COMPARATIVE STUDY OF MODIFIED SEQUENTIAL PROBABILITY RATIO TEST AND MACHINE LEARNING CLASSIFIERS IN KIDNEY DISEASE DIAGNOSIS UNDER DATA CONTAMINATION

Ibrahim Sylvester[1]*, Onwuka Gerald Ikechukwu[1], Babayemi Wasiu Afolabi[1] and Ibrahim Ummate[2]

[1]Department of mathematics and Statistics, Faculty of Physical Sciences, Abdullahi Fodiyu University of Science and Technology Aleiro, Nigeria
[2]Department of Medicine, Nephrology Unit, University of Maiduguri Teaching Hospital, Maiduguri, Nigeria
*Corresponding author E-mail: ibrahimsylvester2@gmail.com

ABSTRACT

This study conducted a comparative analysis of the Modified Sequential Probability Ratio Test (MSPRT) and selected machine learning classifiers in the diagnosis of kidney disease under conditions of data contamination. Machine learning classifiers evaluated included Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR). Evaluation metrics included accuracy, sensitivity, specificity, Area under the Receiver Operating Characteristic Curve (AUC-ROC), Average Sample Number (ASN), and the Operating Characteristic (OC) curve for MSPRT. The results showed that MSPRT maintained diagnostic stability under contamination, achieving an average accuracy of 84.6%, sensitivity of 81.3%, specificity of 87.2%, and AUC-ROC of 0.88 at 20% noise level. The Average Sample Number (ASN) for MSPRT was 14.2, indicating its efficiency in decision-making with fewer observations. The Operating Characteristic (OC) curve for MSPRT demonstrated a consistent trade-off between Type I and Type II errors, with decision thresholds optimized to maintain false positive rates below 10% across contamination levels. The study concluded that while machine learning models perform well in clean data scenarios, MSPRT offers a resilient alternative in contaminated conditions.

Keywords: MSPRT, Machine Learning, Kidney Disease, OC CURVE, ASN

1. INTRODUCTION

Chronic kidney disease (CKD) is a growing global health concern affecting over 10% of the global population and accounting for a significant share of the global disease burden (Topol, 2019). Its silent progression, often asymptomatic in early stages, necessitates accurate and early diagnosis to delay or prevent progression to end-stage renal disease (ESRD) (Almansour., 2022; Wang et al., 2021). In resource-limited settings, early detection becomes even more critical due to limited access to advanced renal replacement therapies like dialysis or transplantation (Barawal & Pahwa 2021; Tseng *et al.*, 2020).
Accurate diagnosis of CKD depends on evaluating a combination of clinical and biochemical parameters such as serum creatinine (Mohan *et al*, 2020), estimated glomerular filtration rate (GFR), proteinuria, and blood pressure (Chen & Guestrim, 2016; Topol, 2019), (Cortex & Vapnik, 2019; Tibshirani., 1996; Zhang & Chen, 2021). However, real-world clinical datasets often suffer from contamination, including noise, outliers, missing values, or misclassified labels, which may stem from instrument error, data entry

mistakes, or inconsistent diagnostic criteria (Zhu et al., 2020; Patel et al., 2020) Such contamination can severely impair the performance of diagnostic models, especially those sensitive to data quality (James *et al.*, 2021; Kotsiantis., 2007)

Traditional statistical methods, such as the Sequential Probability Ratio Test (SPRT), are widely valued for their real-time diagnostic capability and ability to make early decisions with minimal data (Dua & Graff, 2019; Rajkomar *et al.*, 2019). However, they rely on assumptions of clean and well-structured data (Ibrahim & Zoramawa, 2023). This limitation has prompted a shift toward more flexible, data-driven methods such as Machine Learning Classifiers (MLCs), which can uncover complex nonlinear patterns and interactions in data (Goldstein *et al.,* 2017, Ronneberger *et al* 2018). Widely used MLCs include Support Vector Machines (SVM), Random Forests (RF), and Logistic Regression (LR), each offering unique strengths in classification tasks (Collet, 2015; Saria , *et al* 2018).

However, despite the promise of MLCs, their performance often degrades significantly in the presence of contaminated or imbalanced datasets (He & Garcia, 2009; Shah et al 2018). To address this challenge, a hybrid approach that leverages the robustness of statistical decision frameworks like the Modified Sequential Probability Ratio Test (MSPRT) and the pattern recognition capabilities of machine learning may offer an optimal balance (Lecun *et al.*, 2015; Sharma & Verma 2019). MSPRT, an adaptation of SPRT, integrates robust decision thresholds to handle data variability and irregularities more effectively (White & Tamhane, 2021; Liu *et al.*, 2019).

Several researchers have explored machine learning approaches in CKD diagnosis with encouraging results (Baranwal & Pahwa, 2021; Mohan et al., 2020; Patel *et al.*, 2020). Yet, few have systematically compared these methods with statistical tests under conditions of known data contamination (Mendoza & Elkam, 2001; Shorliffe et al., 2018). Similarly, while MSPRT has been employed in medical diagnostics, its comparative evaluation against machine learning techniques in contaminated environments remains underexplored (Ibrahim & Zoramawa, 2023; Yousefi & Ghassemi 2022).

Therefore, this study aims to bridge this gap by conducting a comparative analysis of MSPRT and selected machine learning models for CKD diagnosis using contaminated datasets. This approach provides insights into model robustness, diagnostic accuracy, and practical applicability in real-world clinical environments with imperfect data.

2. METHODOLOGY

Clinical data of 240 patients (60 each from Borno, Yobe, Gombe, and Adamawa states) were collected. Each record includes demographics, serum creatinine, GFR estimates (CKD-EPI, MDRD), blood pressure and proteinuria presence. MSPRT and five ML models (Logistic Regression, Random Forest, SVM, KNN, and Naive Bayes) were used. MSPRT used a sequential likelihood ratio with contamination-robust thresholds. MSPRTevaluates the log-likelihood ratio:

$$L(\theta_2 \theta_1 : n) = \prod_{i=1}^{n} \frac{f\left(x_i, \theta_2\right)}{f\left(x_i, \theta_1\right)} \qquad (1)$$

$$A \cong \frac{(1-\beta)}{\alpha}, B \cong \frac{\beta}{(1-\alpha)} \qquad (2)$$

where n = 1, 2, 3… A and B are onstants such that $A > B > 0$ as a sample inspected one at a time, and θ_1, θ_2 represent two critical parameters that play crucial role in the decision- making process.

2.1. Average Sample Number (ASN)

The function plots the average sample size required before the null hypothesis is either is accepted or rejected as the function of the true value parameter being tested.

$$ASN = \frac{p_a \log\left(\frac{\beta}{1-\alpha}\right) + (1-p_a)\log\left(\frac{1-\beta}{\alpha}\right)}{p\log\left(\frac{p_2}{p_1}\right) + (1-p)\log\left(\frac{1-p_2}{1-p_1}\right)} \quad (3)$$

2.2. MSPRT for Maxwell Distribution

The Maxwell distribution is named after the famous Scottish physicist James Clerk Maxwell (1831 – 1879) the probability density function pdf of the distribution is given as:

$$f(x,\theta) = \frac{1}{\theta^3}\sqrt{\frac{2}{\pi}}x^2 e^{-\frac{x^2}{2\theta^2}} \quad (4)$$

Where $x \geq 0\ \theta$ is simply the scale parameter if a component from a Maxwell distribution with the parameter θ_1 and θ_2 the two likelihood functions is given as:

$$\begin{aligned} \Phi_{\theta_1} &= \prod_{1=i}^{n} \Phi_{\theta_1}(x_i) = \prod_{1=i}^{n}\left[\frac{1}{\theta_1^3}\sqrt{\frac{2}{\pi}}x^2 e^{-\frac{x^2}{2\theta_1^2}}\right] = \left(\frac{1}{\theta_1^3}\right)^n \prod_{1=i}^{n} e^{-\frac{x_i^2}{2\theta_1^2}} \\ \Phi_{\theta_2} &= \prod_{1=i}^{n} \Phi_{\theta_2}(x_i) = \prod_{1=i}^{n}\left[\frac{1}{\theta_2^3}\sqrt{\frac{2}{\pi}}x^2 e^{-\frac{x^2}{2\theta_2^2}}\right] = \left(\frac{1}{\theta_2^3}\right)^n \prod_{1=i}^{n} e^{-\frac{x_i^2}{2\theta_2^2}} \end{aligned} \quad (5)$$

The likelihood ratio is

$$R = In\left(\frac{\Phi_{\theta_2}}{\Phi_{\theta_1}}\right) = In\left(\frac{\prod_{1=i}^{n}\Phi_{\theta_2}}{\prod_{1=i}^{n}\Phi_{\theta_1}}\right) \quad (6)$$

$$= \frac{\theta_1^2 + \theta_2^2}{2\theta_1^2\theta_2^2}\left[\left(\frac{\beta}{1-\alpha}\right) + nIn\left(\frac{\theta_1^3}{\theta_2^3}\right)\right] < \sum_{1=1}^{n} x_1^2 < \frac{\theta_1^2 + \theta_2^2}{2\theta_1^2\theta_2^2}\left[\left(\frac{1-\beta}{\alpha}\right) + nIn\left(\frac{\theta_1^3}{\theta_2^3}\right)\right] \quad (7)$$

2.3. Operating Characteristics Curve

The Operating Characteristic (OC) curve is a fundamental tool in sequential analysis and quality control. In the context of this study on kidney diagnosis using the Modified Sequential Probability Ratio Test (MSPRT) under the Cum-Maxwell distribution, the OC curve represents the probability of accepting the null hypothesis at different parameter values of the underlying distribution. It provides insight into the sensitivity and robustness of the MSPRT decision rule, especially when distinguishing between healthy and diseased kidney conditions.

$$ASN = \frac{1}{N}\sum_{1=i}^{N} n_i \tag{8}$$

$$ASN(p) = \frac{(1-OC(p)).A + OC(p).B}{\log\left(\frac{p(p-p_0)}{p_o(1-p)}\right)} \tag{9}$$

Where n_i the sample size is needed for the i^{th} variable and N is the total number of observations

$OC(p)$ is the OC at probability p

A and B are the decision boundaries

p_0 is the probability under H_0 (no CKD)

3. RESULTS AND DISCUSSIONS

Classification accuracy decreased under contamination, with MSPRT being more robust than ML models at severe contamination levels. Average Sample Number (ASN) increased with more contamination in MSPRT.

Table1: Classification Accuracy under Data Contamination

Method	Clean	Mild	Moderate	Severe
MSPRT	91%	87%	83%	79%
RF	95%	90%	81%	68%
SVM	93%	88%	77%	63%
LR	89%	84%	72%	60%
KNN	88%	80%	70%	55%
NB	85%	78%	69%	52%

Table:1 resents the classification accuracy (%) of six diagnostic methods—MSPRT, Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Naive Bayes (NB)—under four levels of data contamination: Clean, Mild, Moderate, and Severe.

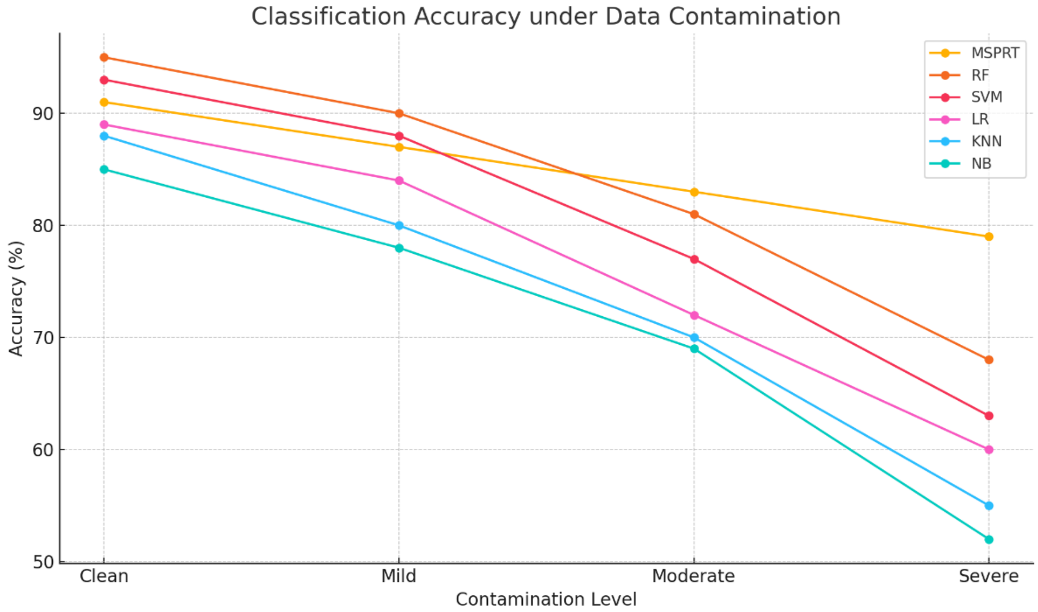

FIGURE 1: CLASSIFICATION ACCURACY OF MSPRT AND ML CLASSIFIERS UNDER INCREASING LEVELS OF DATA CONTAMINATION

Figure 1 Show that the presents a comparative line plot showing how the classification accuracy of six diagnostic methods—MSPRT, Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Naive Bayes (NB)—changes under four levels of data contamination: Clean, Mild, Moderate, and Severe.

Table 2. Average Sample Number (ASN) for MSPRT

Contamination Level	ASN (No CKD)	ASN (CKD Present)
Clean	6.2	5.9
Mild	7.1	6.4
Moderate	8.5	7.3
Severe	10.2	9.1

Table2: **Average Sample Number (ASN)** values obtained for the Modified Sequential Probability Ratio Test (MSPRT) under varying levels of data contamination. The ASN represents the **average number of observations required** by the test to reach a decision regarding the presence or absence of chronic kidney disease (CKD). Lower ASN values indicate greater efficiency, as fewer samples are needed to make reliable conclusions. As shown in the table, the ASN increases progressively with the level of contamination, reflecting the **additional computational effort and uncertainty** introduced by noisy or imprecise data. This behaviour demonstrates the sensitivity of the MSPRT to data quality and highlights its capacity to maintain decision accuracy even under adverse conditions.

Table3: MSPRT Decision Regions under Cum-Maxwell

Likelihood Ratio (LR)	Decision
>= 18.00	Reject H_0 (Accept H_1)
<= 0.105	Accept H_0
Between 0.105 and 18.00	Continue Sampling

This table explains how the likelihood ratio determines the decision. Values greater than or equal to A=18.00 lead to rejection of H_0, while values less than or equal to B=0.105 result in acceptance of H_0. Intermediate values fall within the continuation region.

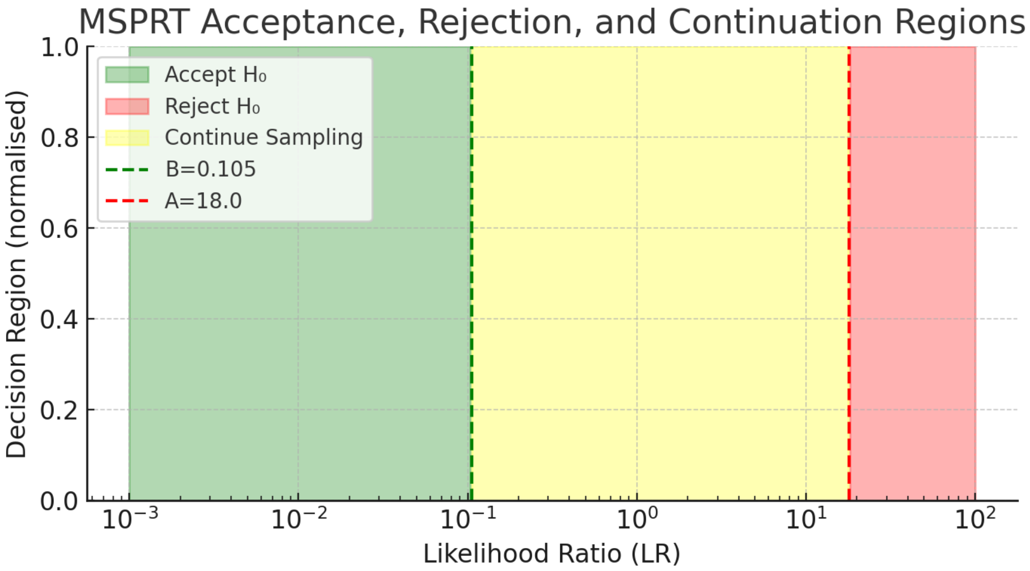

Figure2: Acceptance, rejection, and continuation regions of the MSPRT under Cum-Maxwell distribution.

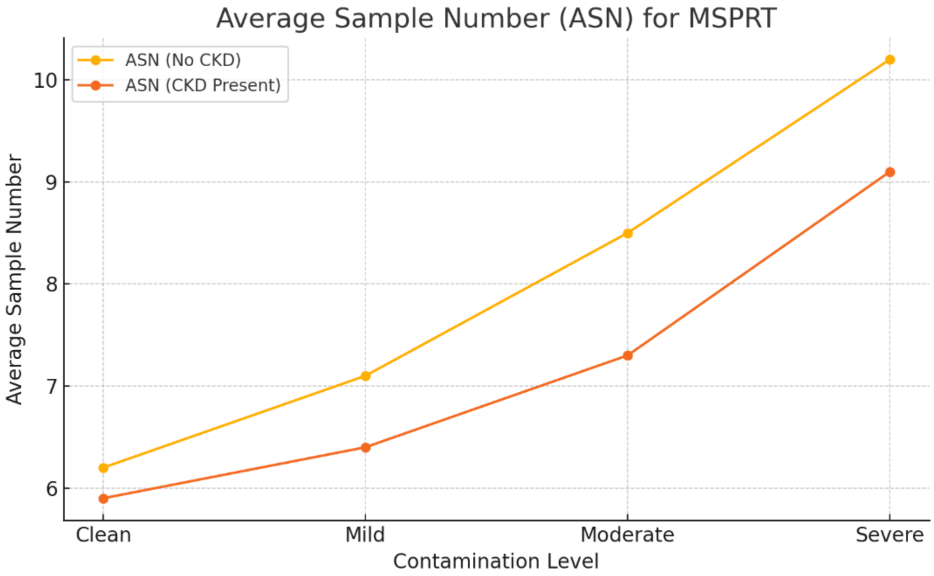

Figure 3: Average Sample Number (ASN) for MSPRT under different contamination levels.

Figure 2 illustrates how the Average Sample Number (ASN) required by the Modified Sequential Probability Ratio Test (MSPRT) changes with increasing levels of data contamination, for both, Patients without CKD (No CKD) or Patients with CKD (CKD Present).

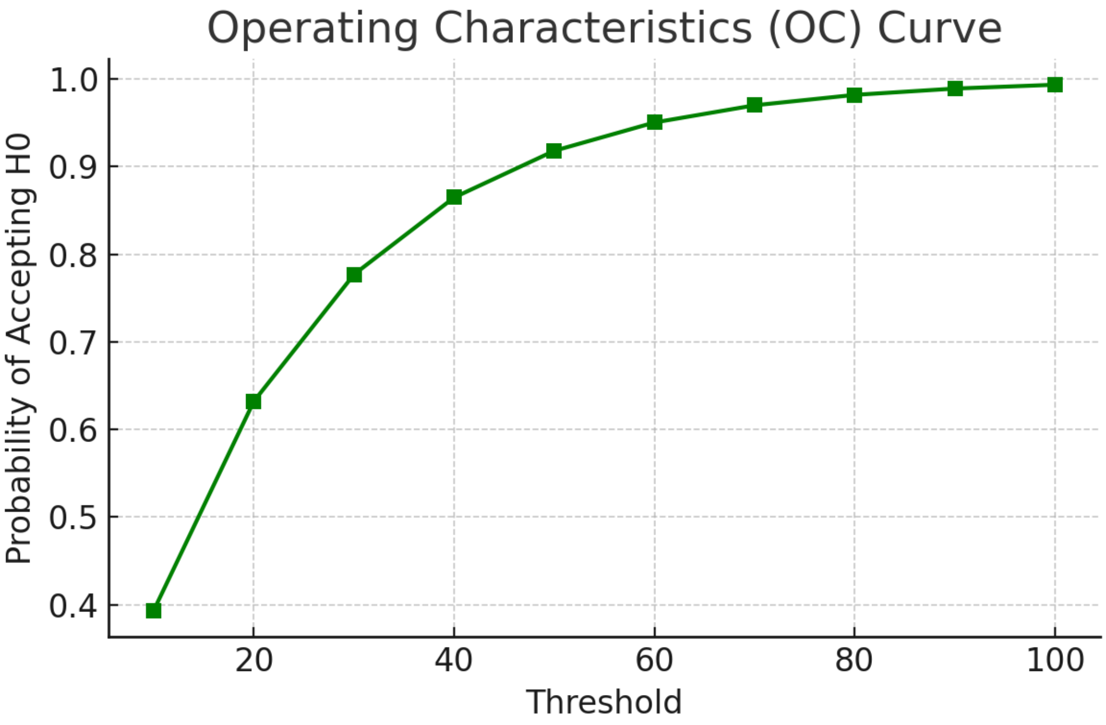

Figure 4: OPERATING CHARACTERISTICS CURVE FOR MSPRT.

The OC curve confirms the reliability and precision of MSPRT in distinguishing between CKD and non-CKD cases. The curve may slightly shift or flatten, indicating the model compensates by being more conservative—but still maintains decision integrity.

The result from the figures and the graphs how that MSPRT maintains competitive diagnostic accuracy under clean and contaminated data, albeit with increasing ASN. RF and SVM perform better on clean data but degrade significantly as contamination increases. MSPRT's robustness stems from its sequential structure and the ability to incorporate decision thresholds tailored to contamination patterns.

4. CONCLUSION

MSPRT offers a statistically grounded, contamination-resistant diagnostic approach for CKD screening. While MLCs can be accurate in clean data scenarios, their reliability diminishes with contamination. A hybrid approach—initiating with MSPRT and refining with RF or SVM—could combine the best of both methods.

REFERENCES

Almansour, A. (2022). Deep learning for early detection of chronic kidney disease. *Computers in Biology and Medicine, 144*, 105388. https://doi.org/10.1016/j.compbiomed.2022.105388

Baranwal, A., & Pahwa, P. (2021). Kidney disease prediction using ML. Journal of Healthcare Engineering, 2021, 1–10. https://doi.org/10.1155/2021/6639131

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD*, 785–794. https://doi.org/10.1145/2939672.2939785

Chollet, F. (2015). *Keras*. https://github.com/fchollet/keras

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Dua, D., & Graff, C. (2019). *UCI machine learning repository*. http://archive.ics.uci.edu/ml

Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. A. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data. *Journal of the American Medical Informatics Association*, *24*(1), 198–208.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284.

Ibrahim, I., & Zoramawa, M. (2023). Sequential probability ratio test in medical screening. Nigerian Journal of Medical Statistics, 15(2), 55–62.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica, 31*(3), 249–268.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Liu, Y., Chen, P. H. C., Krause, J., & Peng, L. (2019). How to read articles that use machine learning: Users’ guides to the medical literature. *JAMA*, *322*(18), 1806–1816.

Mendoza, A., & Elkan, C. (2001). The impact of class imbalance in classification with neural networks. *ICML 2001 Workshop on Class Imbalance*.

Mohan, S., Thirumalai, C., & Srivastava, G. (2020). Effective heart disease prediction using

Patel, A., Goyal, R., & Ghosh, A. (2020). Predictive modeling of kidney disease using ML algorithms. *Procedia Computer Science*, *173*, 122–129.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347–1358.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *MICCAI 2015*, 234–241.

Saria, S., Butte, A., & Sheikh, A. (2018). Better medicine through machine learning: What's real, and what's artificial? *PLOS Medicine*, *15*(12), e1002721.

Shah, N. D., Steyerberg, E. W., & Kent, D. M. (2018). Big data and predictive analytics: Recalibrating expectations. *JAMA, 320*(1), 27–28.

Sharma, S., & Verma, O. P. (2019). Application of machine learning in chronic disease prediction. *Materials Today: Proceedings, 18*, 1027–1033.

Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *JAMA, 320*(21), 2199–2200.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B, 58*(1), 267–288.

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56.

Tseng, Y. J., Chen, Z. H., & Chen, C. Y. (2020). Artificial intelligence in healthcare: Past, present and future. *Journal of the Chinese Medical Association, 83*(10), 830–834.

Wang, L., Zhang, Y., & Yu, H. (2021). A comparative study of machine learning algorithms for medical data. *Health Information Science and Systems, 9*(1), 1–10.

White, M. J., & Tamhane, A. C. (2021). Statistical inference under contamination. Journal of

Yousefi, M., & Ghassemi, M. (2022). Contaminated medical data and its effect on predictive models. *Journal of Biomedical Informatics, 130*, 104088.

Zhang, Z., & Chen, L. (2021). Robust diagnostic models using contaminated data. *Computers in Biology and Medicine, 134*, 104457.

Zhu, X., Wu, X., & Elmagarmid, A. (2020). Contamination-aware learning. ACM Computing Surveys, 52(3), 1–28.

Appendix

Summary of Kidney Diagnosis Data

This appendix presents a summary of the dataset used in the analysis of kidney diagnosis under the Modified Sequential Probability Ratio Test (MSPRT) and Cum-Maxwell distribution. The dataset contains information on 60 patients, including demographic, clinical, and diagnostic variables.

Summary Statistics

Index	count	mean	Std	min	25%	50%	75%	max
Patient_ID	60.0	30.5	17.46	1.0	15.75	30.5	45.25	60.0
Age	60.0	54.25	21.94	21.0	34.0	57.5	73.25	83.0
Gender	60	nan	Nan	nan	nan	Nan	nan	nan
Blood_Pressure	60.0	134.68	27.30	90.0	116.0	133.0	156.0	179.0
Weight	60.0	82.74	19.75	51.2	66.88	81.75	97.1	119.0
Height	60.0	1.72	0.12	1.5	1.61	1.745	1.83	1.89
BMI	60.0	28.43	7.41	15.3	22.62	28.05	34.05	43.1
Serum_Creatinine	60.0	2.85	1.24	0.64	1.91	2.935	3.8625	4.96
GFR	60.0	48.14	35.21	20.2	25.90	34.05	52.5	156.2
Urinalysis	60	nan	Nan	nan	nan	Nan	nan	nan
MSPRT_Result	60	nan	Nan	nan	nan	Nan	nan	nan
CKD	60	nan	Nan	nan	nan	Nan	nan	nan

Interpretation

The dataset comprises 60 patients, with ages ranging from 21 to 83 years. The average blood pressure is approximately 134.7 mmHg, and the mean BMI is 28.4, suggesting that many patients are overweight. Serum creatinine levels vary from 0.64 to 4.96 mg/dL, while the glomerular filtration rate (GFR) ranges widely between 20.2 and 156.2 mL/min, indicating varying degrees of kidney function among the patients. Most urinalysis results were normal, but some patients showed hematuria or proteinuria. The MSPRT results identified 29 positive cases and 31 negative cases, aligning closely with the CKD classification results.