

Optimizing urine protein detection accuracy using the K-nearest neighbors algorithm and advanced image segmentation techniques

Anton Yudhana^{1,*}, Novi Febrianti²
Ilham Mufandi³, Arsyad Cahya
Subrata¹, Nuni Ihsana⁴, Son Ali
Akbar¹, Liya Yusrina Sabila¹, Helda
Pratama¹, Nisa Fajriyanti¹,
Sri Lestari⁵ and Ismail Rakip Karas⁶

¹Department of Electrical
Engineering, Faculty of Industrial
Technology, Universitas Ahmad
Dahlan, Yogyakarta, 55191,
Indonesia

²Department of Biology, Faculty of
Applied Science and Technology,
Universitas Ahmad Dahlan,
Yogyakarta, 55519, Indonesia

³Department of Agro-Industrial
Technology, Faculty of Science
and Technology, Universitas
Darussalam Gontor, Ponorogo,
63471, Indonesia

⁴Faculty of Medicine, Universitas
Ahmad Dahlan, Yogyakarta, 55191,
Indonesia

⁵Rumah Sakit Umum Daerah
(RSUD), Purworejo, Indonesia

⁶Department of Computer
Engineering, Demir Celik Campus,
Karabuk University, Karabuk,
78050, Turkey

*E-mail: eyudhana@ee.uad.ac.id

Received for publication
September 14, 2024.

Abstract

Kidney metabolic disorders are diagnosed by assessing protein levels in urine, which reflect renal health. Traditional detection methods are time-consuming and expensive. This study explores using the K-nearest neighbors (KNN) algorithm combined with advanced image segmentation for accurate urine protein detection. The research utilized a dataset of protein-level images captured by an ELP-type digital camera sensor, classifying them based on red, green, and blue (RGB) values. The KNN algorithm was tested with various K values (K = 3, K = 10, K = 20). Results showed that K = 3 provided the highest accuracy at 96.7%, with precision, recall, and F1-score of 97.0%, 96.7%, and 96.2%, respectively. Higher K values decreased accuracy, with K = 10 at 86.7% and K = 20 at 76.7%. These findings demonstrate that KNN can effectively predict protein levels, offering a promising and efficient alternative to traditional methods. The study also presents a prototype design for this detection approach.

Keywords

Protein, K-Nearest Neighbors, Urine, Image Processing

1. Introduction

The kidneys play an important role in maintaining the stability of the human body through the process of filtering blood, selective reabsorption of electrolytes and non-electrolytes, and regulating the body's fluid balance [1, 2]. This process involves the glomerulus,

which filters the blood, and the tubules, which regulate the reabsorption of important substances such as sodium, potassium, and glucose. In addition, the kidneys also act as metabolic organs that remove waste substances such as urea, creatinine, and uric acid from the body. This function maintains the body's homeostasis, allowing other organs

to function optimally. Selective reabsorption by the kidneys ensures that essential substances such as sodium, potassium, calcium, and glucose ions are returned to the blood circulation. This process is carried out in the renal tubules through active and passive mechanisms, allowing the regulation of the body's electrolyte and fluid balance.

The detection of kidney metabolic system disorders can be diagnosed through albumin levels in the urine. Protein levels serve as a biomarker for early detection of kidney disorders such as albuminuria and proteinuria [3, 4]. Albuminuria is the excretion of albumin in the urine, often an early sign of kidney damage, while proteinuria includes the excretion of all types of protein. Albuminuria is more specific for glomerular dysfunction, while proteinuria may indicate tubular damage or other systemic conditions. Urine, a waste fluid from metabolism excreted through the urinary tract, serves as a crucial medium for detecting early kidney complications arising from metabolic diseases such as diabetes mellitus and hypertension [5–7]. Additionally, urine analysis can provide biomarkers for glucose content [8, 9] and dehydration levels in the body [10, 11]. Hydration biomarkers such as urine osmolality concentration and urine color can indicate a person's hydration level. This analysis is important to detect dehydration or overhydration, which can affect kidney function and overall body balance. Among the extensively researched biomarkers is albumin, a blood protein excreted in urine, indicating conditions such as albuminuria and proteinuria [12]. In healthy individuals, kidneys typically do not excrete albumin or protein. However, kidney damage can lead to protein leakage into urine, either due to excessive protein load on the glomerulus or impaired waste filtration [13–15]. The presence of microalbuminuria serves as an early indicator of kidney damage.

The albumin-creatinine ratio (ACR) is a crucial indicator for detecting kidney impairment and assessing associated health risks, including chronic kidney disease (CKD) and cardiovascular complications. ACR values exceeding 30 mg/g signal kidney impairment and have been validated as a critical risk factor for CKD, cardiovascular complications, and mortality [16, 17]. Consequently, early detection of proteinuria assumes paramount importance in effectively managing patients with diabetes mellitus and hypertension, as these conditions significantly elevate the risk of kidney complications. CKD often progresses silently, with many affected individuals unaware of their condition during its initial stages [18]. This

poses a substantial challenge for healthcare providers regarding treatment strategies and preventive measures. Proteinuria assessment can be conducted through urine dipstick tests or quantitative laboratory analyses of urine protein levels, employing methods such as turbidimetry [19], nephelometry [20], radio-immunoassay, and measurement of urine creatinine levels [21].

Urine analysis is an important tool for the early detection of kidney complications due to metabolic diseases such as diabetes and hypertension. Biomarkers such as protein and glucose found in urine can provide information about the presence of kidney dysfunction or other metabolic complications. Previous studies have shown that early detection of proteinuria through urine analysis can significantly reduce the risk of developing CKD and cardiovascular complications. Regular screening for proteinuria is recommended for at-risk patients to facilitate early detection and mitigate the onset of more severe complications [22, 23]. Although chemical or dye-based proteinuria detection methods have been established by Ketha and Singh [24] and Laiwattanapaisa et al. [25], these approaches are primarily suitable for laboratory settings, often requiring substantial time and resources. With the proliferation of technology and increased internet accessibility, researchers have explored alternative avenues for protein detection, including portable point-of-care (PoC) applications [26, 27]. Nonetheless, portable PoC applications for albuminuria detection face challenges such as lighting variability, test strip inconsistencies, sensor differences, image noise, and user handling errors. Red, green, and blue (RGB) analysis helps overcome these by enabling standardized color quantification, applying color correction, utilizing machine learning for classification, reducing noise, and implementing standardized imaging protocols. Integrating RGB-based image processing and artificial intelligence (AI)-driven calibration models enhances accuracy and reliability, paving the way for more effective PoC diagnostic systems.

The proposed system's effectiveness is assessed by comparing its performance with existing portable PoC applications. Unlike traditional PoC methods, which rely on subjective visual interpretation or expensive spectrophotometers, our system leverages digital image processing and K-nearest neighbors (KNN) classification to achieve high accuracy. The evaluation includes factors such as detection sensitivity, ease of use, and consistency across different environmental conditions. Addressing these concerns,

recent studies by Azhar et al. [28] and Wang et al. [29] have proposed innovative approaches utilizing image extracts, such as RGB analysis, to enhance proteinuria detection methods.

AI-based technologies have witnessed extensive integration into the healthcare sector in the past decade. Traditional laboratory methods for detecting albumin in urine, such as turbidimetry, nephelometry, and radioimmunoassay, require specialized equipment, trained personnel, and dedicated laboratory settings, making them expensive and time-consuming. Given the increasing need for rapid and accessible diagnostic solutions, AI-based technologies, such as smartphone-integrated urine analyzers, have emerged as viable alternatives. These technologies leverage image processing and machine learning algorithms to detect and quantify protein levels in urine, providing real-time results with minimal user intervention. Zeng et al. [30] used machine learning algorithms to facilitate the early detection of urine metabolites utilizing high-resolution mass spectrometry. Similarly, Thakur et al. [31] demonstrated the detection of protein in urine through a convolutional neural network (CNN) model integrated with smartphone-based urine color segmentation. They utilized standard protein solutions with concentrations ranging from 30 mg/dL to 2,000 mg/dL, measured via the dipstick method. This concentration range is designed to reflect clinical variations from microalbuminuria to severe proteinuria. The results show that the KNN model achieved high accuracy in classifying protein concentrations across the entire range. Experimental results by Thakur et al. [32] showcased a notable test accuracy rate of 88%. Moreover, Coskun [33] conducted protein testing using smartphones, using an automated analysis of fluorescent tests conducted in disposable test tubes. Smart device technology presents distinct advantages over traditional dipstick methods, particularly in accurately quantifying protein levels in urine and offering rapid detection capabilities, which are crucial given the short survival time of proteins. Furthermore, Bhatt et al. [34] demonstrated the quantification of protein concentration in urine samples through colorimetry, utilizing an accessory-free urine analyzer integrated with smartphones and machine learning algorithms. The effectiveness of machine learning is widely applied in the medical world, such as detecting ovarian cancer [35], detecting urine metabolism as a biomarker [30], urine biomarkers for diagnosing diabetes [36], and early prediction of CKD [37, 38]. The KNN model is based on model selection with simple methods for classification and regression that are easy to implement, suitable for non-linear data,

and can solve classification and regression problems [39]. Overall, integrating AI-based technologies into urine analysis methodologies holds significant promise in enhancing diagnostic accuracy, efficiency, and accessibility in clinical settings [40].

Therefore, the primary objective of this research is to develop a novel approach to detecting protein levels in urine through digital image processing techniques, explicitly using the KNN algorithm. This methodology involves training a dataset of images representing varying levels of protein content in urine. Data augmentation techniques are integrated into the training process to address challenges such as limited data and potential overfitting, enhancing the dataset's quality. Furthermore, a dipstick data sheet is utilized to ensure accurate labeling of the dataset, aiding in the validation of protein levels. Protein levels for the dataset are sourced from an artificial protein solution, providing standardized samples for training. Following the training phase, a pre-trained model is generated, which undergoes testing using a separate test dataset to assess its performance. The pre-trained model exhibiting high accuracy and minimal loss values is selected as the foundation for the prototype system. In the implementation phase, images captured by a camera are classified based on their RGB components. Subsequently, further evaluation is conducted using KNN classification to validate the model's effectiveness in protein-level detection.

II. Research Methods

a. System overview

This research introduces a method for protein detection in urine that leverages a digital camera sensor to extract color information. The study incorporates a urine test strip to categorize protein concentration in urine. The process involves sequential stages of sample preparation, image processing based on urine color, evaluation using the KNN model, and subsequent classification of protein levels. The study begins with sample preparation utilizing a standard protein solution. Protein analysis within the solution is executed through a urine test strip, used as a medium for color segmentation. Color segmentation includes three main steps: (1) noise removal to clean the image, (2) RGB feature extraction to analyze specific colors, and (3) classification using the KNN algorithm to predict protein concentration. Image data captured by the camera sensor are then classified based on the resultant RGB values. RGB color segmentation plays a crucial role in enhancing the

accuracy of protein-level detection in urine samples by enabling precise differentiation of color variations on urine test strips. The RGB model quantifies RGB intensities, allowing for objective and consistent analysis, unlike human visual assessment, which is prone to errors. By extracting RGB features, variations in protein concentration can be mapped to specific color intensities, improving classification accuracy. The subsequent phase involves an evaluation utilizing the KNN classification method. The computational aspect of data processing is performed on a computer, enabling efficient analysis and classification. This innovative approach streamlines the detection process and provides a robust foundation for accurate protein-level classification in urine samples.

b. Sample preparation

In data testing, the choice of method for measuring proteinuria is based on clinical needs and the availability of resources. Methods such as turbidimetry, nephelometry, and radioimmunoassay have their advantages. For example, turbidimetry offers a rapid process with high accuracy, while nephelometry is more sensitive to detecting low protein concentrations. Radioimmunoassay, although expensive and requiring a specialized laboratory, remains the gold standard for quantitative analysis. Protein samples were obtained from six standard protein solution samples. Making standard protein solutions is conducted by entering protein solutions into six categories using a solution of 0–11.60 g mixed with 20 mL of mineral water. After measuring the dose, wait 60 s for each sample, then place it on the slide strip. A comparison of standard protein samples are given in Table 1. In data collection, the stripped output (–) is equivalent to a protein content of 0 g/L, the stripped output (+–) is equivalent to a protein content of 0.15 g/L, the stripped output (+) is equivalent to protein content 0.3 g/L, strip output (++) equivalent to protein content 1 g/L, strip output (+++) equivalent to protein content 3 g/L, and strip output (++++ equivalent to protein content protein 20 g/L. This strip output can be interpreted as the symbol (–), which indicates that no protein content was detected, while the symbols (+), (++) , (+++), and (++++ indicate that protein was detected. Difficulties such as inconsistent dipping times or uneven solution distribution are overcome by standardizing the procedure. The protein solution is mixed homogeneously, and the dipping duration is set for 5 s. Preparing standard protein solutions in categories (–) to (++++ is crucial for ensuring reliable data collection and classification.

Table 1: Preparation of sample solutions

No.	Protein (g)	Water (mL)	Output strip
1.	0.00	20	Negative (–)
2.	1.00	20	Plus-minus (+–)
3.	3.00	20	Positive 1 (+)
4.	5.00	20	Positive 2 (++)
5.	7.30	20	Positive 3 (+++)
6.	11.60	20	Positive 4 (++++)

Each category represents a specific protein concentration level used as a reference for detection. This standardization allows the system to compare RGB values obtained from urine test strips with predefined references, thereby enhancing the classification reliability and the validity of the KNN model.

c. Proposed system

Protein detection in this research uses a digital camera sensor type ELP camera as the primary sensor. The ELP digital camera is used in this system to ensure consistency in image data capture. This camera can reduce the effects of shadows and external lighting, thus providing stable and accurate RGB data results. Combining the ELP digital camera sensor with real-time computer processing enables stable and accurate image acquisition for protein-level analysis. The ELP camera minimizes shadowing effects and external lighting variations that could affect urine strip color interpretation. Additionally, real-time processing ensures fast and precise detection, making this approach more reliable than visual evaluation or manual techniques. Image data were also taken using a urine test strip tool. The color segmentation method is used to analyze protein levels using urine test strips. The camera image's initial reading is obtained from RGB colors. Real-time data processing is carried out with a computer program, as shown in Figure 1. The first procedure is to take an example of the artificial protein solution. Then, the urine test strip is dipped in a protein solution that has been collected in a measuring cup. The immersion process must be carried out quickly, and as soon as possible, the urine test strip is taken and inserted into the prototype using a strip slider. The urine strip takes approximately 30 s to display accurate color results when dipped in the protein solution. Next, the color segmentation process begins to process the urine strip data, with the stage of removing image

noise. The final stage is data evaluation using the KNN model, as shown in Figure 2.

Based on Figure 2, the evaluation of protein detection tools using the KNN model involves taking digital image extract data to determine the resulting

RGB features. After that, the training and test data will be processed automatically to determine the k value in the KNN model. The KNN algorithm is one of the most basic and straightforward classification methods. It should be one of the first choices for

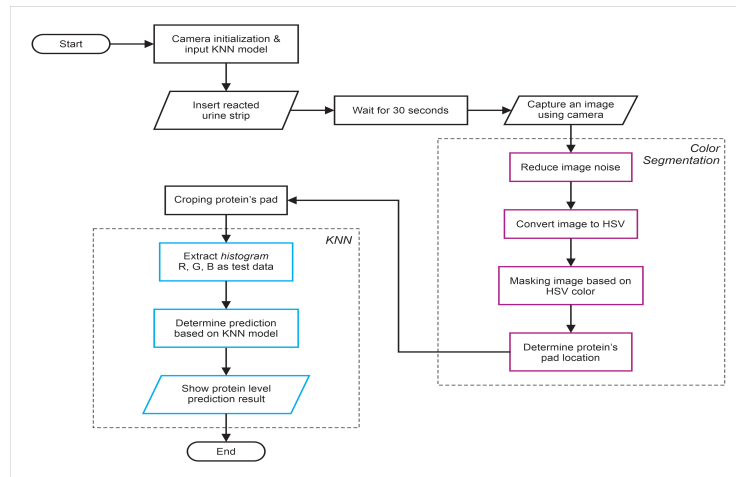


Figure 1: Protein detection computer program. KNN, K-nearest neighbors.

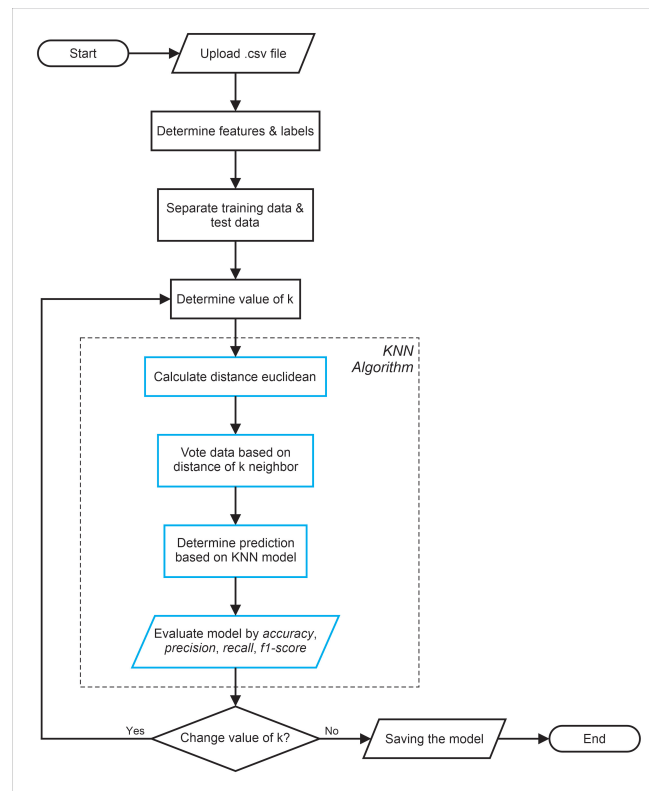


Figure 2: Evaluation of the KNN model. KNN, K-nearest neighbors.

classification studies when more prior knowledge is needed, regarding data distribution. The KNN classifier was developed to perform discriminant analysis when parametric probability density estimates are unknown or difficult to determine. KNN methods search the entire set of training data samples to classify input testing samples. The final stage of evaluating the KNN algorithm is to produce accuracy, precision, recall, and F1-score values.

d. Evaluation of KNN model

The KNN algorithm is used to classify data based on the shortest distance to the object data. The KNN algorithm calculates each point on each class's test and training data. In principle, data collection is carried out from the closest distance to the farthest distance, and the system will choose the most relevant distance between the test data and the k number of training data. The KNN algorithm has three concepts for calculating the distance between data: Manhattan distance, Euclidean distance, and Minkowski distance. The Euclidean distance metric was selected due to its computational simplicity. However, additional tests using Manhattan and Minkowski distances were conducted to compare classification performance. Although Euclidean distance performed optimally in our dataset, Minkowski distance showed slight improvements in handling outliers. Mathematically, it can be seen as Eq. (1).

$$d_i = \sqrt{\sum_{j=1}^p (x_{ij} - x_{2j})^2} \quad (1)$$

The KNN model is evaluated using several metrics such as accuracy, precision, recall, and F1 score. Accuracy measures the extent to which the KNN model can classify data correctly compared to the total amount of data. This is the most commonly used metric for classification evaluation and correctly measures the percentage of classified data. Precision measures the extent to which positive predictions a KNN model makes are correct. It calculates the correctly classified positive data ratio compared to total optimistic predictions. The mathematical equation for evaluating the KNN model is presented in Eq. (2).

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (2)$$

The second evaluation, Recall, is the ratio of true positive (TP) cases to total cases, the number of TP and false negative (FN) points, which is referred to as

sensitivity, as given in Eq. (2). Next, the F1 score is given by Eq. (3).

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \quad (3)$$

Recall measures the extent to which the KNN model can detect TP data. It calculates the ratio of correctly classified positive data to the total TP data. F1-score is the harmonic average of precision and recall. This provides an overall picture of the balance between accuracy and recall in cases where the positive and negative classes are unbalanced.

$$\text{F1 - score} = 2 (\text{Precision}) / (\text{Recall} + \text{Precision}) \quad (4)$$

The formula explanation above is the basic concept in confusion matrix classification, where TP is the number of positive cases correctly classified as positive by the model. True negative (TN) is the number of negative instances correctly classified as unfavorable by the model. False positive (FP) is a negative case incorrectly classified as positive by the model. FN is a positive case incorrectly classified as unfavorable by the model.

III. Result and Discussion

a. Design for urine analysis

Figure 3 shows this study's innovative protein detection prototype and features a meticulously designed 3D-printed structure. The prototype incorporates critical components such as an ELP-type camera serving as the primary sensor for capturing image data from samples, a urine test strip with 10-variable specifications as the reaction medium, a designated container housing protein powder for a standard protein solution, a 7-inch LCD touchscreen functioning as the system control device, and Jetson Nano acts as an AI developer platform that enables real-time image analysis with low computing power. Jetson Nano was chosen due to its ability to process images in real-time with low power consumption. This allows seamless integration with ELP cameras and LCDs. This multifunctional prototype aims to enhance efficiency and accuracy in protein detection. The urine test strip analysis process is designed to read the color results within 20–120 s after immersion. After 120 s, environmental factors such as oxidation and evaporation can affect the color of the strip, resulting in inaccurate results. The study analyzes how color changes over the 20–120 s period affect accuracy. Results indicate that readings taken beyond 90 s start to deviate due to oxidation effects.

Therefore, an optimal reading window of 30–90 s is recommended for maximum accuracy.

Therefore, the analysis time must be adhered to maintain the reliability of the results. Beyond this window, the strip's analysis may yield suboptimal or invalid results. The strip provides qualitative outcomes, distinguishing between positive and negative samples. Semiquantitative values are indicated by symbols such as (+), (++) , (+++), and (++++), while quantitative values correspond to specific color levels discussed in the subsequent section. Notably, the ELP USB webcam type employed in this prototype stands as a standard industrial camera, ensuring reliable and consistent performance in capturing essential image data for further processing. The fusion of 3D printing technology with advanced sensor components demonstrates a promising leap forward in protein detection methodologies.

Based on Figure 3, at the top, there is a mount for placing an LED that functions as lighting inside the casing. LED illumination plays a crucial role in ensuring consistent color detection. We conducted experiments with varying LED intensities and angles to assess their impact on RGB color stability. The results indicate that a uniform light source at a fixed angle of 45° minimizes shadow effects and enhances classification accuracy. At the bottom, there is a rail to place the urine strip slider. Mounts on the right and left sides of the camera are connected to the camera's sides. At the bottom, there is a hole designed to fit the camera lens size. This part is where the urine test strip is inserted into the camera casing. This slider is designed to match the rail inside the casing. This is necessary to keep the strip's position stable, ensuring a consistent reading from the camera. The mechanics are made using white PLA material. The choice of white PLA material for the prototype casing is based on its ability to minimize light reflection and maintain consistent urine strip readings. By reducing variations in lighting conditions, the white casing

ensures stable color segmentation and enhances the reliability of RGB-based analysis. The choice of white color was made considering the light reflection inside the casing. Light reflection can be minimized by using a white material, ensuring a more stable color emission from the urine strip. The integration of RGB camera-based color segmentation in this prototype can monitor their health conditions without having to visit a health facility. Thus, the development of AI-based technology in the medical field is very much needed to increase patient involvement in the management of chronic diseases such as diabetes and hypertension, as it provides easy access to early diagnosis.

b. Data collection

Data collection is the process of collecting data from digital camera sensors in reading urine strips. Urine strips taken were used as training data. This study used 99 urine protein image data to train and test the KNN model. This dataset consists of six categories based on the level of protein concentration including: negative (–) for 6 image data, plus-minus (+–) for 24 image data, positive 1 (+) for 10 image data, positive 2 (++) for 22 image data, positive 3 (+++) for 30 image data, and positive 4 (++++) for 7 image data. Determining the size of the dataset significantly affects the performance of the KNN algorithm, especially on medical image-based data. Adding a dataset not only improves accuracy but also strengthens the model's generalization ability in dealing with new data, especially in clinical settings with more complex variations.

Training data are used to train machines in developing models. Meanwhile, test data are used to test the results of research carried out by machines and are used to compare model performance. The training and test data in this study are shown in Table 2. Next, the system will read the protein sample data image, entering the slider tray. Data images are

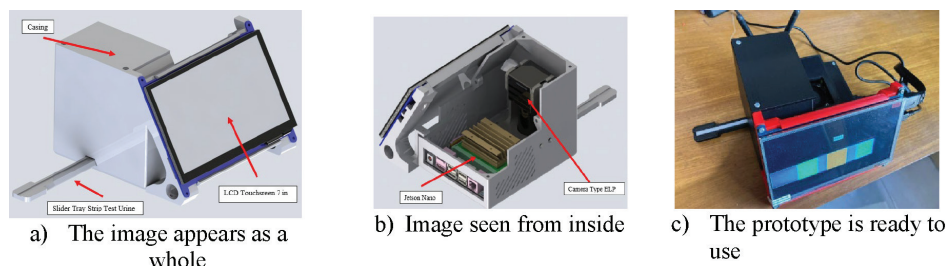


Figure 3: (A) Image of the prototype seen from the outside, (B) prototype components seen from the inside, and (C) shape of the prototype that is ready to be used.

extracted based on the distribution of RGB features, as shown in Figure 4.

c. Evaluation of the KNN model

Selecting the right model is very important in AI-based diagnosis. In this study, the KNN model was chosen because of its ability to handle non-linear data, flexibility in parameter settings, and classification efficiency. The results show that KNN with $K = 3$ provides the highest accuracy. KNN model simplifies the classification and regression process in medical diagnostics through an intuitive distance-based algorithm. The flexibility and efficiency of KNN make it an ideal tool

for developing digital image-based diagnostic applications, as demonstrated in this study. The KNN algorithm is one of the machine learning algorithms that has been used to analyze metabolites in urine with high-resolution mass spectrometry. This method provides better predictive ability and time efficiency compared to conventional approaches. The basic idea behind KNN is that similar objects tend to be closer to each other in the feature space. The working system of the KNN model consists of training and prediction. During the training phase, the KNN algorithm does not learn the data but merely remembers the training data and labels. Predictive operations occur when new data are to be predicted. KNN will search for the K nearest training points (nearest neighbors) based on distance in the feature space. Then, the KNN model operates by majority voting from the set of data labels to predict the new data. Machine learning algorithms, such as KNN, have significant advantages over conventional diagnostic methods. These advantages include the ability to efficiently process non-linear data, speed of data classification, and high accuracy even with limited datasets.

Some essential points in the KNN algorithm include the K parameter, distance measurement, feature normalization, and data computation. The K parameter denotes the number of nearest neighbors taken in the prediction process. A K value that is too small might lead to overfitting, while an overly large

Table 2: Training and test data

Label/class/category	Amount of data	Information protein content (g/L)
–	6	0
+–	24	0.15
+	10	0.3
++	22	1
+++	30	3
++++	7	20

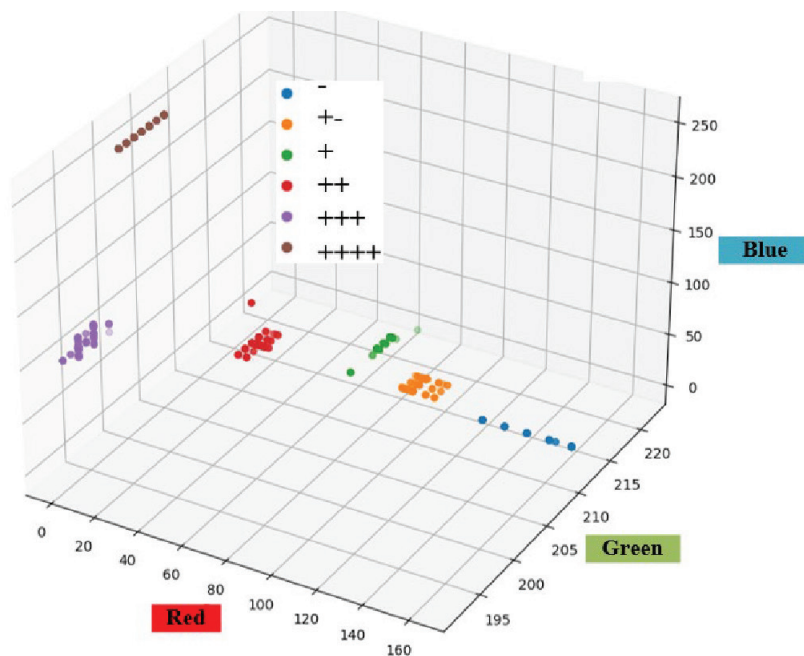


Figure 4: Distribution of 30 test data.

K value might lead to underfitting. Distance measurement calculates the distance between points in the feature space. The choice of this metric can influence KNN results; in this study, the Euclidean distance measurement was used. Feature normalization is vital as KNN is sensitive to feature scaling; hence, normalization or standardization of features might be needed before training the model. Data computation means that KNN requires distance calculation for every pair of data, which can be computationally intensive with large datasets.

In this study, once introduced into the urine test strip, protein solution samples were tested within the created prototype system. The camera sensor then initialized the data. A reaction urine strip requires 30 s before the camera capturing process begins. Image data are then classified based on RGB. The resulting RGB values are evaluated using the KNN model, displaying predicted protein levels. The evaluation results, which include accuracy, precision, recall, and F1-score, are determined using the confusion matrix. The KNN algorithm in this study used the Euclidean distance method for classification and regression. The first step in the KNN model is to determine the K value, which is the variable number of nearest neighbors taken for the classification process. Setting the K value is a balancing act, as different K values can lead to overfitting or underfitting. Lower K values have higher variance, while larger K values can lead to biased classification or lower variance. In this study, the classification parameter, number of neighbors (K), was set as $K = 3$, $K = 10$, and $K = 20$. The performance of the KNN model based on these K values (3, 10, and 20) is shown in Table 3.

Based on Table 3, the KNN model classification system for protein detection can be used to predict data. Overall, it is used to classify protein data into the categories negative (–), trace (+–), positive 1 (+), positive 2 (++) class, positive 3 (+++) class, and positive 4 (++++). Furthermore, the KNN model classification can be realized in the form of a detailed confusion matrix for multiclassification, as shown in Figure 5 for the value of $K = 3$, Figure 6 for the value of $K = 10$, and Figure 7

for the value of $K = 20$. The blue color in the confusion matrix shows the amount of data in the resulting matrix column. The confusion matrix shows that the darker the blue color, the more test data there is in the matrix column, and the more correct data it shows. This confusion matrix makes it easier to read the comparative distribution of expected and predicted data. A confusion matrix is used to determine the suitability of truth and test data.

Based on Figure 5, the x-axis represents the prediction results from the KNN model, and the y-axis represents the classification results using the KNN model. The distribution consists of 30 data points across all classes, which include the negative (–) class, trace (+–) class, positive 1 (+) class, positive 2 (++) class, positive 3 (+++) class, and positive 4 (++++). The number of neighbors is set to 3 ($K = 3$), which means the average of the three nearest neighbors to the test data is calculated. The K value here indicates that the higher the K value, the more neighboring data points are considered in the calculation. Based on the study results, $K = 3$ showed excellent evaluation results with an accuracy rate of 96.7%, precision of 97.00%, recall of 96.7%, and an F1-score of 96.2%. In the test with $K = 3$, one error or misclassified data point was in the upper right.

The combination of the KNN algorithm with digital image processing improves the accuracy of protein detection in urine samples. This study shows that KNN, especially with a value of $K = 3$, provides the best results with an accuracy of 96.7%. The advantages of KNN are its ability to handle non-linear data and provide fast predictions based on the closest distance, making it an ideal tool for clinical applications. Subsequent tests were conducted for $K = 10$ and $K = 20$. The test results show that at $K = 10$, there were five misclassified data points: two data points were (–) but predicted as (++++), and two data points were (+++) but predicted as (++) class. At $K = 20$, there were even more misclassifications: two negative data points were predicted as (++++), three trace (+–) data points were predicted as (++++), and two (++) data points were predicted as (++) class. The selection of

Table 3: Evaluation of the KNN model

K value	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
3	96.7	97.0	96.7	96.2
10	86.7	75.8	86.7	80.7
20	76.7	60.9	76.7	67.3

KNN, K-nearest neighbors.

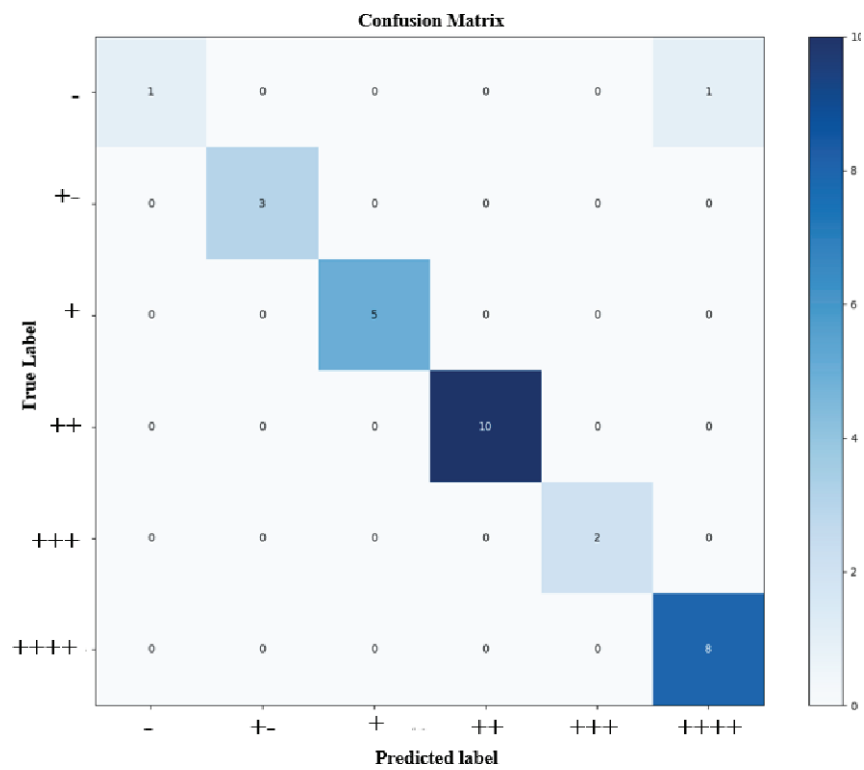


Figure 5: Results of confusion matrix values at $K = 3$.

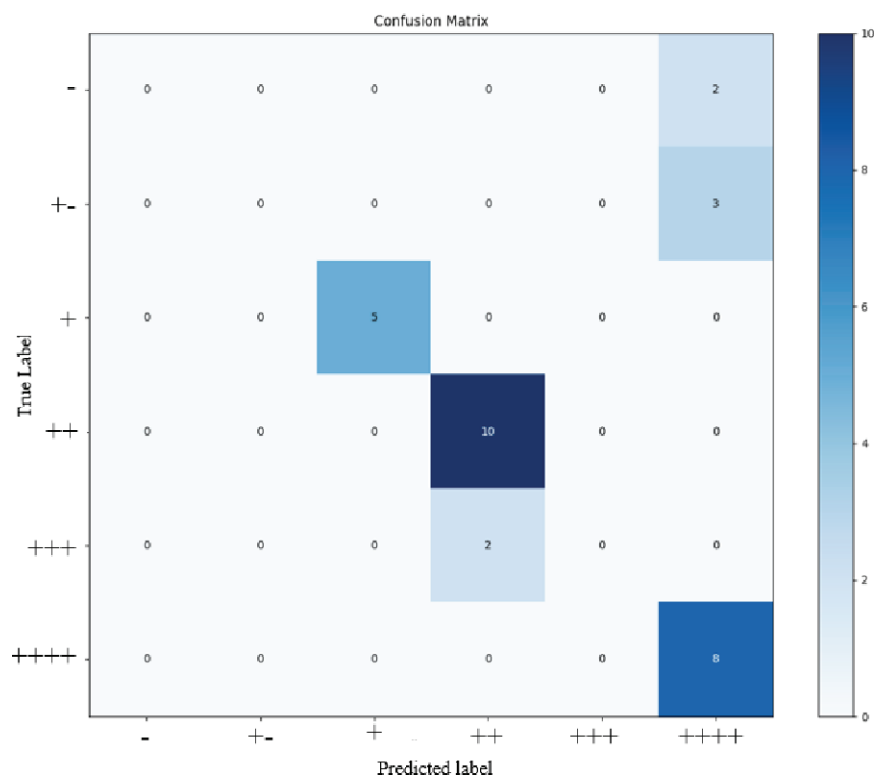


Figure 6: Results of confusion matrix values at $K = 10$.

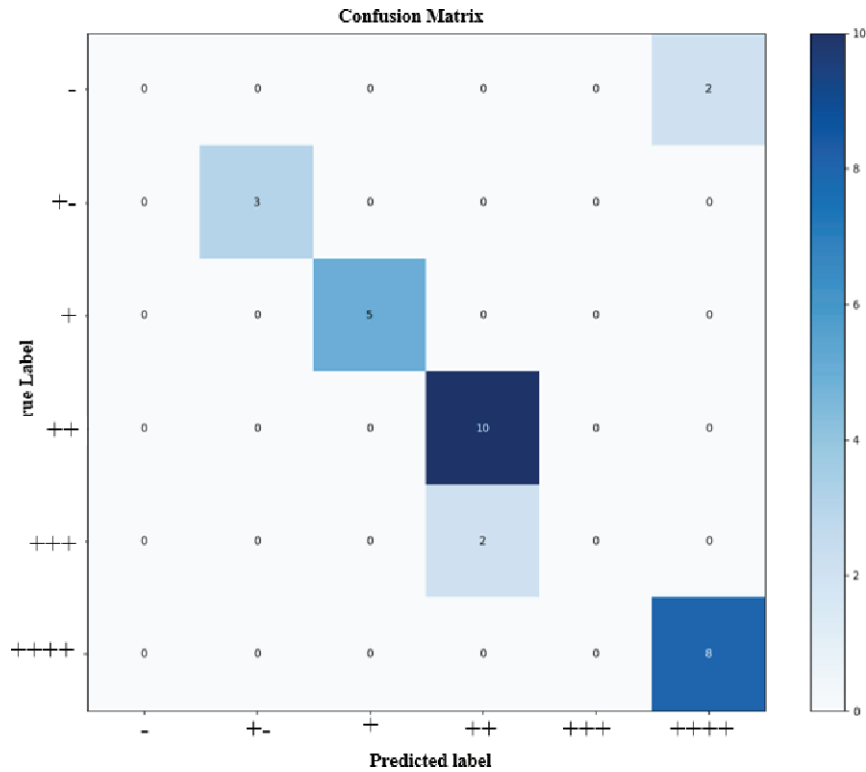


Figure 7: Results of confusion matrix values at $K = 20$.

the K value has a significant impact on the balance between overfitting and underfitting. A small K value (e.g., $K = 3$) provides higher sensitivity but risks overfitting, making the model highly responsive to noise. A larger K value (e.g., $K = 20$) smooths classification but may underfit by ignoring finer variations in protein concentration. This study found that $K = 3$ offered the best accuracy, indicating an optimal balance between sensitivity and generalization. This indicates that as the number of neighbors (K value) increases, the accuracy decreases. From the research results, $K = 10$ had an accuracy rate of 86.7%, precision of 75.8%, recall of 86.7%, and an F1-score of 80.7%, while $K = 20$ had lower results compared to $K = 3$ and $K = 10$ with an accuracy of 76.7%, precision of 60.9%, recall of 76.7%, and an F1-score of 67.3%.

The colors in Figures 5–7, as indicated by the confusion matrix, represent the amount of data in each column. The darker the blue, the more data are present in the corresponding matrix column. This is to make it easier for readers to compare the distribution of comparisons of expected data and predicted results. Comparing machine learning algorithms with other conventional methods provides an opportunity to improve the accuracy and efficiency of analysis,

especially in the analysis of urine biomarkers in metabolic conditions. Machine learning provides a powerful tool for managing chronic diseases through urine biomarker analysis. For example, the detection of micro metabolites in urine using the KNN algorithm can help in the early diagnosis of diabetes and hypertension, thereby increasing the effectiveness of medical interventions.

d. Comparison result

Many researchers have developed proteinuria detection technology in various situations. The obstacles encountered in protein detection are also increasingly complex. Protein detection was initially developed using chemical/dye-based enzyme methods. Turbidimetry, nephelometry, and radioimmunoassay methods have unique advantages in measuring albuminuria. Turbidimetry is suitable for rapid clinical settings, nephelometry provides high sensitivity for low concentrations, while radioimmunoassay offers the most accurate quantitative results, although it requires a specialized laboratory. However, this tool requires professional medical personnel and can only be used in a laboratory setting. Researchers

Table 4: Comparison of research results

No.	Biomarker	Author and year	Color classification	Work principle	Ref.
1.	Albumin	Thakur (2021)	RGB, HSV, and Lab	RF algorithm to estimate albumin concentration using a smartphone	[32]
2.	Albumin	Thakur (2022)	RGB, HSV, and Lab	CNN algorithm for classifying Color in detecting albumin using a smartphone.	[41]
3.	Albumin	Kim (2022)	RGB	RGB extraction uses machine learning and iPhone 11 as a means of detecting color in urine.	[42]
4.	Protein	This study (2023)	RGB	Protein detection equipped with a digital color sensor type ELP camera. Image data are classified based on RGB and evaluated using the KNN algorithm	

CNN, convolutional neural network; KNN, K-nearest neighbors; RF, random forest; RGB, red, green, and blue.

have turned to machine learning technology to process large amounts of data efficiently and accurately, resulting in better predictions and decisions. This research created a protein detection prototype equipped with an ELP camera-type digital color sensor. Next, the camera sensor will initialize the data on the urine strip, which has reacted with the urine sample. The initialization results in the form of image data are then classified based on RGB. The resulting RGB values will be evaluated using the KNN model algorithm, which displays the predicted results of protein levels. Evaluation results in the form of accuracy, precision, recall, and F1-score data will be carried out using the confusion matrix. The results of this research are related to those of previous research, as shown in Table 4.

Based on Table 4, the use of machine learning technology is essential to overcome the problem of determining color in urine. Unpredictable color changes in urine are due to contamination with other substances. Apart from that, analysis using the eyes is less effective because the effects of shadows or ambient light influence it. The difference in this research lies in the use of color retrieval technology, which uses an ELP sensor installed in a series in the protein detection prototype. The ELP camera sensor in the prototype can reduce or minimize shadow effects when capturing colors. Color segmentation for protein detection has been shown to improve diagnostic accuracy. This technology uses a smart camera to capture the color of chemical reactions on a urine strip and analyze its RGB values, using a

machine learning algorithm, such as CNN. A study by Thakur et al. [41] showed 88% accuracy in detecting albumin using this method, making it a practical and affordable solution for healthcare settings.

IV. Conclusion

This study successfully demonstrates the effectiveness of integrating the KNN algorithm with advanced image segmentation techniques for the accurate detection of proteins in urine samples. Protein detection through image segmentation with the KNN algorithm approach has been applied in this research. A prototype design for protein detection has been successfully developed, with the main devices being an ELP-type camera sensor and urine test strips. This prototype helps minimize interference from the effects of shadows and light from outside when taking pictures of urine samples. Image data are classified based on RGB and evaluated using the KNN algorithm based on categories: negative (–), trace (+–), positive 1 (+), positive 2 (++), positive 3 (+++), and positive 4 (++++). From the results of tests carried out with a value of $K = 3$, a value of $K = 10$, and a value of $K = 20$, it can be concluded that the value of K provides the best accuracy value of 96.7% compared to the value of $K = 10$ and the value of $K = 20$. This shows how important it is to select the optimal K value in the KNN algorithm and the role of the confusion matrix as an essential tool for visualizing model performance. The KNN model approach to detecting image-based protein levels in urine is

indeed promising for selecting appropriate parameters. Advances in AI, such as the KNN development algorithm, have not only improved the accuracy and speed of diagnosis but have also expanded access to affordable diagnostic tools, especially through smartphone-based devices.

Acknowledgment

This research received financial support from the Directorate of Research, Technology and Community Service (DRTPM) of the Ministry of Education, Culture, Research and Technology (Kemendikbudristek) of the Republic of Indonesia with Contract No. 181/E5/PG.02.00.PL/2023.

Competing Interests

The authors have stated no conflict of interest.

References

- [1] H. Beng-Ongey, J. S. Robinson, and M. Moxey-Mims, "Chronic kidney disease emerging trends in children and what to do about it," *J. Natl. Med. Assoc.*, vol. 114, no. 3, Supplement 2, pp. S50–S55, 2022, doi: 10.1016/j.jnma.2022.05.002.
- [2] C. A. Herzog *et al.*, "Cardiovascular disease in chronic kidney disease. A clinical update from Kidney Disease: Improving Global Outcomes (KDIGO)," no. July, pp. 572–586, 2011, doi: 10.1038/ki.2011.223.
- [3] D. S. Gardner *et al.*, "Urinary Trace Elements Are Biomarkers for Early Detection of Acute Kidney Injury," *Kidney Int. Reports*, vol. 7, no. 7, pp. 1524–1538, 2022, doi: 10.1016/j.ekir.2022.04.085.
- [4] N. U. Khan *et al.*, "Insights into predicting diabetic nephropathy using urinary biomarkers," *Biochim. Biophys. Acta - Proteins Proteomics*, vol. 1868, no. 10, p. 140475, 2020, doi: 10.1016/j.bbapap.2020.140475.
- [5] M. Z. Bidin, A. M. Shah, J. Stanslas, and C. T. S. Lim, "Blood and urine biomarkers in chronic kidney disease: An update," *Clin. Chim. Acta*, vol. 495, no. April, pp. 239–250, 2019, doi: 10.1016/j.cca.2019.04.069.
- [6] M. Hatada, E. Wilson, M. Khanwalker, D. Probst, J. Okuda-Shimazaki, and K. Sode, "Current and future prospective of biosensing molecules for point-of-care sensors for diabetes biomarker," *Sensors Actuators B Chem.*, vol. 351, no. October 2021, p. 130914, 2022, doi: 10.1016/j.snb.2021.130914.
- [7] R. López-Cortés, B. B. Gómez, S. Vázquez-Estévez, D. Pérez-Fentes, and C. Núñez, "Blood-based protein biomarkers in bladder urothelial tumors," *J. Proteomics*, vol. 247, no. March, 2021, doi: 10.1016/j.jprot.2021.104329.
- [8] A. Yudhana, L. Y. Sabila, A. C. Subrata, H. H. Pratama, and M. S. Akbar, "Non-Invasive Approach for Glucose Detection in Urine Quality Using Its Image Analysis," V. Asadpour and S. Karakuş, Eds. Rijeka: IntechOpen, 2022, p. Ch. 5.
- [9] A. Yudhana, F. Warsino, S. A. Akbar, F. Nuraisyah, and I. Mufandi, "Identification of glucose levels in urine based on classification using -nearest neighbor algorithm method," *Int. J. Smart Sens. Intell. Syst.*, vol. 16, no. 1, 2023, doi: 10.2478/ijssis-2023-0006.
- [10] A. Yudhana and S. A. Akbar, "Self-Health Monitoring Based on Dehydration Level and Glucose Content in Urine using Thermal Sensing," *Int. Conf. Eng. Emerg. Technol.*, no. October, pp. 27–28, 2021.
- [11] A. Yudhana *et al.*, "Multi sensor application-based for measuring the quality of human urine on first-void urine," *Sens. Bio-Sensing Res.*, p. 100461, Oct. 2021, doi: 10.1016/J.SBSR.2021.100461.
- [12] V. Paeder, V. Musi, L. Hvozda, S. Herminjard, and H. P. Herzig, "Detection of protein aggregation with a Bloch surface wave based sensor," *Sensors Actuators, B Chem.*, vol. 157, no. 1, pp. 260–264, 2011, doi: 10.1016/j.snb.2011.03.060.
- [13] R. Z. Alicic, M. T. Rooney, and K. R. Tuttle, "Diabetic Kidney Disease," *Clin. J. Am. Soc. Nephrol.*, vol. 12, no. 18, 2017, doi: 10.2215/CJN.11491116.
- [14] E. Nah, S. Cho, S. Kim, and H. Cho, "Comparison of Urine Albumin-to-Creatinine Ratio (ACR) Between ACR Strip Test and Quantitative Test in Prediabetes and Diabetes," *Ann. Lab. Med.*, vol. 37, pp. 28–33, 2017.
- [15] J. I. Park, H. Baek, B. R. Kim, and H. H. Jung, "Comparison of urine dipstick and albumin: creatinine ratio for chronic kidney disease screening: A population-based study," *PLoS One*, vol. 12, no. 2, pp. 1–12, 2017, doi: 10.1371/journal.pone.0171106.
- [16] K. International, "KDIGO clinical practice guidelines for the diagnosis, evaluation, prevention, and treatment of chronic kidney disease-mineral and bone disorder (CKD-MBD)," *Kidney Int.*, vol. 31, no. 2, pp. 1–150, 2013, doi: 10.1038/ki.2009.188.
- [17] G. R. D. Jones, "Laboratory reporting of urine protein and albumin," *Clin. Biochem. Rev.*, vol. 32, no. 2, pp. 103–107, 2011.
- [18] A. Bellasi, L. Di Lullo, and B. Di Iorio, "Chronic Kidney Disease: The Silent Epidemic," *J. Clin. Med.*, vol. 8, pp. 1–8, 2019, doi: 10.3390/jcm8111795.
- [19] A.-S. Bargnoux *et al.*, "Evaluation of five immunoturbidimetric assays for urinary albumin quantification and their impact on albuminuria categorization," *Clin. Biochem.*, vol. 47, no. 16, pp. 250–253, 2014, doi: 10.1016/j.clinbiochem.2014.07.014.
- [20] J. W. Brinkman *et al.*, "Which method for quantifying urinary albumin excretion gives what outcome? A comparison of immunonephelometry with HPLC," *Kidney Int.*, vol. 66, pp. S69–S75, 2004, doi: 10.1111/j.1523-1755.2004.09219.x.

- [21] A. R. Kahkoska et al., "The early natural history of albuminuria in young adults with youth-onset type 1 and type 2 diabetes," *J. Diabetes Complicat.*, vol. 32, no. 12, pp. 1160–1168, 2019, doi: 10.1016/j.jdiacomp.2018.09.018.The.
- [22] M. E. Haas et al., "Genetic Association of Albuminuria with Cardiometabolic Disease and Blood Pressure," *Am. J. Hum. Genet.*, vol. 103, no. 4, pp. 461–473, 2018, doi: 10.1016/j.ajhg.2018.08.004.
- [23] J. R. Mejia et al., "Diagnostic accuracy of urine dipstick testing for albumin-to-creatinine ratio and albuminuria: A systematic review and meta-analysis," *Heliyon*, vol. 7, no. 11, 2021, doi: 10.1016/j.heliyon.2021.e08253.
- [24] H. Ketha and R. J. Singh, "Quantitation of Albumin in Urine by Liquid Chromatography Tandem Mass Spectrometry," *Clin. Appl. Mass Spectrom. Biomol. Anal.*, pp. 31–36, 2016, doi: 10.1007/978-1-4939-3182-8_4.
- [25] W. Laiwattanapaisai et al., "On-Chip Immunoassay for Determination of Urinary Albumin," *Sensors*, vol. 1, pp. 10066–10079, 2009, doi: 10.3390/s91210066.
- [26] M. A. Cassia et al., "Proteinuria and albuminuria at point of care," *Nephrology*, vol. 2, no. 1, pp. 8–16, 2016, doi: 10.5301/pocj.5000194.
- [27] D. Zhang et al., "Protein detecting with smartphone-controlled electrochemical impedance spectroscopy for point-of-care applications," *Sensors Actuators B Chem.*, vol. 222, pp. 994–1002, 2016, doi: 10.1016/j.snb.2015.09.041.
- [28] M. N. Azhar, A. Bustam, and F. S. Naseem, "Improving the reliability of smartphone-based urine colorimetry using a colour card calibration method," *Digit. Heal.*, vol. 9, pp. 1–11, 2023, doi: 10.1177/20552076231154684.
- [29] T. Wang, H. Huang, R. Wang, H. Zhou, and P. Luo, "Talenta A feasible image-based colorimetric assay using a smartphone RGB camera for point-of-care monitoring of diabetes," *Talanta*, vol. 206, no. August 2019, p. 120211, 2020, doi: 10.1016/j.talanta.2019.120211.
- [30] T. Zeng et al., "Application of machine learning algorithms to screen potential biomarkers under cadmium exposure based on human urine metabolic profiles," *Chinese Chem. Lett.*, vol. 33, pp. 5184–5188, 2022, doi: 10.1016/j.cclet.2022.03.020.
- [31] R. Thakur, P. Maheshwari, S. Kumar Datta, and S. Kumar Dubey, "Smartphone-based, automated detection of urine albumin using deep learning approach," *Meas. J. Int. Meas. Confed.*, vol. 194, no. January, 2022, doi: 10.1016/j.measurement.2022.110948.
- [32] R. Thakur, P. Maheshwari, and S. K. Datta, "Machine Learning-Based Rapid Diagnostic-Test Reader for Albuminuria Using Smartphone," *IEEE Sens. J.*, vol. 21, no. 13, pp. 14011–14026, 2021, doi: 10.1109/JSEN.2020.3034904.
- [33] A. F. Coskun, "Albumin testing in urine using a smart-phone," *NIH Public Access*, vol. 13, no. 21, pp. 4231–4238, 2014, doi: 10.1039/c3lc50785h.Albumin.
- [34] S. Bhatt, S. Kumar, M. K. Gupta, S. K. Datta, and S. K. Dubey, "Colorimetry-based and smartphone-assisted machine-learning model for quantification of urinary albumin," *Meas. Sci. Technol.*, vol. 35, no. 1, pp. 1–14, 2023, doi: 10.1088/1361-6501/acfd4c.
- [35] F. H. Juwono, W. K. Wong, H. T. Pek, S. Sivakumar, and D. D. Acula, "Ovarian cancer detection using optimized machine learning models with adaptive differential evolution," *Biomed. Signal Process. Control*, vol. 77, no. April, p. 103785, 2022, doi: 10.1016/j.bspc.2022.103785.
- [36] J. Yu et al., "Using metabolomics and proteomics to identify the potential urine biomarkers for prediction and diagnosis of gestational diabetes," *eBioMedicine*, vol. 101, p. 105008, 2024, doi: 10.1016/j.ebiom.2024.105008.
- [37] R. K. Halder et al., "ML-CKDP: Machine learning-based chronic kidney disease prediction with smart web application," *J. Pathol. Inform.*, vol. 15, no. December 2023, p. 100371, 2024, doi: 10.1016/j.jpi.2024.100371.
- [38] K. Hema, K. Meena, and R. Pandian, "Analyze the impact of feature selection techniques in the early prediction of CKD," *Int. J. Cogn. Comput. Eng.*, vol. 5, no. May 2023, pp. 66–77, 2024, doi: 10.1016/j.ijcce.2023.12.002.
- [39] A. S. Zamani, A. H. A. Hashim, A. S. A. Shatat, M. M. Akhtar, M. Rizwanullah, and S. S. I. Mohamed, "Implementation of machine learning techniques with big data and IoT to create effective prediction models for health informatics," *Biomed. Signal Process. Control*, vol. 94, p. 106247, 2024, doi: 10.1016/j.bspc.2024.106247.
- [40] M. Mustafizur Rahman, M. Al-Amin, and J. Hossain, "Machine learning models for chronic kidney disease diagnosis and prediction," *Biomed. Signal Process. Control*, vol. 87, p. 105368, 2024, doi: 10.1016/j.bspc.2023.105368.
- [41] R. Thakur, P. Maheshwari, S. Kumar, and S. Kumar, "Smartphone-based, automated detection of urine albumin using deep learning approach," *Measurement*, vol. 194, no. January, p. 110948, 2022, doi: 10.1016/j.measurement.2022.110948.
- [42] S. C. Kim and Y. S. Cho, "Predictive System Implementation to Improve the Accuracy of Urine Self-Diagnosis with Smartphones: Application of a Confusion Matrix-Based Learning Model through RGB Semiquantitative Analysis," *Sensors*, vol. 22, no. 14, 2022, doi: 10.3390/s22145445.