regression to detect chronic kidney disease on the UCI benchmark. We tested it with checks that are less emphasised in the existing literature on this dataset. The base model achieved 98.33 percent accuracy, 98.40 percent precision, 98.33 percent recall, an F1 score of 98.34 percent and an area under the receiver operating characteristic curve of 100.00 percent. We performed a grid search over the inverse regularisation strength and the penalty term and selected an L2 penalty with C=2.8072; however, no improvement was achieved on the held-out data, so the benchmark cannot be further optimised in this way.

This contribution is based on the smaller panel. The mean accuracy was 97.11 percent across 30 random splits of the eight predictors a primary health centre already records, compared with 98.53 percent for the full 24-item panel. Moving the decision point to 0.30 will save one case at the expense of two false alarms and therefore favour screening. Further research is needed to test the reduced panel on prospective Nigerian clinic records and to evaluate the performance of the reduced logistic model compared with gradient boosting and support vector machines when both models use the reduced input set rather than the full benchmark.

Acknowledgement

The authors thank Rubini, Soundarapandian and Eswaran for making the Chronic Kidney Disease dataset available through the UCI Machine Learning Repository [27].

Funding: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflict of Interest: The authors declare no conflict of interest.

References

1. Almasoud M, Ward TE. Detection of chronic kidney disease using machine learning algorithms with least number of predictors. International Journal of Advanced Computer Science and Applications, 2019;10(8).

2. Bhatta RP, Air DK. Diabetes prediction using logistic regression machine learning algorithm. Research Journal of Multidisciplinary Engineering Technologies, 2025;4(6):1 to 14. doi:10.58924/rjmet.v4.iss6.p1
3. Burden of chronic kidney disease and its attributable risk factors in 204 countries and territories, 1990 to 2021: results from the Global Burden of Disease Study 2021. Medicine, 2025. doi:10.1097/MD.0000000000050111
4. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002;16:321 to 357.
5. Chittora P, Chaurasia S, Chakrabarti P, et al. Prediction of chronic kidney disease: a machine learning perspective. IEEE Access, 2021;9:17312 to 17334. doi:10.1109/ACCESS.2021.3053763
6. Chronic kidney disease detection using machine learning: from analysis to framework development. Biomedical and Pharmacology Journal, 2024;17(3).
7. Chronic kidney disease prediction based on machine learning algorithms. Journal of Pathology Informatics, 2023.
8. Chukwuonye II, Ogah OS, Anyabolu EN, et al. Prevalence of chronic kidney disease in Nigeria: systematic review of population-based studies. International Journal of Nephrology and Renovascular Disease, 2018;11:165 to 172.
9. Economic cost of end stage renal disease in Africa: a systematic review. BMC Nephrology, 2025. doi:10.1186/s12882-025-04478-5
10. Erlin E, Marlim YN, Junadhi, Suryati L, Agustina N. Early detection of diabetes using machine learning with logistic regression algorithm. Jurnal Nasional Teknik Elektro dan Teknologi Informasi, 2022;11(2):88 to 96.
11. Evwiekpaefe AE, Abdulkadir N. A predictive model for diabetes mellitus using machine learning techniques: a study in Nigeria. The African Journal of Information Systems, 2023;15(1).
12. George C, Stoker S, Okpechi I, Woodward M, Kengne AP. The chronic kidney disease in Africa collaboration: lessons from a new pan-African network. BMJ Global Health, 2021;6(8):e006454.
13. Global burden of chronic kidney disease and risk factors, 1990 to 2021: an update from the Global Burden of Disease Study 2021. Frontiers in Public Health, 2025.
14. Hasan MK, Alam MA, Das D, Hossain E, Hasan M. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 2020;8:76516 to 76531.
15. Islam MA, Akter S. Risk Factor Prediction of Chronic Kidney Disease [dataset]. UCI Machine Learning Repository, 2020. doi:10.24432/C5WP64
16. Kaze AD, Ilori T, Jaar BG, Echouffo-Tcheugui JB. Burden of chronic kidney disease on the African continent: a systematic review and meta-analysis. BMC Nephrology, 2018;19:125. doi:10.1186/s12882-018-0930-5
17. Lai H, Huang H, Keshavjee K, Guergachi A, Gao X. Predictive models for diabetes mellitus using machine learning techniques. BMC Endocrine Disorders, 2019;19:101. doi:10.1186/s12902-019-0436-6
18. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 2017;18(17):1 to 5.

19. Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. Journal of Global Health, 2018;8(2):020303. doi:10.7189/jogh.08.020303
20. Panesar A. Machine Learning and AI for Healthcare. Apress, 2019. doi:10.1007/978-1-4842-3799-1
21. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research, 2011;12:2825 to 2830.
22. Predicting chronic kidney disease progression using small pathology datasets and explainable machine learning models. medRxiv, 2024. doi:10.1101/2024.04.08.24305414
23. Prevalence and associated factors of chronic kidney disease among adults with type 2 diabetes mellitus in sub-Saharan Africa: a systematic review and meta-analysis, 2025.
24. Prevalence of cognitive impairment and its predictors among chronic kidney disease patients: a systematic review and meta-analysis. PLOS ONE, 2024.
25. The public health impacts of chronic kidney disease in sub-Saharan Africa: a call for action. Global Medicine, 2025. doi:10.1016/j.glmedi.2025.100033
26. Qin J, Chen L, Liu Y, Liu C, Feng C, Chen B. A machine learning methodology for diagnosing chronic kidney disease. IEEE Access, 2020;8:20991 to 21002.
27. Rubini L, Soundarapandian P, Eswaran P. Chronic Kidney Disease [dataset]. UCI Machine Learning Repository, 2015. doi:10.24432/C5G020
28. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 2019;19:281. doi:10.1186/s12911-019-1004-8

***

ABOUT EMBAR PUBLISHERS

Embar Publishers is an open-access, international research based publishing house committed to providing a 'peer reviewed' platform to outstanding researchers and scientists to exhibit their findings for the furtherance of society to provoke debate and provide an educational forum. We are committed about working with the global researcher community to promote open scholarly research to the world. With the help of our academic Editors, based in institutions around the globe, we are able to focus on serving our authors while preserving robust publishing standards and editorial integrity. We are committed to continual innovation to better support the needs of our communities, ensuring the integrity of the research we publish, and championing the benefits of open research.

Our Journals

1. Research Journal of Education , linguistic and Islamic Culture - 2945-4174
2. Research Journal of Education and Advanced Literature – 2945-395X
3. Research Journal of Humanities and Cultural Studies - 2945-4077
4. Research Journal of Arts and Sports Education - 2945-4042
5. Research Journal of Multidisciplinary Engineering Technologies - 2945-4158
6. Research Journal of Economics and Business Management - 2945-3941
7. Research Journal of Multidisciplinary Engineering Technologies - 2945-4166
8. Research Journal of Health, Food and Life Sciences - 2945-414X
9. Research Journal of Agriculture and Veterinary Sciences - 2945-4336
10. Research Journal of Applied Medical Sciences - 2945-4131
11. Research Journal of Surgery - 2945-4328
12. Research Journal of Medicine and Pharmacy - 2945-431X
13. Research Journal of Physics, Mathematics and Statistics - 2945-4360