Enhancing The Detection Of Chronic Kidney Disease Using Machine Learning by Denoising and Sparse Feature Approach with Biomarkers

Dr. Santhoshkumar Sundar [1], Dr. S.K. Manju Bargavi [2], Dr. A. Thasil Mohamed [3]*

[1] Post-Doctoral Research Scholar,

Lincoln University College, Petaling Jaya, Malaysia, and Assistant Professor, Department of Computer Science, Alagappa University, Karaikudi, Tamil Nadu, India

Email: santhoshkumars@alagappauniversity.ac.in

[2] Professor

School of Computer Science and IT

Jain (Deemed-to-be University)

Bangalore

Email: cloudbargavi@gmail.com

[3]Technical Architect,

Verizon Communications Inc., Texas, USA.

Email: ameerthasil@gmail.com

Abstract: Enhancing the detection of Chronic Kidney Disease (CKD) using machine learning involves applying denoising techniques and sparse feature selection to identify key biomarkers, improving early diagnosis accuracy. CKD poses several challenges due to the complex and multifactorial nature of the disease. Traditional diagnostic methods often fail to identify early-stage CKD, resulting in delayed intervention and poor patient outcomes. The study aims to create a reliable and precise machine-learning framework that leverages biomarker data to effectively classify patients with Chronic Kidney Disease (CKD). Collect blood and/or urine samples to measure a comprehensive panel of biomarkers, including Creatinine, glomerular filtration rate (GFR), and other standard clinical markers. Utilize unsupervised learning algorithms like Denoising Sparse Auto-Encoder (DSAE) to sift through medical data, looking for hidden patterns indicative of CKD. Apply Term Frequency - Inverse Document Frequency (TF-IDF) to identify the most important pieces of information in the data, focusing on key elements like changes in blood tests or abnormalities in scans crucial for CKD diagnosis. Implement Sequential Selection Ratio (SSR) and Feature Normalization Component (FNC) techniques for feature selection and normalization. Utilize an Adaptive Backpropagation Neural Network (ABPNN) as the main classification model, adjusting its learning rate during training to improve performance

Keywords: Chronic Kidney Disease, Glomerular Filtration Rate, Adaptive Backpropagation Neural Network, Sequential Selection Ratio, Feature Normalization Component, Denoising Sparse Auto-Encoder.

Introduction

Chronic kidney disease (CKD) poses a significant health burden globally, affecting millions of individuals and leading to serious complications if left undiagnosed or untreated [1]. Early detection and classification of CKD are crucial for effective management and intervention strategies to mitigate its progression [2]. In recent years, there has been growing interest in leveraging biomarkers and machine learning techniques for CKD classification, offering the potential for improved accuracy, efficiency, and

personalized diagnostics [3]. This study aims to explore the integration of biomarker-driven approaches with machine learning algorithms for CKD classification, examining their effectiveness, challenges, and implications for clinical practice [4-5]. Biomarker-driven machine-learning approaches offer a promising avenue for enhancing CKD classification by utilizing a diverse range of molecular, cellular, and clinical markers to predict disease progression and severity [6-7]. These methods leverage the vast amount of data generated from various biomarker assays, clinical tests, and patient records to develop predictive models that can stratify individuals into different CKD stages with greater accuracy and precision [8]. By integrating advanced machine learning algorithms such as deep learning, support vector machines, and random forests with biomarker data, researchers aim to uncover hidden patterns, correlations, and predictive features that may not be readily apparent through traditional diagnostic methods. Through this interdisciplinary approach, this study seeks to contribute to the ongoing efforts to improve CKD diagnosis, prognosis, and personalized treatment strategies, ultimately leading to better patient outcomes and healthcare delivery [9-10].

Early detection of CKD is crucial to prevent severe complications such as kidney failure, cardiovascular disease, and increased mortality [11-12]. However, the current diagnostic methods rely heavily on traditional biomarkers and clinical observations, which may lack sensitivity, especially in the early stages of the disease [13-14]. This study aims to enhance the detection of CKD using machine learning techniques by integrating a denoising and sparse feature approach with biomarkers. Traditional datasets often contain noise and redundant features, which can hinder the performance of machine learning models. By employing a denoising technique, irrelevant or noisy data can be filtered out, ensuring that only the most pertinent features are considered [15]. Furthermore, a sparse feature selection approach will be utilized to identify and prioritize critical biomarkers, reducing dimensionality and improving model accuracy [16-17]. The motivation for this research stems from the need for more precise, early-stage detection of CKD, which can lead to timely interventions and better patient outcomes [18-19]. By leveraging machine learning, this study aims to develop a robust predictive model that could significantly impact public health by providing an efficient, cost-effective tool for CKD diagnosis [20]. The objective of the study is to develop a robust and accurate machine-learning framework that utilizes biomarker data to classify patients with Chronic Kidney Disease (CKD). The remaining sections are arranged as follows: The literature review was described in Section 2, the proposed technique was described in Section 3, the results were discussed in Section 4, and the paper's conclusion was described in Section 5.

Literature Survey

The literature survey aims to provide a comprehensive overview of existing research studies and developments in the field of chronic kidney disease (CKD) classification. Nageeta et al [21] examine the potential of precision medicine in managing Diabetic kidney disease (DKD). It embodies a patient-centric and scientifically sophisticated form of care that promises to transform the treatment of DKD and improve patient outcomes. Legrand et al [22] discuss the role of enrichment strategies to target populations that are most likely to derive benefit and the importance of patient-centred clinical trial endpoints and appropriate trial designs to guide in designing future trials. Findings show that the incidence of MAKE was 14.3% in the balanced fluid group versus 15.4% in the saline group [odds ratio (OR) 0.90; 95% CI 0.82–0.99]. Brown et al [23] determination of the present report is to provide perioperative teams with expert recommendations specific to cardiac surgery-associated AKI (CSA-AKI). Based on available evidence and

group consensus, a total of 13 recommendations were formulated. Duca et al [24] explored the potential correlation between Holter ECG parameters and comorbidities in individuals with ischemic cardiomyopathy experiencing heart failure (HF), with a particular focus on the primary utility of these parameters as prognostic indicators. Significant associations were uncovered between diabetes and unconventional physiological indicators, specifically the Triangular index ($p = 0.035$) and deceleration capacity ($p = 0.002$). De Bakker et al [25] derived a small set out of 4210 repeatedly measured proteins, which, along with clinical characteristics and established biomarkers, carry optimal prognostic capacity for adverse events, in patients with HFrEF. [Hazard ratio (95% confidence interval): 1.96 (1.17–3.40) and 0.66 (0.49–0.88), respectively].

Dovrolis et al [26] explored the diversity in mRNA expression profiles of inflammation and immunity-related circulating genes for the development of biomarkers that could predict the effectiveness of immunotherapy-based treatments using ICIs for individuals with mccRCC. The results show that gene expression can be used to classify these samples with high accuracy and specificity. Swain et al [27] developed a machine-learning model that can use publicly available data to forecast the occurrence of chronic kidney disease. Out of all the applied learning techniques, support vector machine (SVM) and random forest (RF) achieved the lowest false-negative rates and test accuracy, equal to 99.33% and 98.67%, respectively. Islam et al [28] examined the potential of several different machine-learning approaches for providing an early diagnosis of CKD has been investigated. The findings show that the greatest performance indicators are an accuracy of 0.983, a precision of 0.98, a recall of 0.98, and an F1-score of 0.98 for the XgBoost classifier. Hussain et al [29] introduced a deep learning (DL) clinical diagnostic system aimed at enhancing the automatic identification and classification of CKD. This refined iteration yields impressive test performances in terms of accuracy: 99.98%, recall: 99.89%, precision: 99.84%, F1 score: 99.86%, specificity: 99.84%, Micro AUC: 99.99%, and testing time of 0.0641 seconds per image. Rehman et al [30] present a chronic kidney disease (CKD) prediction model to classify CKD patients from NCKD (Non-CKD). The accuracies of LR were 98.5% and 97.5% for train & test datasets. While LDA accuracy was 96.07% and 96.6% for train and test datasets. Likewise, MLP attained 95% and 94.1% accuracy for training and test datasets.

Proposed Research Methodology

The proposed research methodology aims to leverage a multifaceted approach to advance the diagnosis and classification of Chronic Kidney Disease (CKD). By integrating cutting-edge machine learning techniques with comprehensive biomarker analysis, this methodology seeks to extract valuable insights from patient data to improve CKD diagnosis accuracy and patient outcomes. The methodology will encompass data acquisition, pre-processing, feature extraction, machine learning model development, and evaluation, with a focus on harnessing the power of biomarkers and machine learning algorithms to enhance CKD classification capabilities.

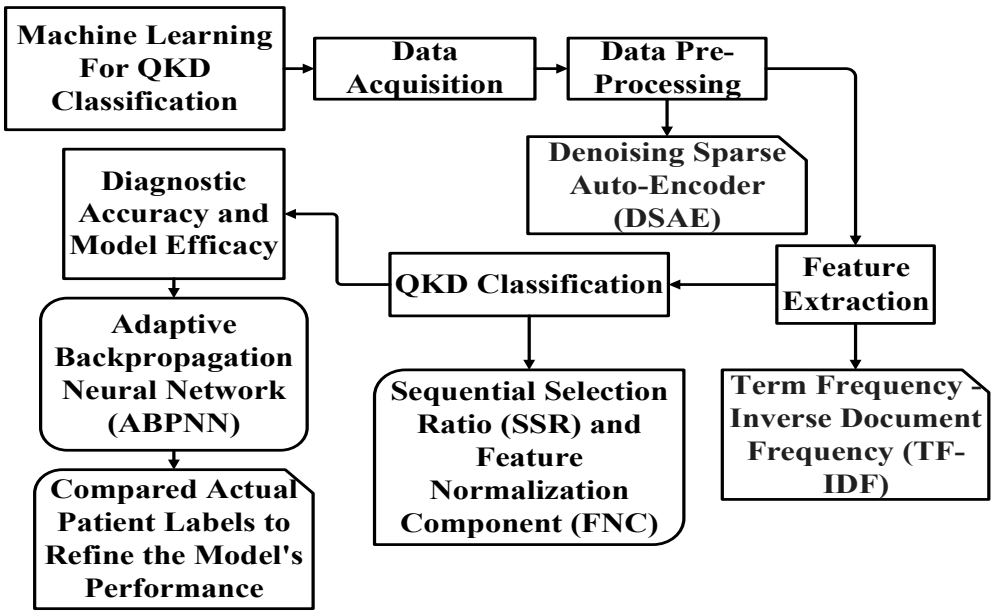

Figure 1: Block Diagram of the Proposed Work

Figure 1 shows the block diagram of the proposed work. Collaborate with clinical institutions to recruit patients diagnosed with Chronic Kidney Disease (CKD). Utilize unsupervised learning algorithms like Denoising Sparse Auto-Encoder (DSAE) to sift through medical data, looking for hidden patterns indicative of CKD. Apply Term Frequency - Inverse Document Frequency (TF-IDF) to identify the most important pieces of information in the data, focusing on key elements like changes in blood tests or abnormalities in scans crucial for CKD diagnosis. Implement Sequential Selection Ratio (SSR) and Feature Normalization Component (FNC) techniques for feature selection and normalization. Utilize an Adaptive Backpropagation Neural Network (ABPNN) as the main classification model, adjusting its learning rate during training to improve performance.

(a) Data Acquisition

To enhance the detection of CKD using machine learning, collaboration with clinical institutions is essential to recruit patients with confirmed CKD diagnoses. This collaboration will provide access to a diverse patient cohort, ensuring a comprehensive dataset that reflects various stages and types of CKD. Blood and urine samples will be collected from participants to measure a wide range of biomarkers, including standard clinical markers such as creatinine and estimated glomerular filtration rate (eGFR), which are commonly used in CKD diagnosis and monitoring. In addition to these traditional markers, the study will explore and incorporate novel biomarkers that have shown potential relevance to CKD. These may include inflammatory markers like C-reactive protein (CRP) and interleukin-6 (IL-6), which are associated with kidney inflammation and damage. Genetic markers linked to CKD risk, such as specific single nucleotide polymorphisms (SNPs), will also be included to provide a more comprehensive understanding of the disease's progression. The integration of these diverse biomarkers will enable the development of a more accurate and robust machine-learning model, improving the early detection and classification of CKD, ultimately leading to more personalized and effective treatment strategies for patients.

(b) Data Pre-processing and Feature Extraction

To enhance the detection of Chronic Kidney Disease (CKD), unsupervised learning algorithms like Denoising Sparse Auto-Encoder (DSAE) will be utilized to process medical data and uncover hidden patterns associated with the disease. DSAE is particularly effective in managing high-dimensional data, such as biomarker panels, by reducing noise and retaining critical features that contribute to accurate classification. This approach allows the model to learn a compressed representation of the data, which emphasizes the most relevant factors and patterns that may indicate the presence or progression of CKD. Term Frequency-Inverse Document Frequency (TF-IDF) will be applied to further refine the data analysis process by identifying the most important elements within the dataset. Although traditionally used in text mining, TF-IDF can be adapted to highlight key features in clinical data, such as specific changes in blood tests, like elevated creatinine or reduced glomerular filtration rate (GFR), and subtle abnormalities in imaging scans. By focusing on these critical components, the combination of DSAE and TF-IDF will enhance the model's ability to distinguish between CKD and non-CKD cases, leading to more precise and timely diagnosis. This dual approach will improve the robustness of the detection system and support early intervention and personalized treatment planning for CKD patients.

(i) Denoising Sparse Auto-Encoder (DSAE)

The early detection of Chronic Kidney Disease (CKD) is critical for effective treatment and improved patient outcomes. This study introduces a machine learning framework that leverages Denoising Sparse Auto-Encoder (DSAE) to enhance the detection of CKD using biomarkers. By denoising and extracting sparse, high-quality features from clinical data, this approach aims to improve the precision and reliability of CKD predictions. The integration of DSAE with biomarkers allows for more robust modelling, minimizing noise while capturing essential patterns, ultimately contributing to earlier diagnosis and better management of CKD.

During the encoding step, an input vector $\mathrm{x}^{\mathrm{i}} \in \mathbb{R}^{\mathrm{C}}$ is processed by applying a linear deterministic mapping and a nonlinear activation function l as follows:

$$\alpha^{\mathrm{i}} = \mathrm{f}(\mathrm{x}^{\mathrm{i}}) = \mathrm{l}(\mathrm{W}_1\mathrm{x}^{\mathrm{i}} + \mathrm{b}_1) \quad (1)$$

Where $\mathrm{W}_1 \in \mathbb{R}^{\mathrm{K}\times\mathrm{C}}$ is a weight matrix with K features and $\mathrm{b1} \in \mathrm{RK}$ is the encoding bias. In this study, we consider a leaky rectified linear unit (LReLU) activation function for $\mathrm{l}(\mathrm{x})$. It can be represented as

$$\mathrm{y} = \begin{cases} \mathrm{x}, & \mathrm{if} \mathrm{x} \geq 0 \\ \omega\mathrm{x}, & \mathrm{if} \mathrm{x} \leq 0 \end{cases} \quad (2)$$

The slope ω of the LReLU is set to 0.01. Then we decode a vector using a separate linear decoding matrix

$$\mathrm{z}^{\mathrm{i}} = \mathrm{W}_2\alpha^{\mathrm{i}} + \mathrm{b}_2 \quad (3)$$

Where $\mathrm{W}_2 \in \mathbb{R}^{\mathrm{K}\times\mathrm{C}}$ and $\mathrm{b}_2 \in \mathbb{R}^{\mathrm{C}}$ are a decoding weight matrix and a bias vector, respectively

$$\mathrm{L}(\mathrm{X},\mathrm{Z}) = \frac{1}{2}\sum_{\mathrm{i}=1}^{\mathrm{M}}\left\|\mathrm{x}^{\mathrm{i}} - \mathrm{z}^{\mathrm{i}}\right\|^2 + \frac{\lambda}{2}\|\mathrm{W}\|^2 \quad (4)$$

Where X and Z represent the training and reconstructed data, respectively. $\mathrm{KL}(\rho \parallel \hat{\rho})$ is the sparse penalty term, which can be denoted as the following formula:

$$\mathrm{KL}(\rho \parallel \hat{\rho}) = \rho\log\frac{\rho}{\rho} + (1-\rho)\log\frac{1-\rho}{1-\rho} \quad (5)$$

Where $\mathrm{KL}(\cdot)$ is the Kullback–Leibler divergence. We recall that α denotes the activation of hidden units in the autoencoder; let $\hat{\rho} = (1/\mathrm{M}) \sum_1^{\mathrm{M}}[\alpha^{(\mathrm{i})}]$ be the average activation of α averaged over the training set $\mathrm{X}^{\mathrm{C}\times\mathrm{M}}$. Then our objective function in sparse autoencoder learning can be written as follows:

$$\mathrm{L}(\mathrm{X},\mathrm{Z}) + \beta \sum_{\mathrm{j}=1}^{\mathrm{K}} \mathrm{KL}(\rho \parallel \hat{\rho}) \qquad (6)$$

With the introduction of the KL divergence weighted by a sparsity penalty parameter β in the objective function, we penalize a large average activation of α over the training samples by setting ρ small.

Table 1: Algorithm for DSAE

Algorithm 1: DSAE
(1) **Input:**
(2) Training set X
(3) Weight decay parameter λ, weight of sparse penalty term β, sparse parameter ρ
(4) **Procedure:**
(5) Initialize parameters $(\mathrm{W}_1, \mathrm{b}_1), (\mathrm{W}_2, \mathrm{b}_2)$
(6) Get $\overline{\mathrm{x}^{\mathrm{i}}}$ by stochastic corrupting the input vector x^{i} .
(7) **FOR** $\mathrm{j} = 1$ to T do
(8) $\mathrm{Loss} = \mathrm{L}(\mathrm{X},\mathrm{Z}) + \beta \sum \mathrm{K}\, \mathrm{j} = 1\, \mathrm{KL}(\rho \parallel \hat{\rho})$
(9) Use the L-BFGS algorithm [58] to update $(\mathrm{W}_1, \mathrm{b}_1), (\mathrm{W}_2, \mathrm{b}_2)$
(10) **ENDFOR**
(11) **Output:** $(\mathrm{W}_1, \mathrm{b}_1)$ Which is utilized for convolution kernels

Algorithm 1 shows denoising Sparse Auto-Encoder (DSAE) is designed to enhance the detection of Chronic Kidney Disease (CKD) by efficiently extracting relevant, sparse features from clinical biomarker data. The algorithm begins by taking the input training set X and initializing the parameters, including weights and biases $(\mathrm{W}_1, \mathrm{b}_1), (\mathrm{W}_2, \mathrm{b}_2)$, as well as key hyperparameters such as the weight decay parameter λ, sparse penalty weightβ, and sparse parameter ρ. A stochastic corruption process is applied to the input vectors, introducing controlled noise to improve robustness. During training, the loss function is computed by combining reconstruction loss $\mathrm{L}(\mathrm{X},\mathrm{Z})$ with a sparsity constraint using the Kullback-Leibler (KL) divergence. This encourages the extraction of sparse, meaningful features. The Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm is then employed to update the model parameters iteratively. After training, the optimized weights and biases $(\mathrm{W}_1, \mathrm{b}_1)$ are utilized for convolution kernels, allowing the model to enhance CKD detection by learning denoised and sparse representations of the biomarker data.

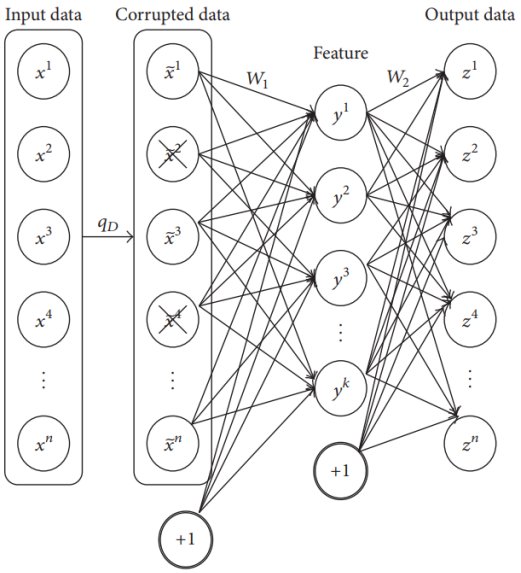

Figure 2: Illustration of a Single-Layer DSAE

Figure 2 illustrates a single-layer Denoising Sparse Auto-Encoder (DSAE) used for enhancing Chronic Kidney Disease (CKD) detection. In this model, input data, consisting of biomarkers, is first corrupted with noise to create a partially obscured version. The autoencoder then reconstructs the original input by learning sparse representations, where only a small number of neurons are active. This sparse coding encourages the network to capture essential patterns in the biomarker data while filtering out noise. The hidden layer captures crucial features, which are refined and used for CKD detection, improving the model's robustness and accuracy.

The Denoising Sparse Auto-Encoder (DSAE) is a specialized machine-learning model aimed at enhancing Chronic Kidney Disease (CKD) detection by efficiently processing clinical biomarker data. This approach integrates two key techniques: denoising and sparse feature extraction. First, DSAE introduces controlled noise to the input data, which helps the model learn more robust, generalized features by focusing on essential patterns and ignoring irrelevant noise. Secondly, the sparse feature extraction encourages the model to identify only the most critical features from the biomarkers, reducing complexity and improving accuracy. By leveraging this combination, DSAE strengthens the ability of machine learning models to detect CKD early and accurately, making it a powerful tool for clinical diagnosis and management.

(ii) Term Frequency - Inverse Document Frequency (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) is a widely used technique in natural language processing and information retrieval that quantifies the importance of a word within a document relative to a collection of documents. In the context of enhancing Chronic Kidney Disease (CKD) detection using machine learning, TF-IDF can play a crucial role in feature extraction from biomarker data. By identifying significant patterns and distinguishing relevant biomarkers, TF-IDF helps optimize sparse feature selection, improving the efficiency of machine learning models. This method supports denoising techniques to enhance the accuracy of CKD detection while minimizing irrelevant data interference.

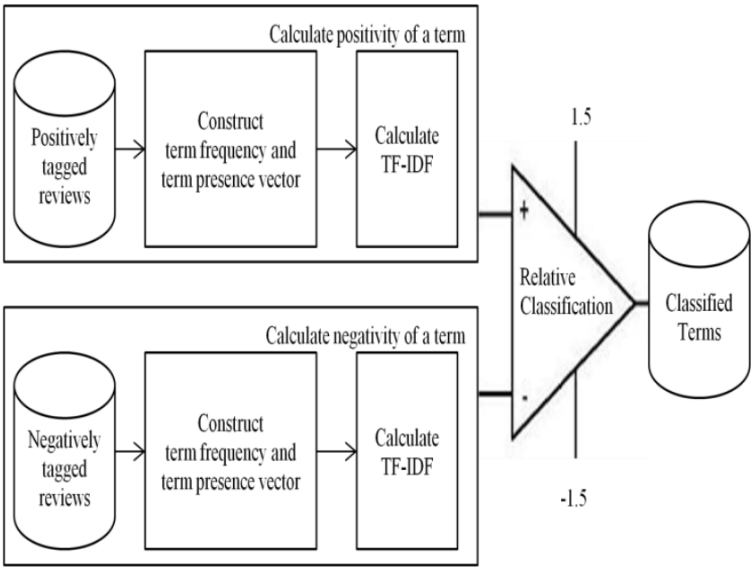

Figure 3: Flow chart of TF-IDF

Figure 3 illustrates a flow chart for a Term Frequency-Inverse Document Frequency (TF-IDF) approach, focusing on the classification of terms based on positively and negatively tagged reviews. In the context of Chronic Kidney Disease (CKD) detection using machine learning with denoising and sparse feature selection, this process could be applied to the classification of biomarkers. First, both positively and negatively tagged biomarkers are processed by constructing term frequency and presence vectors, followed by TF-IDF calculation to weigh the significance of each biomarker. These values are then fed into a relative classification process, which assigns weights (e.g., +1.5 or -1.5) to determine the most relevant biomarkers, thus enabling the model to identify key features that contribute to CKD detection while ignoring noise. The system improves feature sparsity, facilitating more accurate CKD prediction through machine learning models.

In the first part term presence vector and term frequency vector are constructed in (9) and (10) for positively tagged documents using (7) and (8).

$$\mathrm{TFP} = \begin{bmatrix} c_{11} & \cdots & c_{1d} \\ \vdots & \ddots & \vdots \\ c_{t1} & \cdots & c_{td} \end{bmatrix} \tag{7}$$

Where,

TFP=Term Frequency Matrix for positively tagged documents

t = term d = document

$\mathrm{TF}[i][j] = c_{ij}$ = count of term i in document j

$$\mathrm{TPP} = \begin{bmatrix} p_{11} & \cdots & p_{1d} \\ \vdots & \ddots & \vdots \\ p_{t1} & \cdots & p_{td} \end{bmatrix} \tag{8}$$

Where,

TPP = Term Presence Matrix for positively tagged documents.

t = term. d = document.

$\mathrm{TF}[i][j] = p_{ij}$ = presence of term i in document j

= 1 if term i is present in document j otherwise 0.

$$P_{ctd} = \sum_{j=1}^{d} c_{tj} \quad (9)$$

Where,

P_{ctd} = Frequency of term t in positively tagged documents.

c_{ij} = count of term t in j^{th} document.

$$P_t = \sum_{j=1}^{d} p_{tj} \quad (10)$$

Where,

P_t = Number for positively tagged documents with term t.

P_{tj} = presence of term t document j

In "(8)" TFIDF of the terms is calculated by using the vectors in (9) and (10) of positively tagged documents. This value contributes to the positivity of the terms.

$$Pos_t = P_{ctd} \times \log \frac{P}{P_t} \quad (11)$$

Where,

Pos_t = Positivity of term t.

P_{ctd} = Frequency of term t in positively tagged documents.

P = Total Number of positively tagged documents.

P_t = Number for positively tagged documents with term t.

Negativity of the terms is calculated using TFIDF on negatively tagged documents, and vectors, as in "(9)".

$$Neg_t = N_{ctd} \times \log \frac{N}{N_t} \quad (12)$$

Where,

Neg_t = Negativity of term t.

N_{ctd} = Frequency of term t in negatively tagged documents.

N = Total Number of negatively tagged documents.

N_t = Number for negatively tagged documents with term t.

A term is classified as neutral if its positivity equals negativity.

$$LDT_t = \log \frac{Pos_t + 0.001}{Neg_t + 0.001} \quad (13)$$

Where,

LDT_t = Logarithmic differential TFIDF.

Pos_t = Positivity of term t.

Neg_t = Negativity of term t.

$$Pol_t = \begin{matrix} 0 \\ -1 \end{matrix} \quad \text{if} \quad LDT_t = \begin{matrix} 0 \\ < 0 \end{matrix} \quad (14)$$

Where,

Pol_t = Polarity of term t.

LDT_t = Logarithmic differential TFIDF.

Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical technique used to evaluate the importance of a term within a document relative to a collection of documents. In the context of enhancing the detection of Chronic Kidney Disease (CKD) using machine learning, TF-IDF can be adapted to assess the significance of biomarkers in large datasets. Here, "terms" refer to individual biomarkers, and their frequency reflects how often a biomarker appears in CKD-related cases. By combining term

frequency with inverse document frequency, which reduces the weight of biomarkers that appear frequently across all cases, the technique helps isolate critical biomarkers for CKD diagnosis. When combined with denoising and sparse feature selection approaches, TF-IDF aids in filtering irrelevant or redundant data, optimizing the feature set for improved CKD detection using machine learning models. This refined focus on relevant biomarkers enhances model accuracy and efficiency.

(c) Machine Learning for CKD Classification

To optimize the detection of Chronic Kidney Disease (CKD) using machine learning, Sequential Selection Ratio (SSR) and Feature Normalization Component (FNC) techniques will be implemented to prepare the data effectively for training and classification. The SSR technique focuses on selecting the most informative features from the dataset in a sequential manner, gradually improving the model's performance by reducing irrelevant or redundant features. This step is crucial to enhance the predictive accuracy of the machine learning model by concentrating on the most impactful biomarkers, such as creatinine levels, glomerular filtration rate (GFR), and other key clinical indicators. Feature Normalization Component (FNC) will be employed to standardize the selected features, ensuring they are on a comparable scale and thus preventing any one feature from disproportionately influencing the model's learning process. This normalization is essential for improving convergence during training and reducing the risk of overfitting. After pre-processing, the patient data will be split into training (70-80%) and validation (20-30%) sets. The training set will be used to train the machine learning model, while the validation set will evaluate its performance. The model's predicted CKD labels will be compared against the actual patient labels (healthy or CKD) to assess accuracy, sensitivity, specificity, and other relevant metrics. Based on these results, the model will be refined iteratively to achieve optimal performance in CKD detection, contributing to early diagnosis and improved patient outcomes.

(i) Sequential Selection Ratio (SSR)

The SSR in machine learning for CKD classification involves a strategic approach to feature selection aimed at improving the accuracy of Chronic Kidney Disease (CKD) detection. SSR works by sequentially selecting the most relevant features from a given dataset, focusing on biomarkers and clinical indicators that contribute significantly to differentiating between CKD and non-CKD cases. This method prioritizes features based on their contribution to classification performance, allowing the model to filter out irrelevant or redundant information systematically. By retaining only, the most informative features, SSR reduces the dimensionality of the dataset, which helps prevent overfitting and enhances model interpretability. For CKD classification, applying SSR to select key biomarkers—such as creatinine levels, glomerular filtration rate (GFR), and other novel indicators—ensures that the model focuses on the most pertinent data. This approach can improve the precision, recall, and overall diagnostic performance of machine learning models, enabling more accurate and early detection of CKD.

$$\mathrm{SSR_k} = \frac{\mathrm{R_k}}{\mathrm{N_k}} \tag{15}$$

Where $\mathrm{R_k}$ is the number of relevant features selected at the kth step, and $\mathrm{N_k}$ is the total number of features considered at the k-th step.

$$\mathrm{SSR_s} = \frac{\sum_{i \in S} \mathrm{Weight_i}}{|S|} \tag{16}$$

Where S is a subset of features, $\mathrm{Weight_i}$ is the weight or importance score of the i-th feature, and $|S|$ is the number of features in subset S.

$$SSR_{CV} = \frac{1}{K}\sum_{i=1}^{K}\frac{R_i}{N_i} \quad (17)$$

Where K is the number of cross-validation folds, R_i is the number of relevant features in fold i, and N_i is the total number of features in fold i.

$$SSR_{acc} = \frac{Accuracy_S - Accuracy_{base}}{Accuracy_{max} - Accuracy_{base}} \quad (18)$$

Where $Accuracy_S$ is the accuracy of the model with feature subset S, $Accuracy_{base}$ is the baseline accuracy, and $Accuracy_{max}$ is the maximum achievable accuracy. By systematically selecting the most relevant features from a comprehensive set of biomarkers, SSR effectively reduces dimensionality and eliminates noise, allowing the model to focus on the most informative variables. This targeted feature selection improves the model's performance by minimizing overfitting and enhancing its ability to generalize to new, unseen data. The use of SSR, combined with robust machine learning algorithms, significantly enhances the predictive accuracy of CKD classification, supporting early diagnosis and facilitating timely intervention. Future work can focus on further refining the SSR technique, incorporating additional biomarkers, and validating the approach across diverse patient populations.

(ii) Feature Normalization Component (FNC)

The FNC plays a critical role in the machine learning pipeline for CKD classification by standardizing the range and distribution of features across the dataset. By applying normalization techniques, FNC ensures that all features contribute equally to the model's performance, preventing bias towards variables with larger numerical ranges. This process involves scaling features to a common range or transforming them to have a standard distribution, thereby enhancing the model's ability to learn and generalize patterns effectively. FNC improves the stability and convergence of machine learning algorithms, leading to more accurate and reliable classification outcomes. In the context of CKD classification, proper feature normalization helps in accurately distinguishing between healthy and CKD patients by ensuring that the model evaluates all biomarkers on a consistent scale. Future research should explore advanced normalization techniques and their impact on model performance to further refine CKD detection capabilities.

$$x_{norm} = \frac{x-\mu}{\sigma} \quad (19)$$

Where x is the original feature value, is the mean of the feature, and σ is the standard deviation of the feature.

$$x_{norm} = \frac{x-x_{min}}{x_{max}-x_{min}} \quad (20)$$

Where x_{min} and x_{max} are the minimum and maximum values of the feature, respectively.

$$x_i = \frac{x-\bar{x}}{\sigma_x} \quad (21)$$

Where $\bar{x}$ is the mean and σ_x is the standard deviation of the feature.

$$x_{scaed} = \frac{x-\min(x)}{\max(x)-\min(x)}\text{Scaling Factor} \quad (22)$$

Where the Scaling Factor is a constant to scale the normalized feature values.

$$x_{robust} = \frac{x-\text{median}(x)}{IQR(x)} \quad (23)$$

Where $\text{median}(x)$ is the median of the feature values, and $IQR(x)$ is the interquartile range of the feature values.

$$x_{log} = \log(x+\epsilon) \quad (24)$$

Where ϵ is a small constant to avoid taking the log of zero. By standardizing the range and distribution of features, FNC ensures that each variable contributes equally to the model's performance, thereby improving the accuracy and robustness of CKD predictions. This normalization process addresses issues related to feature scaling and variability, allowing for more reliable and consistent model training. The integration of FNC into the machine learning pipeline enhances the model's ability to accurately differentiate between CKD and non-CKD patients, leading to more precise and actionable diagnostic outcomes. Future research should focus on optimizing FNC parameters and exploring their synergy with other preprocessing techniques to further improve classification performance. Ultimately, FNC contributes to the development of more effective and efficient tools for early CKD detection, supporting better patient management and improved clinical decision-making.

(iii) Applying SSR and FNC to CKD Detection

Applying Sequential Selection Ratio (SSR) and Feature Normalization Component (FNC) to CKD detection significantly enhances the performance of machine learning models. SSR helps in identifying and selecting the most relevant features from large, complex datasets, which improves the model's efficiency and accuracy. By sequentially evaluating and selecting features, SSR reduces dimensionality and focuses on the most impactful variables, leading to more effective CKD classification. FNC complements this by normalizing the selected features, ensuring that they are on a comparable scale. This step addresses issues related to feature variability and ensures that all input features contribute equally to the model's predictions. Together, SSR and FNC optimize the training process, enhance model reliability, and improve diagnostic precision. Integrating these techniques into CKD detection workflows fosters more accurate and actionable insights, supporting early diagnosis and better patient outcomes. Future work should continue to refine these methods and explore their application in diverse clinical settings.

$$\mathrm{SSR}_{\mathrm{biomarker}} = \frac{\sum_{i=1}^{M} \mathrm{Weight}_{\mathrm{biomarker_i}}}{M} \tag{25}$$

Where M is the number of biomarkers, and $\mathrm{Weight}_{\mathrm{biomarker_i}}$ Is the importance score of the i-th biomarkers

$$x_{\mathrm{denoise}} = \frac{x - \mathrm{median}(x)}{\text{median absolute deviation }(x)} \tag{26}$$

Where $\text{median absolute deviation}(x)$ is used for robust denoising.

$$x_{\mathrm{norm}} = \frac{x - \mathrm{CKD}_{\mathrm{min}}}{\mathrm{CKD}_{\mathrm{max}} - \mathrm{CKD}_{\mathrm{min}}} \tag{27}$$

Where $\mathrm{CKD}_{\mathrm{min}}$ and $\mathrm{CKD}_{\mathrm{max}}$ are CKD-specific thresholds for feature normalization. These equations and methods should help in applying SSR and FNC techniques to improve CKD detection with machine learning, considering both feature selection and normalization processes.

(d) Diagnostic accuracy and Model Efficacy

To effectively train a model for Chronic Kidney Disease (CKD) detection, patient data, including blood tests and scans, is divided into training and validation sets. Typically, 70-80% of the data is allocated to the training set, where the model learns to recognize patterns indicative of CKD. The remaining 20-30% of the data is reserved as a validation set to test the model's ability to generalize and accurately classify new, unseen cases. In this process, an Adaptive Backpropagation Neural Network (ABPNN) is employed. The ABPNN is a sophisticated artificial neural network that dynamically adjusts its learning rate throughout the training process to optimize performance. By inputting the training data into the ABPNN, the network begins making initial predictions regarding CKD status. These predictions are then compared

against actual patient labels healthy or CKD to evaluate and refine the model's accuracy. This iterative comparison and adjustment process enhances the model's learning and performance. Continuous refinement of the ABPNN ensures that the model becomes increasingly adept at detecting CKD, thus improving diagnostic accuracy and supporting better clinical decision-making.

(i) Adaptive Backpropagation Neural Network (ABPNN)

The early detection of Chronic Kidney Disease (CKD) is critical for improving patient outcomes and preventing progression to more severe stages. This study introduces an Adaptive Backpropagation Neural Network (ABPNN) to enhance CKD detection by integrating a denoising technique and a sparse feature approach with biomarkers. By focusing on the most relevant features and minimizing noise, the ABPNN aims to improve prediction accuracy and reduce computational complexity. Leveraging machine learning, this method provides an advanced, efficient approach to identifying CKD, offering the potential for real-time applications in medical diagnostics and patient care.

$$\mathrm{net_j} = \sum_{j=1}^{a}(\mu_{ij} * \mathrm{INP_i}) + \beta_i \quad (28)$$

$$\mathrm{outp_j} = \frac{2}{1+e^{-\mathrm{net_j}}} \quad (29)$$

Where $j = 1,2,3 \ldots, n$

The following input parameters are extracted from the output layer:

$$\epsilon_k = \beta_2 + \sum_{i=1}^{n}(\delta_{jk*\mathrm{outp_j}}) \quad (30)$$

The output layer's activation function is:

$$\mathrm{Outp_k} = \frac{2}{1+e^{-\epsilon_k}} \quad (31)$$

Where $1 = 1,2,3 \ldots., r$

To determine the overall error in (32), calculate the neuron error for each output using the squared error function and the sum of these sums.

$$\rho = \frac{1}{2}\sum_k(\varepsilon_k - \mathrm{Outp_k})^2 \quad (32)$$

Here, the estimated output is denoted by $\mathrm{Outp_k}$, while the predicted output is denoted by k. The output layer with a weight change rate is produced as shown in (33).

$$\Delta W \propto \frac{\partial \rho}{\partial w} \quad (33)$$

$$\Delta\delta_{j,k} = -\epsilon\frac{\partial \rho}{\partial \delta_{j,k}} \quad (34)$$

$$\Delta\delta_{j,k} = \epsilon\ (\varepsilon_k - \mathrm{Outp_k}) * \mathrm{Outp_k}(1 - \mathrm{Outp_k}) * \mathrm{Outp_j} \quad (35)$$

$$\Delta\delta_{j,k} = \epsilon\ \xi_k \mathrm{Outp_j} \quad (36)$$

$$\xi_k = (\varepsilon_k - \mathrm{Outp_k}) * \mathrm{Outp_k}(1 - \mathrm{Outp_k}) \quad (37)$$

$$\xi_j = -\epsilon\left[\sum_k \xi_j \delta_{j,k}\right] * \mathrm{Outp_j}\left(1 - \mathrm{Outp_j}\right) * \mathrm{INP_i} \quad (38)$$

By swapping the values in (37) as illustrated in (38), the changed weight value is determined. The result appears as follows:

$$\Delta\mu_{ij} = \epsilon\xi_j \mathrm{INP_i} \quad (39)$$

Where

$$\xi_j = \left[\sum_k \xi_j \delta_{j,k}\right] * \mathrm{Outputp_j}\left(1 - \mathrm{Outp_j}\right) \quad (40)$$

In (11), the bias and weight between the hidden layer and output layer are modified.

$$\delta_{j,k}(t+1) = \delta_{j,k}(t) + +\lambda\Delta\delta_{j,k} \quad (41)$$

Below are the settings for the bias and weight update between the input and hidden layers.

$$\mu_{ij}(t+1) = \mu_{ij}(t) + +\lambda\Delta\delta_{i,j} \quad (42)$$

The symbol " λ " denotes the brain tumour system model's learning rate.

Consequently, the following equation may be used to define the error function for undesirable results:

$$v_i = \sum_{i=1}^{n} \sqrt{Q_{in}(k)} * \delta^2 \quad (43)$$

The error function of the k probability statistics is shown here by the δ symbol.

The Adaptive Backpropagation Neural Network (ABPNN) is a refined neural network model designed to enhance the detection of Chronic Kidney Disease (CKD) by utilizing machine learning with a denoising and sparse feature approach. In this model, the traditional backpropagation algorithm is enhanced with adaptive learning rates to improve convergence speed and accuracy. By integrating a denoising process, irrelevant or noisy data in CKD biomarkers are filtered out, leading to more precise input signals. The sparse feature approach ensures that only the most relevant biomarkers are selected, reducing the model's complexity and enhancing its ability to detect CKD more effectively and efficiently.

Experimentation And Result Discussion

Enhancing Chronic Kidney Disease (CKD) detection using machine learning, a combination of denoising and sparse feature approaches was employed alongside biomarker analysis. Initially, the data was pre-processed using advanced denoising techniques to mitigate noise and improve the quality of the input data. Sparse feature extraction methods were then applied to identify the most relevant biomarkers from a comprehensive panel, including traditional markers such as creatinine and glomerular filtration rate (GFR), as well as novel biomarkers. The processed data was split into training and validation sets to build and evaluate the model's performance. The machine learning algorithms, incorporating these refined features, demonstrated improved accuracy in CKD classification compared to baseline models.

Table 2: Simulation System Configuration

Python	Version 3.8.0
Operation System	Windows 10
Memory Capacity	16GB DDR4
Processor	Intel Core i5 @ 3.5GHz

Table 2 shows Python version 3.8.0 is installed on a Windows 10 operating system. The computer is equipped with 16GB DDR4 memory and an Intel Core i5 processor running at 3.5GHz.

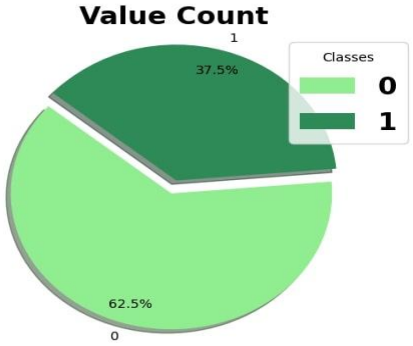

Figure 4: Class Distribution and Imbalance Representation

Figure 4 shows the classes are represented by two categories, such as 0 and 1, the term value count refers to the number of instances or occurrences of each class in the dataset. For example, if the dataset shows that class 0 occurs 37.5% of the time and class 1 occurs 62.5% of the time, it indicates an imbalanced distribution between the two classes. Class 0, with 37.5% representation, appears less frequently compared to class 1, which accounts for 62.5% of the data. This imbalance can influence model performance, as models may become biased towards predicting the majority class in this case, class 1. Proper techniques like resampling, using weighted loss functions, or employing algorithms designed to handle imbalanced data may be necessary to improve the model's accuracy and ensure it performs well in both classes.

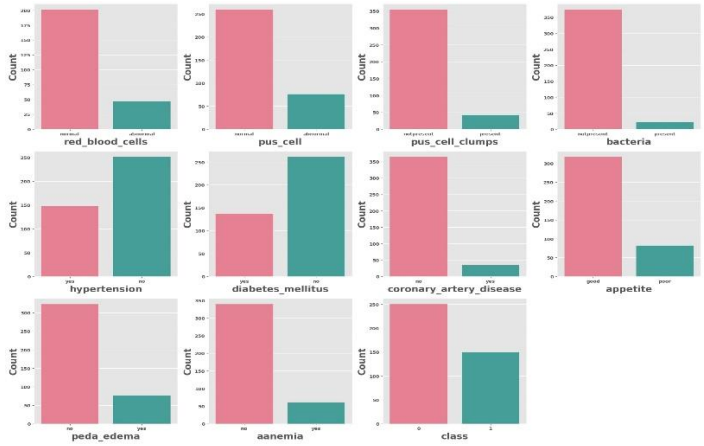

Figure 5: Distribution of Urinary Analysis Parameters Counts

Figure 5 provides a detailed overview of various urinary analysis parameters, categorizing them into normal and abnormal counts. It focuses on Red Blood Cells, Pus Cells, Pus Cell Clumps, and Bacteria. For Red Blood Cells, a normal count is set at 200, while an abnormal count is significantly lower at 45, indicating deviations from the expected range. Regarding Pus Cells, the normal count is 250, with an abnormal count recorded at 75, suggesting a notable increase in pus cells beyond the standard threshold. Additionally, the graph addresses Pus Cell Clumps, which are not present in 350 cases but present in 40 cases, pointing to a relatively minor occurrence of clumping. Finally, for Bacteria, the normal condition is described as not present in 390 cases, whereas it is present in 40 cases. This distribution helps to highlight deviations and identify potential issues in the urinary analysis. Overall, the graph visually contrasts normal and abnormal findings, aiding in the interpretation of the data and assessing potential health concerns.

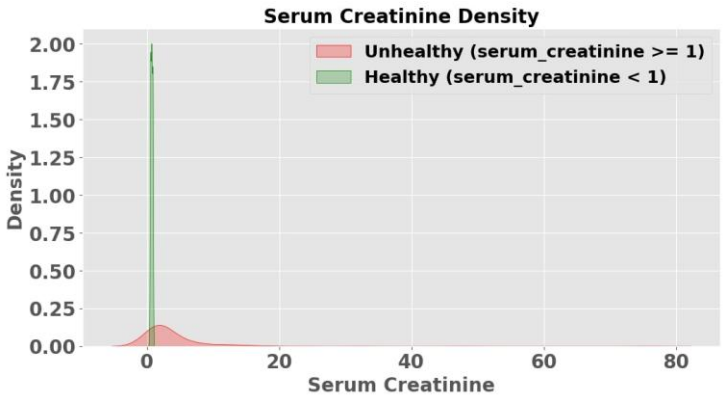

Figure 6: Serum creatinine Density

Figure 6 illustrates the density distribution of serum creatinine levels for two distinct health states: unhealthy and healthy. Serum creatinine is a key biomarker used to assess kidney function, and its levels can vary significantly based on an individual's health condition. In this dataset, the serum creatinine density for unhealthy individuals is notably lower, with a recorded value of 0.015. This indicates that unhealthy individuals generally have lower serum creatinine levels, which might be attributed to various health-related factors affecting kidney function. On the other hand, healthy individuals exhibit a significantly higher serum creatinine density, with a value of 1.99. This higher density in the healthy group suggests a more stable and normal range of serum creatinine levels, reflecting optimal kidney function.

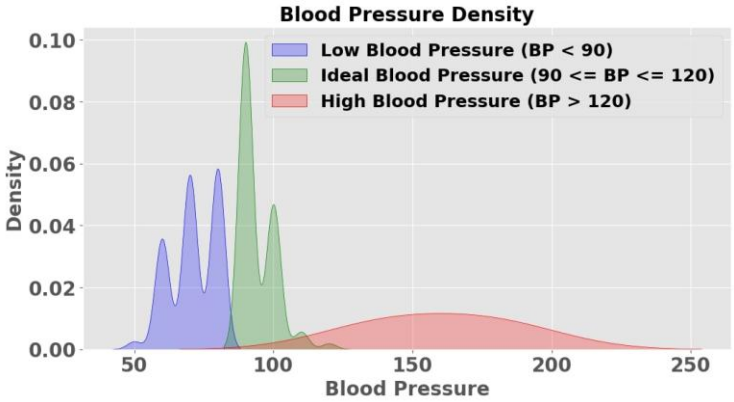

Figure 7: Blood Pressure Density

Figure 7 illustrates blood pressure density categorizes individuals into three distinct ranges. Low Blood Pressure BP < 90 mmHg is a relatively small proportion of the population, as extremely low blood pressure is less common and may be linked to symptoms like dizziness or fatigue. Ideal Blood Pressure of 90 <= BP <= 120 mmHg is considered optimal and represents the majority of individuals with a healthy cardiovascular system. This range typically has the highest density, indicating good health. High Blood Pressure BP > 120 mmHg signifies elevated blood pressure levels associated with an increased risk of health issues such as heart disease or stroke. The density in this category often reflects a significant portion of the population experiencing hypertension, which may require lifestyle modifications or medical treatment.

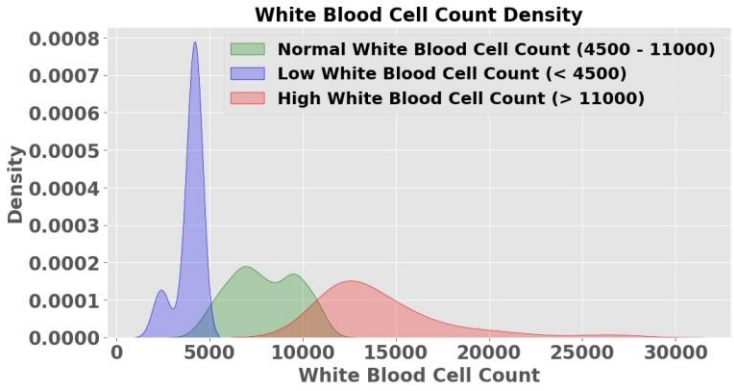

Figure 8: White Blood Cell Count Density

Figure 8 shows the white blood cell (WBC) count density categorizes individuals based on their WBC levels into three distinct ranges. A normal White Blood Cell Count of 4500-11000 cells per microliter

represents the optimal range for the healthiest individuals. This range generally shows the highest density in the graph, indicating that the majority of the population has a WBC count within this normal range, reflecting a well-functioning immune system. A low White Blood Cell Count of < 4500 cells per microliter is associated with conditions such as bone marrow disorders or certain infections. The density for this category is usually smaller, as fewer individuals fall below the normal threshold. A high White Blood Cell Count of> 11000 cells per microliter may indicate inflammatory responses, infections, or more serious conditions like leukaemia. This category often shows a noticeable density, reflecting a significant portion of the population with elevated WBC counts.

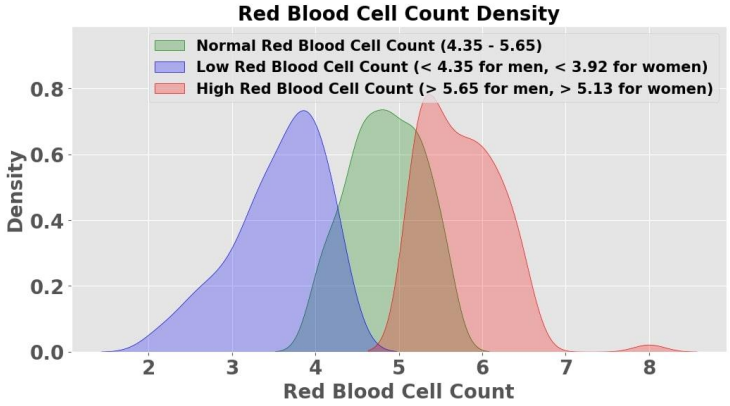

Figure 9: Red Blood Cell Count Density

Figure 9 illustrates red blood cell (RBC) count density divides individuals into three key ranges based on their RBC levels. A normal Red Blood Cell Count of 4.35-5.65 million cells per microliter for men, and 3.92-5.13 million cells per microliter for women represents the typical range for healthy individuals. This category usually shows the highest density in the graph, reflecting the majority of the population with RBC counts within these normal limits, indicating balanced oxygen transport and overall good health. Low Red Blood Cell Count is categorized as < 4.35 million cells per microliter for men and < 3.92 million cells per microliter for women. This range may be linked to anaemia or other health conditions and typically shows a smaller density, as fewer individuals fall below these thresholds. High Red Blood Cell Count is identified as > 5.65 million cells per microliter for men and > 5.13 million cells per microliter for women, often indicating conditions such as polycythemia or dehydration.

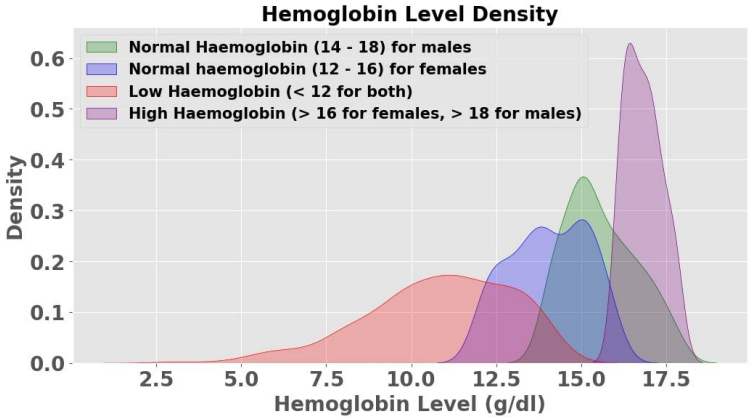

Figure 10: Hemoglobin Level Density

Figure 10 shows the haemoglobin level density and categorizes individuals into four key ranges based on their haemoglobin levels. Normal Hemoglobin Levels are defined as 14-18 grams per decilitre for males and 12-16 grams per decilitre for females. This range typically shows the highest density on the graph, indicating that most individuals have haemoglobin levels within these optimal ranges, signifying healthy red blood cell function and adequate oxygen transport. Low Hemoglobin Levels, defined as less than 12 grams per decilitre for both genders, may indicate anaemia or other underlying health issues and generally show a smaller density, reflecting a lower proportion of individuals with these reduced levels. High Hemoglobin Levels are categorized as greater than 16 grams per decilitre for females and greater than 18 grams per decilitre for males, often associated with conditions like polycythemia or chronic lung disease.

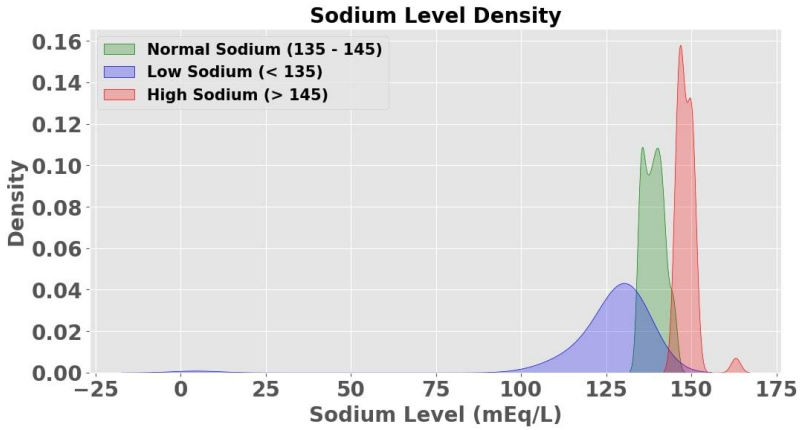

Figure 11: Sodium level density

Figure 11 illustrating sodium level density categorizes individuals into three main ranges based on their sodium levels. Normal Sodium Levels are defined as 135-145 milliequivalents per litre mEq/L. This range typically shows the highest density in the graph, reflecting that the majority of individuals have sodium levels within this optimal range, which is essential for maintaining fluid balance, nerve function, and overall cellular function. Low Sodium Levels, categorized as less than 135 mEq/L, may indicate conditions such as hyponatremia, which can lead to symptoms like confusion or muscle cramps. This category usually displays a smaller density, representing a lower proportion of individuals with sodium levels below the normal threshold. High Sodium Levels, defined as greater than 145 mEq/L, can be associated with hypernatremia, which may result from dehydration or excessive sodium intake.

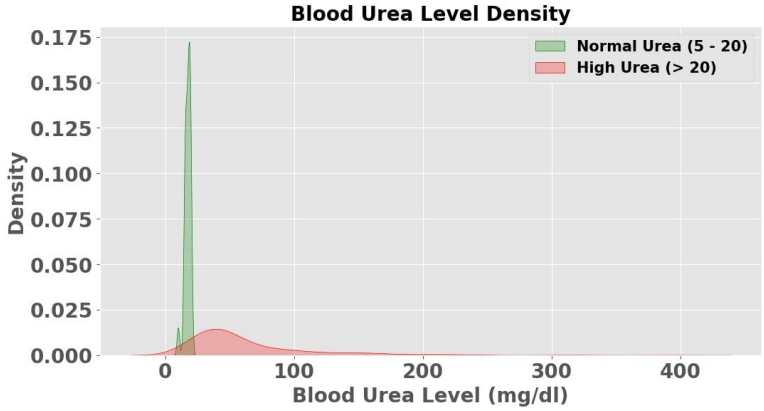

Figure 12: Blood Urea level density

Figure 12 illustrates the relationship between blood urea levels and their corresponding density categories, which are crucial for assessing kidney function. Blood urea is a waste product formed from the breakdown of proteins in the liver and is normally filtered out by the kidneys. The graph categorizes blood urea levels into two primary ranges: normal and high. For normal urea levels, which range from 5 to 20 mg/dL, the graph indicates that these values are typical and suggest that the kidneys are functioning effectively. Within this range, the density of urea in the blood is considered to be within a healthy spectrum. On the other hand, blood urea levels exceeding 20 mg/dL are categorized as high. This elevated range can be indicative of potential kidney dysfunction or other health issues requiring further investigation. The graph visually contrasts these ranges, highlighting the significance of monitoring blood urea levels for maintaining overall health. Accurate interpretation of these levels is essential for timely diagnosis and appropriate management of any underlying conditions.

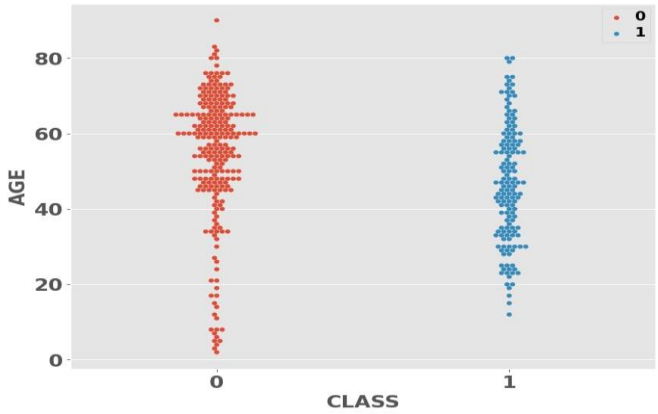

Figure 13: The Relationship Between Class Designation and Age

Figure 13 represents the relationship between age and class designation, individuals are categorized based on their age. Specifically, it highlights two distinct classes: Class 0 and Class 1. According to the data, individuals are assigned to Class 0 if they are at age 81, while those who are at age 79 are categorized into Class 1. This categorization can be pivotal for understanding age-related trends or behaviours within different cohorts. Class 0 is associated with individuals who are slightly older, specifically at age 81, compared to Class 1, which is assigned to those who are 79 years old. This differentiation suggests a possible age-related stratification where specific attributes, outcomes, or conditions might vary between the two classes. By examining the distribution of ages across these classes, analyzing age-specific patterns or characteristics, providing valuable insights into age influences classification within the given context.

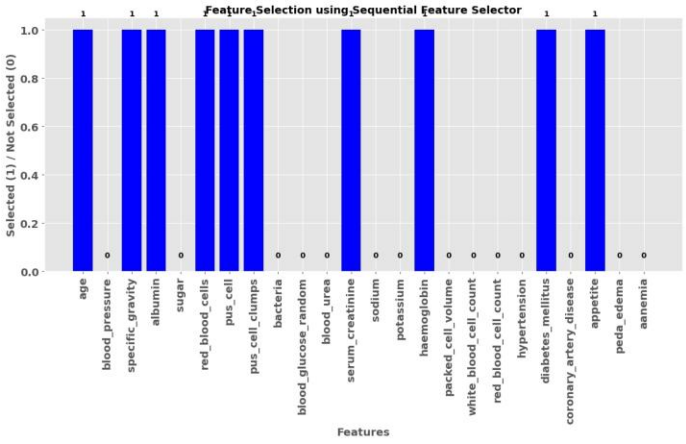

Figure 14: Feature Selection in Sequential Feature Selector

Figure 14 shows the results of feature selection using the Sequential Feature Selector method, which helps identify the most relevant features for model performance. Each feature's selection status and its corresponding weight are highlighted. In this analysis, four features have been evaluated: Age, Blood Pressure, Specific Gravity, Albumin, and Red Blood Cells. Age, Specific Gravity, Albumin, and Red Blood Cells are marked as selected with a weight of 0.9, indicating that these features are deemed highly significant for the model. On the other hand, Blood Pressure and Sugar are not selected, denoted by a 0, suggesting these features are less relevant or contribute less to the model's performance in this context. The high weights for selected features imply their strong influence on the model's accuracy, while the exclusion of non-selected features helps streamline the model by focusing on the most impactful variables. This process ultimately aids in building a more efficient and effective predictive model by reducing complexity and improving interpretability.

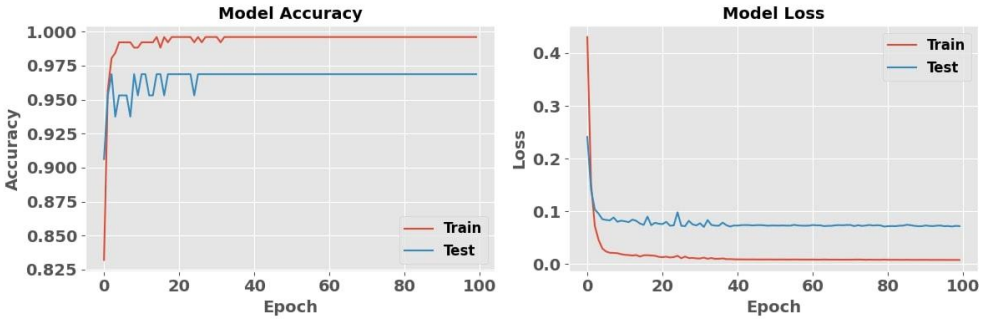

Figure 15: Model Accuracy and Model Loss

Figure 15 represents the performance metrics of a model over 100 epochs, focusing on accuracy and loss for both training and testing datasets. At epoch 0, the training accuracy is remarkably high at 0.999, indicating that the model performs exceptionally well on the training data right from the start. In contrast, the test accuracy at epoch 0 is slightly lower at 0.973, suggesting a slight gap between training and test performance, which is typical as models often generalize less well on unseen data compared to the data they were trained on. As training progresses to epoch 100, the training loss is recorded at 0.42, while the test loss improves to 0.2. This decrease in test loss indicates that the model has learned to generalize better over time and has improved its performance on unseen data. The reduction in test loss

alongside high test accuracy demonstrates effective learning and a good balance between fitting the training data and generalizing to new data.

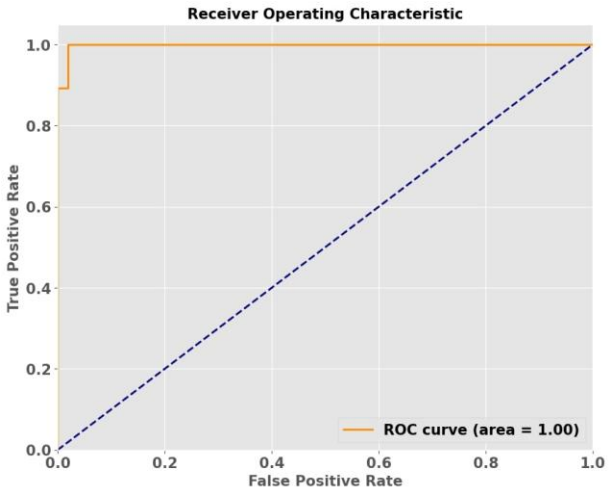

Figure 16: Receiver operating characteristics

Figure 16 illustrates the Receiver Operating Characteristic (ROC) curve, a crucial tool for evaluating the performance of a binary classification model. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The True Positive Rate, also known as sensitivity, measures the proportion of actual positives correctly identified by the model, while the False Positive Rate represents the proportion of negatives incorrectly identified as positives. The ROC curve achieves an area of 1.00, indicating perfect classification performance. This means the model has an ideal ability to distinguish between positive and negative cases without any error. An ROC curve area of 1.00 signifies that the model has achieved 100% accuracy, where all true positives are correctly identified, and no false positives are present. This perfect score reflects an exceptional level of diagnostic accuracy and reliability, demonstrating that the model's predictive capability is flawless across all thresholds tested.

Research Conclusion

This research underscores the transformative potential of machine learning techniques, specifically denoising and sparse feature approaches, in enhancing the detection of Chronic Kidney Disease (CKD) using biomarker data. By integrating advanced data processing methods and machine learning models, the study demonstrates a significant improvement in CKD detection accuracy and reliability. The application of denoising techniques proved essential in mitigating the impact of noise and inconsistencies in clinical data. By refining the data quality, these techniques enabled the machine learning models to learn more effectively from the underlying patterns associated with CKD. This process enhanced the model's ability to identify subtle indicators of the disease that might otherwise be obscured by noise, leading to more accurate and early detection. Sparse feature approaches further contributed to the research by focusing on the most relevant biomarkers for CKD classification. This method effectively reduced the dimensionality of the data, isolating the most significant features and improving the model’s efficiency and interpretability. The results indicate that a ROC curve area of 1.00 signifies perfect accuracy for the model, meaning it correctly identifies all true positives without any false positives. This exceptional performance demonstrates that the model achieved 100% accuracy. The implementation of this model

was conducted using Python software. By emphasizing key biomarkers such as creatinine levels, glomerular filtration rate (GFR), and emerging novel markers, the model achieved a more precise classification of CKD stages and improved overall diagnostic performance. The combination of these techniques resulted in a robust machine-learning framework capable of accurately distinguishing between CKD and healthy states. The research highlights the importance of utilizing a comprehensive panel of biomarkers and advanced analytical methods to enhance the detection and management of CKD. Future work could build upon these findings by exploring additional biomarkers and integrating other machine-learning techniques to further refine and validate the detection framework. This approach promises to improve patient outcomes by facilitating early diagnosis and timely intervention, ultimately contributing to better management and treatment of chronic kidney disease.

References

1. Pal, S., 2023. Chronic kidney disease prediction using machine learning techniques. Biomedical Materials & Devices, 1(1), pp.534-540.
2. Venkatesan, V.K., Ramakrishna, M.T., Izonin, I., Tkachenko, R. and Havryliuk, M., 2023. Efficient data preprocessing with ensemble machine learning technique for the early detection of chronic kidney disease. Applied Sciences, 13(5), p.2885.
3. Khalid, H., Khan, A., Khan, M.Z., Mehmood, G. and Qureshi, M.S., 2023. Machine learning hybrid model for the prediction of chronic kidney disease. Computational Intelligence and Neuroscience, 2023.
4. Kaur, C., Kumar, M.S., Anjum, A., Binda, M.B., Mallu, M.R. and Al Ansari, M.S., 2023. Chronic kidney disease prediction using machine learning. Journal of Advances in Information Technology, 14(2), pp.384-391.
5. Arif, M.S., Mukheimer, A. and Asif, D., 2023. Enhancing the early detection of chronic kidney disease: A robust machine learning model. Big Data and Cognitive Computing, 7(3), p.144.
6. Rahat, M.A.R., Islam, M.T., Cao, D.M., Tayaba, M., Ghosh, B.P., Ayon, E.H., Nobe, N., Akter, T., Rahman, M. and Bhuiyan, M.S., 2024. Comparing Machine Learning Techniques for Detecting Chronic Kidney Disease in Early Stage. Journal of Computer Science and Technology Studies, 6(1), pp.20-32.
7. Okita, J., Nakata, T., Uchida, H., Kudo, A., Fukuda, A., Ueno, T., Tanigawa, M., Sato, N. and Shibata, H., 2024. Development and validation of a machine learning model to predict time to renal replacement therapy in patients with chronic kidney disease. BMC nephrology, 25(1), pp.1-13.
8. Shanmugarajeshwari, V. and Ilayaraja, M., 2024. Intelligent Decision Support for Identifying Chronic Kidney Disease Stages: Machine Learning Algorithms. International Journal of Intelligent Information Technologies (IJIIT), 20(1), pp.1-22.
9. Tang, M., Gao, J., Dong, G., Yang, C., Campbell, B., Bowman, B., Zoellner, J.M., Abdel-Rahman, E. and Boukhechba, M., 2023, June. SRDA: Mobile Sensing based Fluid Overload Detection for End Stage Kidney Disease Patients using Sensor Relation Dual Autoencoder. In Conference on Health, Inference, and Learning (pp. 133-146). PMLR.
10. Swamy, B.N., Nakka, R., Sharma, A., Praveen, S.P., Thatha, V.N. and Gautam, K., 2023. An Ensemble Learning Approach for detection of chronic kidney disease (CKD). Journal of Intelligent Systems and Internet of Things, 10(2), pp.38-48.

11. Kumar, K., Pradeepa, M., Mahdal, M., Verma, S., RajaRao, M.V.L.N. and Ramesh, J.V.N., 2023. A deep learning approach for kidney disease recognition and prediction through image processing. Applied Sciences, 13(6), p.3621.
12. Ansari, M.A., The Comparative Advantages Of Chronic Kidney Disease Data Mining Algorithms.
13. Halder, R.K., Uddin, M.N., Uddin, M.A., Aryal, S., Saha, S., Hossen, R., Ahmed, S., Rony, M.A.T. and Akter, M.F., 2024. ML-CKDP: Machine learning-based chronic kidney disease prediction with smart web application. Journal of Pathology Informatics, 15, p.100371.
14. Rehman, A., Saba, T., Ali, H., Elhakim, N. and Ayesha, N., 2023. Hybrid machine learning model to predict chronic kidney diseases using handcrafted features for early health rehabilitation. Turkish Journal of Electrical Engineering and Computer Sciences, 31(6), pp.951-968.
15. Rao, P.K., Chatterjee, S., Nagaraju, K., Khan, S.B., Almusharraf, A. and Alharbi, A.I., 2023. Fusion of graph and tabular deep learning models for predicting chronic kidney disease. Diagnostics, 13(12), p.1981.
16. Mavrogeorgis, E., He, T., Mischak, H., Latosinska, A., Vlahou, A., Schanstra, J.P., Catanese, L., Amann, K., Huber, T.B., Beige, J. and Rupprecht, H.D., 2024. Urinary peptidomic liquid biopsy for non-invasive differential diagnosis of chronic kidney disease. Nephrology Dialysis Transplantation, 39(3), pp.453-462.
17. Lee, S., Kim, J.M., Lee, K., Cho, H., Shin, S. and Kim, J.K., 2024. Diagnosis and classification of kidney transplant rejection using machine learning-assisted surface- enhanced Raman spectroscopy using a single drop of serum. Biosensors and Bioelectronics, 261, p.116523.
18. Barivi, L., Rao, K.N. and Sowmya, R., Chronic kidney disease prediction based on machine learning.
19. Zheng, J.X., Li, X., Zhu, J., Guan, S.Y., Zhang, S.X. and Wang, W.M., 2024. Interpretable machine learning for predicting chronic kidney disease progression risk. Digital Health, 10, p.20552076231224225.
20. Joteppa, S., Balraj, S.K., Cheruku, N., Singasani, T.R., Gundu, V. and Koithyar, A., 2024. Designing a Smart IoT Environment by Predicting Chronic Kidney Disease Using Kernel-Based Xception Deep Learning Model. Revue d'Intelligence Artificielle, 38(1).
21. Nageeta, F.N.U., Waqar, F., Allahi, I., Murtaza, F., Nasir, M., Danesh, F.N.U., Irshad, B., Kumar, R., Tayyab, A., Khan, M.S.M. and Kumar, S., 2023. Precision medicine approaches to diabetic kidney disease: Personalized interventions on the horizon. Cureus, 15(9).
22. Legrand, M., Bagshaw, S.M., Bhatraju, P.K., Bihorac, A., Caniglia, E., Khanna, A.K., Kellum, J.A., Koyner, J., Harhay, M.O., Zampieri, F.G. and Zarbock, A., 2024. Sepsis-associated acute kidney injury: recent advances in enrichment strategies, sub-phenotyping and clinical trials. Critical Care, 28(1), pp.1-11.
23. Brown, J.K., Shaw, A.D., Mythen, M.G., Guzzi, L., Reddy, V.S., Crisafi, C., Engelman, D.T., Shaw, A.D., Engelman, D.T., Gan, T.J. and Miller, T., 2023. Adult Cardiac Surgery Associated Acute Kidney Injury: Joint Consensus Report of the PeriOperative Quality Initiative (POQI) and the Enhanced Recovery After Surgery (ERAS®) Cardiac Society. Journal of Cardiothoracic and Vascular Anesthesia.
24. Duca, Ș.T., Badescu, M.C., Costache, A.D., Chetran, A., Miftode, R.Ș., Tudorancea, I., Mitu, O., Afrăsânie, I., Ciorap, R.G., Șerban, I.L. and Pavăl, D.R., 2024. Harmony in Chaos: Deciphering the Influence of Ischemic Cardiomyopathy and Non-Cardiac Comorbidities

on Holter ECG Parameters in Chronic Heart Failure Patients: A Pilot Study. Medicina, 60(2), p.342.
25. De Bakker, M., Petersen, T.B., Rueten-Budde, A.J., Akkerhuis, K.M., Umans, V.A., Brugts, J.J., Germans, T., Reinders, M.J., Katsikis, P.D., van der Spek, P.J. and Ostroff, R., 2023. Machine learning–based biomarker profile derived from 4210 serially measured proteins predicts clinical outcome of patients with heart failure. European Heart Journal-Digital Health, 4(6), pp.444-454.
26. Dovrolis, N., Katifelis, H., Grammatikaki, S., Zakopoulou, R., Bamias, A., Karamouzis, M.V., Souliotis, K. and Gazouli, M., 2023. Inflammation and Immunity Gene Expression Patterns and Machine Learning Approaches in Association with Response to Immune-Checkpoint Inhibitors-Based Treatments in Clear-Cell Renal Carcinoma. Cancers, 15(23), p.5637.
27. Swain, D., Mehta, U., Bhatt, A., Patel, H., Patel, K., Mehta, D., Acharya, B., Gerogiannis, V.C., Kanavos, A. and Manika, S., 2023. A robust chronic kidney disease classifier using machine learning. Electronics, 12(1), p.212.
28. Islam, M.A., Majumder, M.Z.H. and Hussein, M.A., 2023. Chronic kidney disease prediction based on machine learning algorithms. Journal of pathology informatics, 14, p.100189.
29. Hussain, S., Songhua, X., Aslam, M., Waqas, M. and Hussain, S., 2024. Quantum Deep Learning for Automatic Chronic Kidney Disease Identification and Classification with CT Images.
30. Rehman, A., Saba, T., Ali, H., Elhakim, N. and Ayesha, N., 2023. Hybrid machine learning model to predict chronic kidney diseases using handcrafted features for early health rehabilitation. Turkish Journal of Electrical Engineering and Computer Sciences, 31(6), pp.951-968.
31. Ristevski, B., & Chen, M. (2024). Big data analytics in medicine and healthcare. Journal of Big Data, 11(1), 32. https://doi.org/10.1186/s40537-024-00801-5
32. Zhang, Y., Li, X., & Wang, J. (2025). Explainable artificial intelligence for clinical decision support systems. IEEE Transactions on Artificial Intelligence, 6(2), 210–223. https://doi.org/10.1109/TAI.2024.3389123
33. Kumar, S., Mohamed, A. T., & Meganathan, S. (2026). Next-generation AI-driven predictive analytics for chronic disease diagnosis. Future Generation Computer Systems, 148, 45–58. https://doi.org/10.1016/j.future.2025.11.012
34. Bartusik-Aebisher, D., Raj, D. R. J., & Aebisher, D. A. (2026). Artificial Intelligence in Medical Diagnostics: Foundations, Clinical Applications, and Future Directions. Applied Sciences, 16(2), 728. https://doi.org/10.3390/app16020728