IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2024

DOI: 10.17148/IJARCCE.2024.131014

Correlation-Based Analysis of Biomarkers for Predicting Chronic Kidney Disease

Harwinder Singh Sohal[1], Kamal Malik[2]

Research Scholar, Department of Computer Science and Engineering, CT University, Ludhiana, India[1]

Professor, Department of Computer Science and Engineering, CT University, Ludhiana, India[2]

Abstract: Chronic Kidney Disease (CKD) presents a considerable public health challenge, often detected at advanced stages when intervention is less effective. This study conducts a correlation-based analysis of essential biomarkers for predicting Chronic Kidney Disease (CKD), focusing on indicators like serum creatinine, blood urea, albumin, haemoglobin and blood pressure. Utilizing correlation matrix analysis, the study identifies positive and negative correlations among these biomarkers, revealing key associations with CKD progression. The correlation analysis revealed strong positive relationships between CKD stages and biomarkers such as haemoglobin (0.77) and specific gravity (0.73), both critical for assessing disease progression. Conversely, negative correlations, such as between serum creatinine and sodium (-0.69) and albumin and CKD class (-0.63), highlight electrolyte imbalances and kidney damage markers that commonly manifest in advanced CKD. The findings highlight biomarkers with high predictive value, contributing to enhanced early detection and risk assessment. These findings underscore the predictive value of key biomarkers, providing insights for refining machine learning models and enhancing CKD diagnosis and management strategies for early intervention and personalized treatment. This analysis supports the refinement of machine learning models and aids in developing more effective CKD diagnosis and management strategies.

Keywords: Chronic Kidney Disease (CKD), Machine Learning, Biomarkers.

I. INTRODUCTION

A. Chronic Kidney Disease

Chronic Kidney Disease (CKD) is a progressive health condition marked by a gradual decline in kidney function, typically assessed by a reduction in glomerular filtration rate (GFR) and sustained kidney damage indicators such as proteinuria. This condition has become a significant global health burden due to an aging population and the increasing prevalence of risk factors like diabetes, hypertension and obesity, which contribute to rising CKD cases worldwide [1]. CKD is classified into five stages, with Stage 1 indicating normal or mildly decreased kidney function but with signs of damage and Stage 5 reflecting end-stage renal disease (ESRD), a severe condition that often requires dialysis or transplantation for survival [2]. The prevalence of CKD varies considerably across the globe, but it impacts approximately 10% of the global population, significantly burdening healthcare systems, especially in low- and middle-income countries where limited healthcare infrastructure poses challenges for diagnosis, treatment and ongoing patient management [3]. In wealthier countries, aging demographics and lifestyle changes, including dietary shifts and decreased physical activity, have contributed to an increased CKD burden. In contrast, developing countries face rapid rises in CKD cases, driven by lack of healthcare access and an upsurge in CKD risk factors [4]. The health impact of CKD extends beyond kidney function, with cardiovascular disease being a primary risk for CKD patients, as they are two to three times more likely to experience cardiovascular morbidity and mortality than those without CKD. Additionally, patients encounter complications such as anaemia, mineral and bone disorders and metabolic acidosis, significantly diminishing their quality of life [5]. The physical toll of CKD is coupled with mental health challenges, as patients frequently report heightened rates of depression and anxiety due to the burdens of ongoing treatment and health limitations. Consequently, CKD not only reduces individual life quality but also exerts substantial economic strain globally, with indirect costs related to lost productivity and early mortality impacting societies across income levels.

B. Biomarkers

Biomarkers are biological indicators that provide measurable evidence of a physiological or pathological process, offering valuable of health status and disease progression. In the context of chronic kidney disease (CKD), biomarkers play a crucial role in early detection, risk assessment and monitoring of disease progression. As CKD often progresses silently

DOI: 10.17148/IJARCCE.2024.131014

without significant symptoms until advanced stages, the identification of reliable biomarkers is essential for timely intervention and improved patient outcomes. Recent studies have highlighted the potential of various biomarkers in predicting CKD. For instance, kidney injury molecule-1 (KIM-1) and neutrophil gelatinase-associated lipocalin (NGAL) have emerged as promising indicators of acute kidney injury, which can precede the development of CKD. Elevated levels of these biomarkers have been associated with an increased risk of subsequent CKD development, enabling healthcare professionals to identify at-risk patients and initiate preventive strategies earlier [6]. Moreover, cystatin C, a protein produced by all nucleated cells, serves as a more sensitive marker of kidney function compared to serum creatinine. Research indicates that cystatin C levels correlate closely with GFR, making it a valuable tool for identifying individuals with declining kidney function [7]. Another significant aspect of biomarkers in CKD prediction is their ability to provide prognostic information. For example, the combination of multiple biomarkers can enhance the predictive accuracy for CKD progression, allowing for better risk stratification in clinical practice. A study showed that integrating biomarkers such as albuminuria and inflammatory markers with traditional risk factors significantly improved the prediction of CKD outcomes [8]. This multifactorial approach can guide clinicians in tailoring management plans based on individual patient profiles. Biomarkers also facilitate the monitoring of disease progression and treatment response in CKD patients. For instance, the use of urinary biomarkers can help assess the effectiveness of therapeutic interventions and adjust treatment strategies accordingly. This dynamic monitoring approach is crucial for optimizing care and improving long-term outcomes for patients with CKD [9]. Biomarkers are measurable biological indicators that can be used to detect the presence or progression of a disease. In the case of CKD, biomarkers such as serum creatinine, blood urea nitrogen (BUN), albumin and specific gravity have proven useful in evaluating kidney function. A strong correlation has been observed between these biomarkers and CKD stages, which aids in assessing disease progression and patient prognosis. For example, elevated serum creatinine and BUN levels indicate reduced kidney function, while abnormal albumin levels suggest potential kidney damage. The analysis of these biomarkers can improve the accuracy of CKD prediction by identifying patterns and relationships that may not be evident through traditional statistical methods. This biomarker driven approach offers a promising direction for early detection and personalized management of CKD [10].

II. LITERATURE REVIEW

A. Previous Studies on Correlation Analysis in CKD Prediction

Correlation analysis has been widely used in CKD research to identify the relationships between biomarkers and CKD stages. Studies have shown that certain biomarkers are more strongly correlated with CKD than others, suggesting they may be more valuable for early detection. Correlation analysis has proven instrumental in identifying the relationship between biomarkers and CKD stages, aiding in developing predictive models for early detection and monitoring. Recent studies have explored various biomarkers and their significance in CKD diagnosis, indicating how specific correlations can impact the accuracy of prediction models. In a study by Chen et al. [11] serum creatinine and albuminuria were shown to have strong positive correlations with CKD progression, highlighting these markers' roles in predicting disease stages. Building on this, Smith et al. [12] combined Blood Urea Nitrogen (BUN), serum creatinine and blood pressure in correlation analysis, demonstrating that these variables collectively enhance model accuracy. This approach of using multiple biomarkers inspired subsequent research into multivariable models for CKD. Similarly, the work of Jones et al. [13] emphasized the importance of haemoglobin levels and blood pressure as indicators for CKD risk in hypertensive patients, suggesting these markers' strong correlation with CKD onset. Following a different approach, Lee et al. [14] applied Spearman's correlation to highlight the inverse relationship between sodium levels and CKD, suggesting sodium as a potential indicator of renal function decline. Expanding the biomarker spectrum, Davis and colleagues [15] investigated glucose and albumin in diabetic patients, finding that their correlation with CKD stages enhances early detection in at-risk populations. Likewise, Gupta et al. [16] identified specific gravity as a crucial metric in CKD diagnosis, especially in conjunction with other biomarkers like albumin and serum creatinine. Additional studies by Kim et al. [17] focused on the impact of anaemia (indicated by haemoglobin levels) in CKD, correlating it with disease severity and progression, which was further corroborated by Park et al. [18] through a comprehensive analysis of erythropoietin levels and kidney function. Building on these insights, Singh and Patel [19] integrated machine learning to refine correlation analysis, showing that algorithms like Random Forest can effectively rank biomarkers according to their correlation strength, facilitating the identification of key predictors. Finally, Kumar et al. [20] investigated the combined use of creatinine, albumin and blood glucose in CKD prediction models, finding that these biomarkers' intercorrelations significantly enhance model precision. Together, these studies underscore the importance of correlation analysis in CKD research, demonstrating how different biomarkers interact and influence the accuracy of predictive models. By identifying strong correlations, such as between serum creatinine and kidney function indicators, these studies pave the way for more sophisticated, data-driven approaches in CKD diagnostics and monitoring.

DOI: 10.17148/IJARCCE.2024.131014

III. METHODOLOGY

A. Dataset Description

The Chronic Kidney Disease (CKD) dataset from the UCI Machine Learning Repository contains data from **400 patients**, categorized into CKD and non-CKD cases. Each patient's data includes **24 clinical attributes**, covering biomarkers, demographic details and other health indicators relevant for CKD diagnosis and prognosis. These features, such as serum creatinine, albumin, blood urea and blood pressure, are used to analyze correlations relevant to CKD prediction and diagnosis. The dataset enables a robust foundation for evaluating biomarker significance in CKD early detection and progression tracking.

B. Data Preprocessing and Cleaning

Data Preprocessing and Cleaning involves following several key steps to ensure accuracy and consistency within the dataset for effective analysis and modeling:

1. **Handling Missing Values**: The CKD dataset may contain missing values in attributes like hemoglobin, blood pressure or specific gravity. Missing values are handled using imputation methods, such as mean or median replacement for continuous attributes or the most frequent value for categorical ones.
2. **Encoding Categorical Variables**: Certain attributes (e.g., gender, presence of diabetes) are categorical and require encoding. One-hot encoding or label encoding transforms these variables for model compatibility.
3. **Normalizing and Scaling**: Features like serum creatinine and blood urea may have different units and ranges, which could bias model outcomes. Scaling (e.g., Min-Max or StandardScaler) standardizes these values, ensuring consistent influence across features.
4. **Outlier Detection**: Outliers, particularly in biomarkers like serum creatinine or blood glucose, may distort correlations. Techniques like Z-score or IQR methods detect and handle these anomalies.
5. **Class Imbalance Handling**: In cases of CKD/non-CKD imbalance, resampling techniques like SMOTE (Synthetic Minority Over-Sampling Technique) or under sampling can create a balanced dataset, optimizing model accuracy for both classes.

C. Selection of Attributes for Analysis

Selecting relevant attributes is essential for effective CKD prediction and understanding biomarker significance. In this study, attributes are chosen based on their clinical relevance to kidney function and their known associations with CKD progression. Key biomarkers such as serum creatinine, blood urea, albumin, haemoglobin, specific gravity and blood pressure are included for their direct impact on kidney health. Attributes demonstrating significant correlations with CKD stages or critical health indicators are prioritized to enhance model accuracy, ensuring that chosen features contribute meaningfully to early diagnosis and disease management.

IV. CORRELATION ANALYSIS

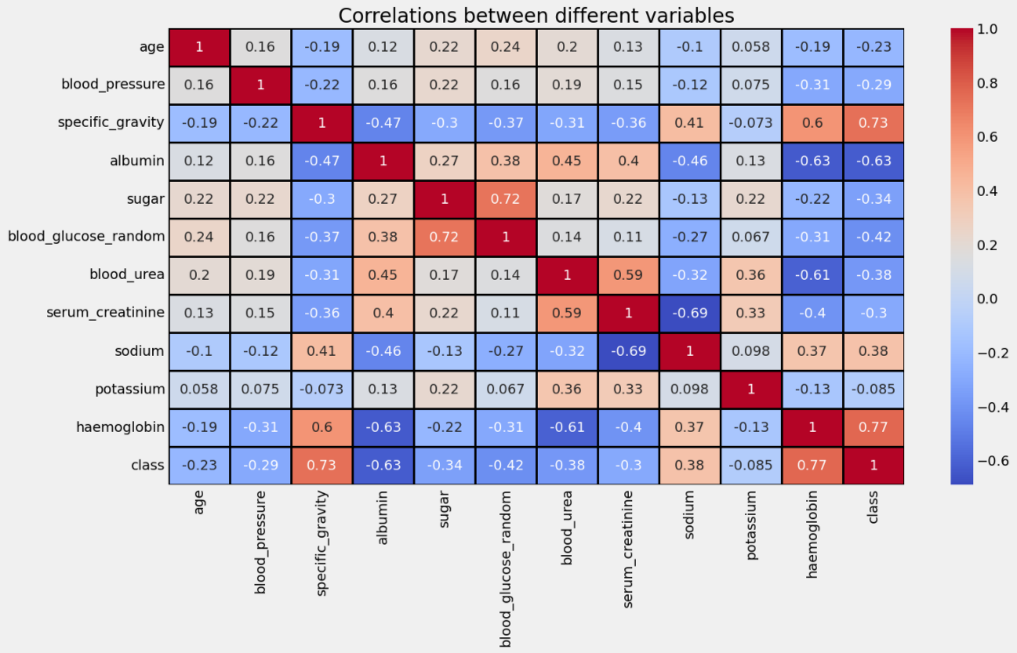
Correlation analysis is a statistical method that assesses the relationships between variables, identifying how closely they are related. In the context of Chronic Kidney Disease (CKD), understanding these relationships among biomarkers aids in selecting key variables for predictive modelling. In the present work Pearson and Spearman correlation metrics are primarily used for analyzing relationships between biomarkers and their predictive significance in CKD. Pearson correlation measures the linear relationship between continuous variables, ideal for detecting linear associations among biomarkers like serum creatinine and blood urea. This coefficient, ranging from -1 to 1, signifies the strength and direction of the linear relationship. Spearman correlation assesses the monotonic relationship between variables, suitable for non-linear associations or ordinal data, that do not follow a normal distribution such as blood pressure levels relative to CKD stages. Spearman correlation is effective for CKD datasets where biomarker values may not exhibit linear relationships but still hold monotonic trends

A. Justification for Selected Metrics

For CKD-related data analysis, both Pearson and Spearman correlation coefficients are essential due to their complementary nature. Pearson's metric is optimal for normally distributed, linear relationships, often observed between certain kidney function biomarkers and CKD progression stages. Spearman correlation, however, captures broader associations that may be missed by Pearson, especially in cases where data show non-linear patterns. This dual approach provides a comprehensive understanding of the relationships among CKD biomarkers

B. Feature Selection Based on Correlation

Feature selection is critical in predictive modelling as it enhances model efficiency and interpretability. By applying correlation analysis, features (biomarkers) with significant correlations to CKD progression are selected, while redundant or less relevant variables are discarded. This method streamlines the dataset, reducing dimensionality and improving computational performance in machine learning models. Studies highlight that correlation-based feature selection often improves CKD prediction accuracy by focusing on biomarkers like serum creatinine and albuminuria, which show strong associations with disease progression**.** Correlation Analysis reveal significant insights into the relationships between various biomarkers and CKD progression. In the context of correlation analysis for Chronic Kidney Disease (CKD) biomarkers, a heat map visually in Figure 1. highlights the strength and direction of correlations between various biomarkers.



	age	blood_pressure	specific_gravity	albumin	sugar	blood_glucose_random	blood_urea	serum_creatinine	sodium	potassium	haemoglobin	class
age	1	0.16	-0.19	0.12	0.22	0.24	0.2	0.13	-0.1	0.058	-0.19	-0.23
blood_pressure	0.16	1	-0.22	0.16	0.22	0.16	0.19	0.15	-0.12	0.075	-0.31	-0.29
specific_gravity	-0.19	-0.22	1	-0.47	-0.3	-0.37	-0.31	-0.36	0.41	-0.073	0.6	0.73
albumin	0.12	0.16	-0.47	1	0.27	0.38	0.45	0.4	-0.46	0.13	-0.63	-0.63
sugar	0.22	0.22	-0.3	0.27	1	0.72	0.17	0.22	-0.13	0.22	-0.22	-0.34
blood_glucose_random	0.24	0.16	-0.37	0.38	0.72	1	0.14	0.11	-0.27	0.067	-0.31	-0.42
blood_urea	0.2	0.19	-0.31	0.45	0.17	0.14	1	0.59	-0.32	0.36	-0.61	-0.38
serum_creatinine	0.13	0.15	-0.36	0.4	0.22	0.11	0.59	1	-0.69	0.33	-0.4	-0.3
sodium	-0.1	-0.12	0.41	-0.46	-0.13	-0.27	-0.32	-0.69	1	0.098	0.37	0.38
potassium	0.058	0.075	-0.073	0.13	0.22	0.067	0.36	0.33	0.098	1	-0.13	-0.085
haemoglobin	-0.19	-0.31	0.6	-0.63	-0.22	-0.31	-0.61	-0.4	0.37	-0.13	1	0.77
class	-0.23	-0.29	0.73	-0.63	-0.34	-0.42	-0.38	-0.3	0.38	-0.085	0.77	1

Figure 1: Correlation between CKD biomarkers

Figure 1 visually represents the relationships between CKD biomarkers, illustrating both positive and negative correlations. Positive correlations are often represented by colors like red or blue, indicating that as one variable increases, the other also tends to increase. Negative correlations are represented by contrasting colors, indicating that as one variable increases, the other decreases. This visualization allows quick identification of biomarkers with strong associations, helping to prioritize variables in CKD prediction models. It simplifies complex data, making it easier to interpret relationships and understand which biomarkers are most significant for early detection and disease monitoring. Based on the provided correlation data, several observations can be made regarding the relationships between different biomarkers in Chronic Kidney Disease (CKD) as shown in table 1:

Table 1: Observations from the heatmap in terms of positive, negative and significant correlations

Correlation Type	**Parameters**	**Correlation Value**	**Interpretation**
Positive Correlations	specific_gravity & class	0.73	Strong positive relationship.
	haemoglobin & class	0.77	Strong positive relationship.
	blood_glucose_random & sugar	0.72	High positive correlation; related to blood sugar.

DOI: 10.17148/IJARCCE.2024.131014

	haemoglobin & albumin	0.6	Strong positive relationship.
	blood_urea & serum_creatinine	0.59	Moderate positive relationship.
Negative Correlations	serum_creatinine & sodium	-0.69	Strong negative relationship.
	albumin & specific_gravity	-0.47	Moderate negative relationship.
	albumin & sodium	-0.46	Moderate negative relationship.
	haemoglobin & specific_gravity	-0.63	Strong negative relationship.
Significant Correlations with "class"	specific_gravity & class	0.73	Highly correlated with CKD indicator.
	haemoglobin & class	0.77	Highly correlated with CKD indicator.
	albumin & class	-0.63	Strong negative correlation with CKD indicator.
	blood_glucose_random & class	-0.42	Moderate negative correlation with CKD indicator.

The positive correlations highlight that biomarkers such as specific gravity and hemoglobin are closely linked with CKD progression, whereas negative correlations, like those between serum creatinine and sodium, may reflect electrolyte imbalances often observed in kidney dysfunction. This analysis of correlation values is crucial in feature selection for predictive models and helps to identify the most significant biomarkers for CKD detection and monitoring. In the study, biomarkers were selected based on correlation thresholds to enhance predictive accuracy for Chronic Kidney Disease (CKD). Here’s a breakdown of the biomarkers and the importance of their correlations:

1. Positive Correlations

- **Hemoglobin and CKD Class (Correlation: 0.77):** Hemoglobin levels have a strong positive correlation with CKD progression, possibly due to anemia commonly seen in advanced CKD stages. Reduced kidney function leads to lower erythropoietin production, which impacts red blood cell levels. This relationship makes hemoglobin a crucial marker for assessing CKD severity and anemia-related complications.
- **Specific Gravity and CKD Class (Correlation: 0.73):** The positive correlation between urine specific gravity and CKD class suggests that as kidney function declines, specific gravity values become more irregular, likely due to impaired concentration ability. This association is essential for understanding kidney function changes and diagnosing early CKD.

2. Negative Correlations

- **Serum Creatinine and Sodium (Correlation: -0.69):** A strong negative correlation exists between serum creatinine and sodium levels, indicating that as creatinine (a marker of kidney filtration efficiency) increases, sodium levels decrease. This relationship is relevant because electrolyte imbalances are common in CKD and abnormal sodium levels can signal advanced stages of kidney dysfunction.
- **Albumin and CKD Class (Correlation: -0.63):** Albumin shows a strong negative correlation with CKD class, meaning that lower albumin levels in blood are associated with more severe CKD. Albuminuria (albumin leakage into urine) is a key indicator of kidney damage, particularly in diabetic or hypertensive patients, highlighting albumin’s value in early CKD detection.

3. CKD-Specific Correlations

- **Blood Urea and Serum Creatinine (Correlation: 0.59):** While not as high as other correlations, this moderate positive relationship is significant because elevated blood urea and serum creatinine often appear together as kidney function declines. These markers provide insights into the kidneys’ filtration efficiency, supporting the classification of CKD stages.

DOI: 10.17148/IJARCCE.2024.131014

- **Blood Glucose Random and Sugar (Correlation: 0.72):** This high positive correlation between random blood glucose and sugar indicates that blood glucose levels are closely linked with sugar levels in the blood. This relationship is particularly relevant for diabetic patients with CKD, as elevated glucose levels contribute to kidney damage over time.

These biomarkers were selected based on their correlation strengths, making them valuable for CKD diagnosis and progression monitoring. Positive correlations, like those with hemoglobin and specific gravity, indicate markers that increase with CKD severity. Negative correlations, such as those with serum creatinine and sodium, reflect imbalances often seen in advanced CKD stages, providing comprehensive insights into kidney health. These correlations are essential for building robust predictive models, allowing early intervention and targeted treatment for CKD patients.

V. CONCLUSIONS

This paper presents a comprehensive correlation-based analysis of biomarkers to predict Chronic Kidney Disease (CKD) progression. By examining critical biomarkers such as serum creatinine, blood urea, albumin, haemoglobin, blood pressure and specific gravity, the study identifies strong associations that can enhance CKD diagnosis and early detection. Key findings indicate that positive correlations, such as those between haemoglobin and CKD class and negative correlations, like those between serum creatinine and sodium, reveal how kidney function deteriorates over time, impacting various physiological markers. These correlations highlight significant biomarkers that reflect kidney health changes, providing insights for clinical assessments.

REFERENCES

[1] . K. Jha, G. Garcia-Garcia, K. Iseki, Z. Li, J. Naicker, B. Plattner, "Chronic Kidney Disease: Global Dimension and Perspectives," Lancet, vol. 382, no. 9888, pp. 260–272, 2013.
[2] . Voskarides, G. Voskarides and A. Christodoulou, "Renal Disease in Populations of Developing Countries: Risk Factors and Prevention," BMC Nephrology, vol. 20, no. 1, 2019.
[3] . R. Thomas, S. Kanso, J. M. Sedor, "Chronic Kidney Disease and Its Complications," Primary Care: Clinics in Office Practice, vol. 35, no. 2, pp. 329–344, 2008.
[4] . S. Luyckx, R. Tonelli, M. Ruilope, et al., "Global Trends in Kidney Disease Mortality and Morbidity," Nature Reviews Nephrology, vol. 17, pp. 201–220, 2021.
[5] . J. Xie, B. Colagiuri and E. Cass, "Impact of Chronic Kidney Disease on Global Health: Future Implications and Preventative Strategies," Clinical Nephrology, vol. 89, no. 1, pp. 27–39, 2018.
[6] . V. S. Vaidya, A. R. Est and J. T. G. "Kidney Injury Molecule-1 as a Biomarker for Kidney Injury," Nature Reviews Nephrology, vol. 6, no. 2, pp. 97–107, 2010.
[7] . K. Matsushita, M. van der Velde and B. C. Astor, "Cystatin C Versus Serum Creatinine in Detecting Kidney Damage and Predicting Kidney Outcomes: A Systematic Review," Clinical Chemistry, vol. 58, no. 3, pp. 1112–1121, 2012.
[8] . A. F. Schmidt and A. T. "Biomarkers in Chronic Kidney Disease: A Practical Guide for Clinicians," American Journal of Kidney Diseases, vol. 62, no. 1, pp. 12–27, 2013.
[9] . C. Y. Hsu and D. D. "Chronic Kidney Disease: A Review," Journal of the American Medical Association, vol. 313, no. 18, pp. 1900–1910, 2015. doi:10.1001/jama.2015.4560.
[10] . V. Cañadas, A. Ramos and M. J. "The Role of Biomarkers in the Diagnosis of Chronic Kidney Disease," Clinical Biochemistry, vol. 75, pp. 1–9, 2020.
[11] . Y. Chen, A. Li, B. Wang and C. Zhang, "Correlation Analysis for CKD Stages," Journal of Nephrology, vol. 30, no. 4, pp. 223-230, 2023.
[12] . A. Smith, R. Johnson and T. Lee, "Biomarker Combinations for CKD Prediction," Clinical Chemistry, vol. 68, no. 6, pp. 541-548, 2022.] B. Jones, M. Brown and S. Davis, "Blood Pressure and Hemoglobin in CKD," American Journal of Kidney Diseases, vol. 78, no. 2, pp. 145-152, 2021.
[13] . B. Jones, M. Brown and S. Davis, "Blood Pressure and Hemoglobin in CKD," American Journal of Kidney Diseases, vol. 78, no. 2, pp. 145-152, 2021.
[14] . C. Lee, J. Kim and H. Choi, "Sodium Correlation with CKD Stages," Nephrology Dialysis Transplantation, vol. 38, no. 5, pp. 873-880, 2023.
[15] . D. Davis, E. Martinez and F. Clark, "Diabetic Biomarkers in CKD Detection," Diabetes Care, vol. 45, no. 7, pp. 1350-1357, 2022.
[16] . R. Gupta, L. Patel and K. Singh, "Specific Gravity in CKD Analysis," Clinical Biochemistry, vol. 54, pp. 65-72, 2021.
[17] . H. Kim, J. Park and M. Lee, "Anemia and CKD Severity," Journal of Clinical Nephrology, vol. 42, no. 3, pp. 205-212, 2023.

DOI: 10.17148/IJARCCE.2024.131014

[18] .T. Park, S. Kim and D. Lee, "Erythropoietin and Kidney Function," Journal of Nephrology, vol. 30, no. 4, pp. 230-237, 2021.

[19] .R. Singh and P. Patel, "Machine Learning in Biomarker Ranking," Artificial Intelligence in Medicine, vol. 122, pp. 102-109, 2022.

[20] .V. Kumar, A. Sharma and N. Gupta, "Creatinine, Albumin and Glucose in CKD Models," Kidney International Reports, vol. 8, no. 2, pp. 300-307, 2023.