Chronic Kidney Disease Early Prediction

Using Machine Learning.

A thesis submitted as a fulfilment of requirements for Master's degree in

Bioinformatics

By:

Sarah Alnokta /sarah180200

Supervised by:

Dr. Yanal Ahmad Alkudsi

2024-2025

Table of Contents

Table of Abbreviations

Abbreviation	Meaning
ACE	The Angiotensin Converting Enzyme
ACR	Albumin-Creatinine Ratio
ADH	Antidiuretic Hormone
AGEs	Advanced Glycation End Products
AI	Artificial Intelligence
AKI	Acute Kidney Injury
Ang I	Angiotensin I
Ang II	Angiotensin II
BP	Blood Pressure
BUN	Blood Urea Nitrogen
CCA	Correlation Coefficient Analysis
CKD	Chronic Kidny Disease
DM	Diabetes Mellitus
DN	Diabetic Nephropathy
EPO	Erythropoietin
ESRD	End-Stage Renal Disease.
GFR	Glomerular Filtration Rate
Hb.	Hemoglobin
HbA1c	Hemoglobin A1C
Hct	Hematocrit

KDIGO	Kidney Disease: Improving Global Outcomes
ML	Machine Learning
PCV	Packed Cell Volume
RAAS	The Renin-Angiotensin-Aldosterone System
ROS	Reactive oxygen species.
RSNA	Renal sympathetic nerve activity
SC	Serum Creatinine.
SG	The specific gravity
SGLT-2	Sodium-Glucose Transport Protein 2
UTI	Urinary Tract Infection
XGBoost	eXtreme gradient boosting

Table of Figures

Table of Tables

Abstract.

In the human body, the kidneys, play the important role of filtering wastes and toxic bodies from the blood. Chronic kidney disease (CKD) is a condition in which the human kidneys are damaged and unable to filter the blood in a proper way. It is a nontransmissible disease that causes mortality of large numbers worldwide and is very expensive to properly detect and diagnose, therefore, CKD patients often reach its chronic stages, especially in countries with limited resources. Furthermore, CKD is a silent killer due to the lack of physical symptoms at the initial stage, but a steady loss of glomerular filtration rate (GFR) occurs over a period of time longer than three months. CKD is a fatal disease if left undetected as it leads to renal failure, in the worst cases. However, the early diagnosis of CDK can significantly reduce the mortality rate. Moreover, if CKD is predicted early and correctly, it results in an increased probability of successful treatment and prolongs the patient's life. The advances in ML, in addition to predictive analytics, provide promising results which in turn prove the capability of prediction in CKD and beyond. The utilization of ML methods in nephrology enables the building of ML models to better detect the at-risk patients of CKD especially in primary care settings. The current study carries out a prediction-based method that helps in early detecting of CKD patients at the early stage. In this study, we utilize on of the boosting method, XGBoost to achieve a higher prediction accuracy for CKD. Various preprocessing steps are employed to achieve better prediction performance, along with suitable hyperparameter tuning and feature selection. We assessed the degree of importance of each feature in the dataset leading to CKD. The performance of the model was evaluated with accuracy. It attained 98 % accuracy for training and testing sets. The way the research was done leads to the conclusion that recent improvements in machine learning, along with the help of predictive modeling, make for an interesting way to find new solutions that can then be used to test the accuracy of prediction in the field of kidney disease and beyond.

Chapter 1 – Chronic Kidney Disease (CKD)

1.1. **Introduction**.

The urinary system, also known as the renal system, comprising the kidneys, ureters, bladder, and urethra, Fig (1) is responsible for the production, storage, and elimination of urine. The system contains two kidneys which are two bean-shaped organs, each about the size of a fist. They are located just below the rib cage, one on each side of your spine. [1] Every day kidneys filter about 120 to 150 quarts of blood to remove wastes and balance fluids. This process produces about 1 to 2 quarts of urine per day. [2]
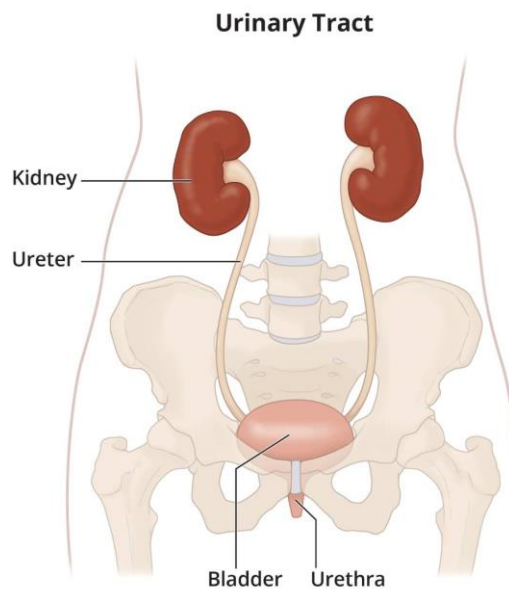
Figure (1): Front view of the Urinary Tract. [2]

1.2. **Kidneys functions.**

1.2.1. Filtration of Blood and urine formation:

- **Filtration:** The kidneys filter waste products, toxins, and excess substances from the blood, such as urea, creatinine, and excess salts. [3] Blood enters the kidneys through the renal arteries. Inside the kidneys, blood flows into tiny structures called glomeruli where small molecules such as water, glucose, salts, and waste products are filtered out of the blood and into the Bowman's capsule. [4] Fig (2)

- **Reabsorption and Secretion:** As the filtrate flows through the renal tubules, many of the filtered substances that are important for the body, such as water, glucose, amino acids, and electrolytes, are reabsorbed back into the bloodstream. [5]

- **Excretion:** After passing through the renal tubules, the remaining fluid is collected in the collecting ducts and eventually drains into the renal pelvis and then into the ureters. From there, urine is transported to the bladder for storage until it is excreted from the body. [4]

Kidney

Unfiltered blood

Filtered blood

Ureter

Urine exits to bladder
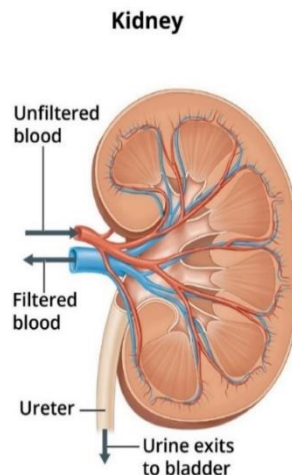
Figure (2): kidney anatomy and urine flow in it. Unfiltered blood flows into your kidneys through the renal artery and filtered blood exits through your renal vein. The ureter carries urine from the kidney to your bladder.[2]

1.2.2. **Acid-Base Balance.**

The kidneys play a critical role in maintaining acid-base balance in the body, ensuring that the pH of the blood remains within a narrow range (7.35-7.45). [6] This is done by: reabsorbing of approximately 85-90% of filtered bicarbonate (HCO_3^-) [7], secreting of hydrogen ions (H^+), which prevents the urine from becoming too acidic [8], generating new bicarbonate, and adjusting these processes in response to changes in blood ph. This intricate regulation ensures the stability of the body's internal environment.[6]

1.2.3. **Blood Pressure Regulation**.

Kidneys play crucial role in regulating blood pressure through the RAAS, controlling blood volume by adjusting sodium and water balance, interacting with the sympathetic nervous system, and responding to natriuretic peptides. These mechanisms ensure stable blood pressure, essential for proper organ function and overall health. Here's a brief look at these mechanisms:

- **Renin-Angiotensin-Aldosterone System (RAAS)**: When blood pressure drops, the juxtaglomerular cells in the kidneys release renin.[9] Renin converts angiotensinogen (from the liver) into angiotensin I, which is then converted to angiotensin II by the angiotensin-converting enzyme (ACE) primarily in the lungs. [10] Angiotensin II is a powerful vasoconstrictor narrows blood vessels, increasing blood pressure, and

11

stimulates aldosterone release from the adrenal glands, leading to sodium (Na^+) and water retention, which further increases blood volume and pressure. [11]

- **Regulation of Blood Volume:** "Sodium Balance": The kidneys adjust the amount of sodium excreted or retained, which directly affects blood volume. High sodium retention increases blood volume and pressure. [12] "Water Balance": The kidneys regulate water excretion through antidiuretic hormone (ADH). Increased ADH leads to more water reabsorption, increasing blood volume and pressure. [13]

- **Sympathetic Nervous System Interaction**: The kidney is extensively innervated by sympathetic nerves playing an important role in the regulation of blood pressure homeostasis. Sympathetic nerve activity is ultimately controlled by the central nervous system (CNS). Norepinephrine, the main sympathetic neurotransmitter, is released at prejunctional neuroeffector junctions in the kidney and modulates renin release, renal vascular resistance, sodium and water handling, and immune cell response. Under physiological conditions, renal sympathetic nerve activity (RSNA) is modulated by peripheral mechanisms interaction between efferent sympathetic nerves, central mechanism, and afferent sensory nerves. RSNA is increased in hypertension and, therefore, critical for the perpetuation of hypertension and the development of hypertensive kidney disease. [14]

1.2.4. **Endocrine Functions of the Kidney.**

The kidneys perform essential endocrine functions by producing hormones and enzymes that regulate various physiological processes. Here are the key endocrine functions:

- **Renin:** Regulates blood pressure and fluid balance. Renin initiates the renin-angiotensin-aldosterone system (RAAS), which increases blood pressure by constricting blood vessels and promoting sodium and water retention.[9]

- **Erythropoietin (EPO):** Erythropoietin (EPO) is a glycoprotein hormone, naturally produced by the peritubular cells of the kidney, in response to low oxygen levels (hypoxia), which acts on the bone marrow to increase red blood cell production, enhancing the blood's oxygen-carrying capacity. [15]

- **Calcitriol (Active Form of Vitamin D):** Vitamin D can be synthesized in the skin or ingested in the diet and is transported to the liver where it is metabolized into 25-hydroxyvitamin D. This vitamin D metabolite is the main storage form of vitamin D, and subsequently is converted to 1,25-dihydroxy vitamin D in the kidney. This is the major active metabolite of vitamin D and is responsible for the classical effects of vitamin D on calcium-phosphorus metabolism, maintenance of skeletal health, and the regulation of parathyroid function. [16]

1.3. **Acute Kidney Injury (AKI).**

Acute Kidney Injury (AKI) is the term that has recently replaced the term Acute renal failure (ARF). Acute kidney injury (AKI) is defined as an abrupt (within hours) decrease in kidney function, which encompasses both injury (structural damage) and impairment (loss of function), [17] leading to an inability to filter waste products from the blood effectively. This can cause an accumulation of waste products and disturbances in electrolyte, acid-base, and fluid balance.[18] Causes of AKI: include, Prerenal AKI- Occurs because plasma flow and intraglomerular pressure are inadequate to maintain filtration capacity, Postrenal AKI- caused by an obstruction of urinary such as, benign prostatic hyperplasia, urethral stricture, pelvic or abdominal cancers, etc. and Renal AKI-may be linked to nephrotoxic drugs, other nephrotoxins, infection, etc. [19]

1.4. **Chronic Kidney Disease (CKD).**

The definition and classification of chronic kidney disease (CKD) have evolved over time, but current international guidelines define this condition as decreased kidney function shown by glomerular filtration rate (GFR) of less than 60 mL/min per 1·73 m2, or markers of kidney damage, or both, of at least 3 months duration, regardless of the underlying cause. [20] CDK is highly prevalent (10-13% of the population), irreversible and progressive. [21] kidney biopsy samples can show definitive evidence of CKD, through common changes such as glomerular sclerosis, tubular atrophy, and interstitial fibrosis [20] Fig (3). Patients with this pathology remain asymptomatic most of the time, presenting the complications typical of renal dysfunction only in more

advanced stages. [21] Diabetes Mellitus (DM) and hypertension are the most common causes of chronic kidney disease (CKD) [20]. Other causes of kidney disease include, polycystic kidney disease (PKD),[22] Neurogenic bladder dysfunction, [23] urinary tract infection (UTI), Drug-Induced Nephrotoxicity, Lupus nephritis, IgA glomerulonephritis disorders in which the body's immune system attacks its own cells and organs, heavy metal poisoning, such as lead poisoning, rare genetic conditions, such as Alport syndrome, hemolytic uremic syndrome in children, and, renal artery stenosis. [22]
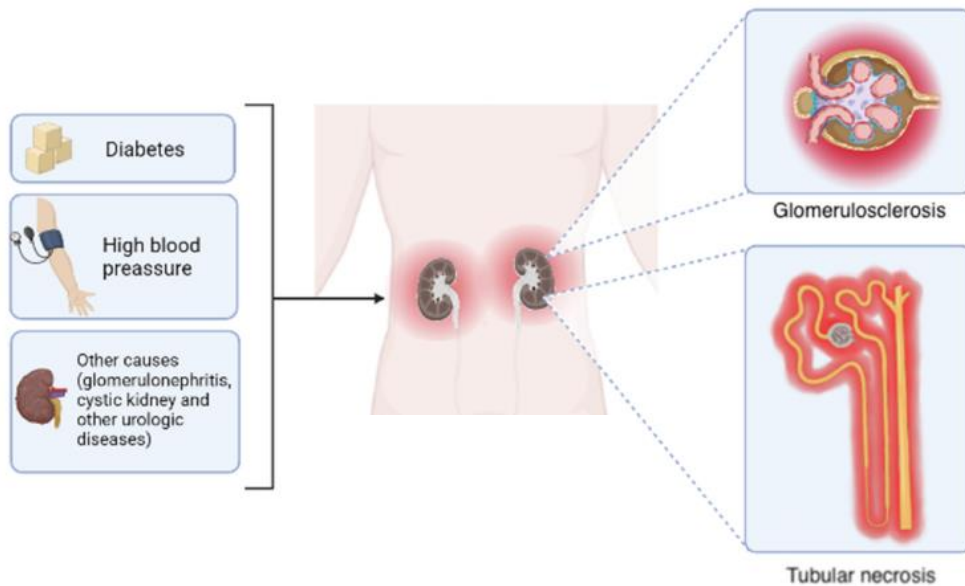
Figure (3): Primary causes and morphological outcomes of CKD. Regardless of the etiology, the progressive reduction in glomerular filtration rate occurs accompanied by two common histological changes: glomerular sclerosis and tubular necrosis.[24]

1.5. **Stages of CKD:**

CKD is classified into six categories based on GFR shown in (Table1) [25], as defined by the Kidney Disease Improving Global Outcomes (KDIGO) guidelines. It also includes the staging based on three levels of albuminuria (A1, A2, and A3) shown in

(Table2) [25], with each stage of CKD being sub-categorized according to the urinary albumin-creatinine ratio. [26] Given the greater risk of disease progression, those with higher risk of disease progression should undergo more frequent monitoring Fig (4). [27]

Table 1. The six categories based on GFR include: [25]

GFR category	GFR (ml/min/1.73 m²)	Terms
G1	≥90	Normal or high
G2	60–89	Mildly decreased*
G3a	45–59	Mildly to moderately decreased
G3b	30–44	Moderately to severely decreased
G4	15–29	Severely decreased
G5	<15	Kidney failure

Abbreviations: CKD, chronic kidney disease; GFR, glomerular filtration rate.

Table 2. The three levels of albuminuria include an albumin-creatinine ratio (ACR): [25]

Category	AER (mg/24 hours)	ACR (approximate equivalent)		Terms
		(mg/mmol)	(mg/g)	
A1	<30	<3	<30	Normal to mildly increased
A2	30-300	3-30	30-300	Moderately increased*
A3	>300	>30	>300	Severely increased**

Abbreviations: AER, albumin excretion rate; ACR, albumin-to-creatinine ratio; CKD, chronic kidney disease.

Albuminuria categories
Description and range

				A1	A2	A3
				Normal to mildly increased	Moderately increased	Severely increased
				<30 mg/g <3 mg/mmol	30–299 mg/g 3–29 mg/mmol	≥300 mg/g ≥30 mg/mmol
GFR categories (ml/min/1.73 m²) Description and range	G1	Normal or high	≥90	Screen 1	Treat 1	Treat 3
	G2	Mildly decreased	60–89	Screen 1	Treat 1	Treat 3
	G3a	Mildly to moderately decreased	45–59	Treat 1	Treat 2	Treat 3
	G3b	Moderately to severely decreased	30–44	Treat 2	Treat 3	Treat 3
	G4	Severely decreased	15–29	Treat* 3	Treat* 3	Treat 4+
	G5	Kidney failure	<15	Treat 4+	Treat 4+	Treat 4+

CKD is classified based on:
• Cause (C)
• GFR (G)
• Albuminuria (A)

Low risk (if no other markers of kidney disease, no CKD)
Moderately increased risk
High risk
Very high risk

Figure (4): Frequency of monitoring glomerular filtration rate (GFR) and albuminuria in people with CKD. Albuminuria and GFR grid reflect the risk of progression by intensity of coloring (green, yellow, orange, red, and deep red). The numbers in the boxes are a guide to the frequency of monitoring (number of times per year). [27]

1.6. Most common causes of CKD.

1.6.1. Diabetes Mellitus (DM).

Diabetes Mellitus (DM) is the leading cause of CKD worldwide. The condition, termed diabetic nephropathy (DN) when referring specifically to kidney damage caused by diabetes, results from the long-term effects of hyperglycemia -high blood sugar levels- kidney function. [28] [29] The pathogenesis of DN development and progression is complex and multifactorial with the involvement of many pathways and mediator, shown in Fig (5). Conventionally, the developmental mechanism of DN is the result of abnormal homeostasis, which includes

hemodynamic abnormalities, metabolic disorders, and hormone synthesis such as

Ang-II. [29]
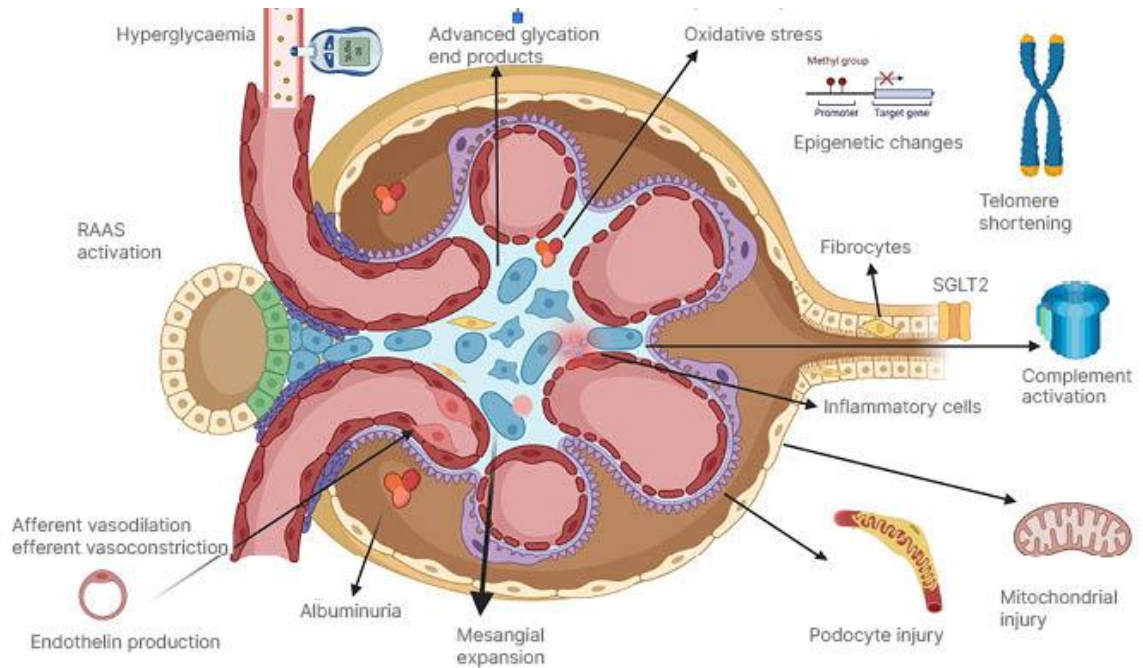
Figure (5): Depicts the glomerulus and tubular cells with some changes that happen in diabetes. [30]

- **Hyperglycemia and Glomerular Hyperfiltration:** Chronic high blood glucose levels cause increased glucose filtration by the kidneys.[31] Initially, increased blood flow and pressure in the glomeruli lead to hyperfiltration, causing damage over time.[32]

- **Formation of Advanced Glycation End Products (AGEs):** Persistent hyperglycemia leads to the formation of AGEs, harmful compounds formed when proteins or fats combine with sugars. These compounds promote inflammation and fibrosis in the kidneys, contributing to structural damage and functional decline. [33]

- **Activation of the Renin-Angiotensin-Aldosterone System (RAAS):** Hyperglycemia triggers RAAS, increasing blood pressure and glomerular damage. Angiotensin II elevated levels promote vasoconstriction, increase glomerular pressure, and stimulate inflammatory and fibrotic pathways.[29]

- **Oxidative Stress:** Conventionally, oxidative stress is a condition of oxidative damage to tissues include, glomerulus, resulting in proteinuria and tubulointerstitial fibrosis due to an imbalance between oxidants and antioxidants. Increased ROS due to hyperglycemia is central to the pathogenesis of DN. [29]

1.6.2. Hypertension.

Hypertension presents in approximately 80–85% of patients with CKD, with the more severe glomerular diseases having a higher incidence of hypertension. For any given cause of CKD including hypertension itself, the elevation in systemic blood pressure (BP) accentuates the rate at which glomerular filtration rate (GFR) declines which makes hypertension an independent risk factor for end-stage renal disease (ESRD). [34] The interaction between hypertension and CKD is complex and increases the risk of adverse cardiovascular and cerebrovascular outcomes. This is particularly significant in the setting of resistant hypertension commonly seen in patient with CKD. [35] Hypertension mainly cause vascular and glomerular lesions as shown in Fig (6). [36] The pathophysiology of CKD associated hypertension is multi-factorial with different mechanisms contributing to hypertension. [35]
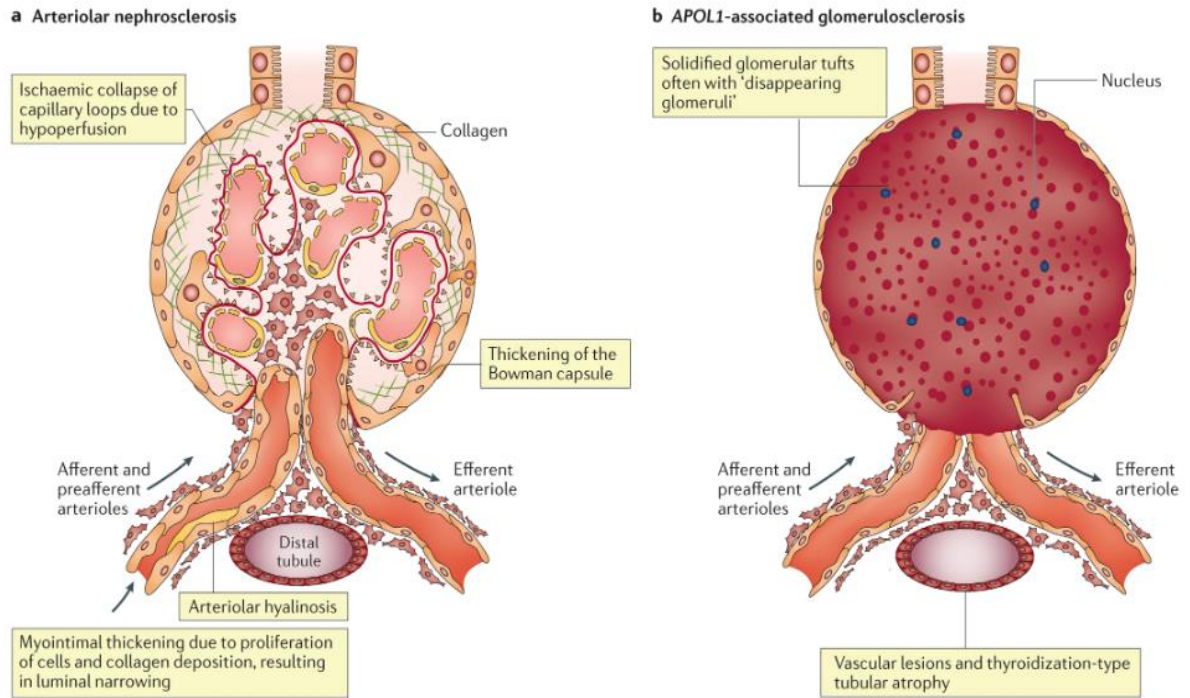
Figure (6): Schematic representation of vascular and glomerular lesions in hypertension-attributed nephropathy. a | Features of arteriolar nephrosclerosis caused by hypoperfusion; thickening of arteries; thickening of the Bowman capsule. b | Features of APOL1-associated glomerulosclerosis include solidified glomerulosclerosis, often with 'disappearing glomeruli' and vascular lesions. [36]

- **The Renin-Angiotensin-Aldosterone System (RAAS) regulation:** It is well known that RAAS is activated in CKD. Renin is secreted from the kidney and converts angiotensinogen to angiotensin I, which in turn is converted to angiotensin II (Ang II) by angiotensin-converting enzyme (ACE). Activation RAAS increases renin secretion ultimately leading to high circulating plasma concentrations of Ang II, a common feature of CKD. [37] Ang II, further increase blood pressure and promote sodium and water retention. Angiotensin II Effects: Elevated angiotensin II levels cause vasoconstriction, increasing glomerular

pressure, and stimulating the production of inflammatory and fibrotic cytokines. [38]

- **Sympathetic nervous system regulation:** Ang II also plays an important role in regulating sympathetic outflow from the brainstem. Ang II potentate norepinephrine release from sympathetic nerve terminals. [38] resulting in decreased urinary sodium excretion. [37] The renal artery is highly innervated, with efferent renal nerves that originate from the central nervous system, and afferent renal nerves that originate from the kidneys. [38]

- **Oxidative Stress:** Elevated levels of Ang II also contribute to overproduction of ROS as Ang II is a potent activator of nicotinamide adenine dinucleotide phosphate (NAD(P)H) oxidase, the primary source of superoxide. Although Ang II receptor blockers and ACE inhibitors are a first-line choice of treatment in CKD patients. [37]

- **Vascular Damage:** Hypertension causes "Arteriolar Thickening" where the walls of the arterioles in the kidneys to thicken and narrow, reducing blood flow to the nephrons (functional units of the kidney). Reduced blood flow leads to ischemia (lack of oxygen) in the kidney tissues, causing tubular and interstitial damage. [38] Hypertension damages the endothelial cells lining the blood vessels, reducing their ability to produce vasodilators like nitric oxide.[39] Reduced vasodilation capacity exacerbates hypertension and renal ischemia, further damaging the kidneys. [34]

- **Pro-inflammatory Cytokines:** Hypertension induces the production of pro-inflammatory cytokines, which contribute to chronic renal inflammation.[40] This

21

chronic inflammation leads to fibrosis (scarring) in the renal glomeruli, impairing kidney function.[41]

1.7. CKD Management:

The Objectives of CKD Management are, "Prevent" or "Delay" the progression of CKD to end-stage renal disease (ESRD), address complications arising from reduced kidney function, maintain patient well-being and functional status, shown in Fig (7), and prepare for "Renal Replacement Therapy", dialysis or kidney transplantation if necessary.[42]
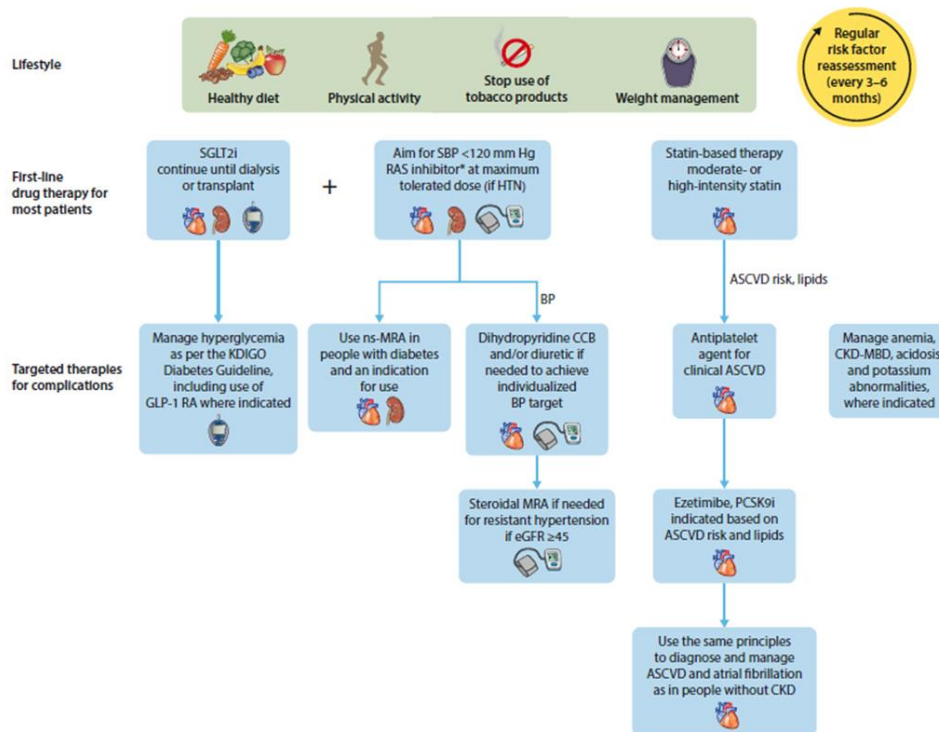
Figure (7). Holistic approach to chronic kidney disease (CKD) treatment and risk modification. [42]

1.7.1. Blood Pressure Control.

- Target BP, Generally, aim for <130/80 mmHg. ACE Inhibitors/ARBs: First-line agents, especially for patients with proteinuria. [42] [43]

- **Diuretics:** Thiazide such as (HCT and chlorthalidone), in early stages, [44] and recent studies chose that they seem to maintain their effectiveness in in patients with advanced CKD too. [43] In CKD Stages 4-5, loop diuretics such as, furosemide should be started at a dose of 40 to 80 mg once daily. [42]

- **Methyldopa:** reduces glomerular filtration rate and increases sodium retention and most commonly used in dialysis CKD patients. [44]

1.7.2. Glycemic Control.

- Target HbA1c: Typically, <7%, individualized based on patient risk factors. [42]

- Glycemic control may delay progression of CKD, with most guidelines recommending the ideal target hemoglobin A1c is approximately ~7.0 %. [46]

- Dose adjustments in oral hypoglycemic agents and Insulin may be necessary. [29]

- Use of specific medication classes such as SGLT-2 inhibitors in those with severely increased albuminuria should be considered, [29] SGLT-2 inhibitors shown to provide renal protection. [42]

1.7.3. Management of Anemia.

- **Erythropoiesis-Stimulating Agents (ESAs):** Initiate when hemoglobin <10 g/dL and aim to keep it between 10-11.5 g/dL. [47]

- **Iron Supplementation:** Oral or IV iron to maintain ferritin >100 ng/mL and transferrin saturation >20%. [48]

- Monitoring: Hemoglobin every 1-3 months, iron studies every 3 months. [42]

1.7.4. **Bone and Mineral Disorder Management**.

- **Phosphate Binders:** 90% of daily phosphate load gets excreted by kidneys; a decrease in renal function causes decreased secretion and increased retention of phosphate. High serum phosphate levels are seen only in the late stages of chronic kidney disease. Activation of compensatory mechanisms, including an increase in fibroblast growth factor 23 and parathyroid hormone secretion, prevent an increase in serum phosphate during the early stages of CKD. As CKD progresses, these mechanisms are unable to overcome the input of phosphate from dietary intake, leading to hyperphosphatemia. In this case, when phosphate levels are very high (greater than 6 mg/dl), phosphate binders are the agent of choice, where they reduce the absorption of dietary phosphate in the gastrointestinal tract, by exchanging the anion phosphate with an active cation (carbonate, acetate, oxyhydroxide, and citrate) to form a nonabsorbable compound that gets excreted in the feces, such as, Calcium-based binders (e.g., calcium carbonate and calcium acetate), and Sevelamer -is a crosslinked polymer that exchanges phosphate with HCl or carbonate in the gastrointestinal tract. [49]

- **Vitamin D Analogues**: To correct secondary hyperparathyroidism. [50]

- Monitoring: Serum calcium, phosphate, PTH levels every 3-6 months. [42]

1.7.5. **Renal Replacement Therapy (RRT).**

- Indications for RRT: Initiate dialysis when eGFR <15 mL/min/1.73m² with symptoms of uremia, fluid overload unresponsive to diuretics, hyperkalemia, or

24

acidosis. [42] Dialysis makes it possible to continue living with end-stage kidney disease for many years or even decades.[51]

- Types of Dialysis: Hemodialysis or peritoneal dialysis based on patient preference and clinical suitability. [42]

- **Hemodialysis:** is the most commonly used type of dialysis. In this method, blood is transported out of the body through tubes and cleaned in a machine using dialysis fluid. The dialysis is typically carried out three times per week at a dialysis center, each session lasts about four to five hours. Hemodialysis usually doesn't lead to any complications. [52]

- **Peritoneal dialysis:** here, the blood isn't cleaned outside the body but on the inside, in the abdominal cavity (the hollow space surrounding the organs in the abdomen), with the help of dialysis fluid. Patients are given a special abdominal catheter: About two liters of the dialysis fluid are transported into the abdominal cavity through this catheter. After some time, this fluid is then removed and replaced with new dialysis fluid. This type of dialysis can also be done at home on your own. The most common complications of peritoneal dialysis have to do with the catheter, like peritonitis, painful irritation of the abdomen. [51]

- **Kidney Transplant:** Evaluation for transplantation as a long-term solution for suitable candidates. [42] A kidney transplant is often the best option for people who have end-stage kidney disease, but it's not always possible. In that case, and while waiting for a donor kidney, it's necessary to have renal (kidney) replacement therapy with dialysis. [51]

1.8. **Prevalence of Chronic Kidney Disease (CKD).**

Chronic Kidney Disease (CKD) is a significant global health issue [53] and emerged as one of the most prominent causes of death and suffering in the 21st century, with rising prevalence, due to increasing rates in risk factors, such as obesity and diabetes mellitus, hypertension, and aging populations. [53] The Global Burden of Disease Study 2017 estimated that the global prevalence of CKD is approximately 9.1%, affecting around **843.6 million** individuals worldwide [54] [55] and represents an especially large burden in low- and middle-income countries, which are least equipped to deal with its consequences, summary CKD epidemiology shown in Fig.8. [55] CKD prevalence increases significantly with age. In individuals over 65, the prevalence can be as high as 40-50% due to age-related decline in kidney function and higher rates of comorbid conditions. Some studies suggest that CKD prevalence is slightly higher in women compared to men, also reported important differences by geographic region classified by income level, with a CKD age-standardized prevalence of 8.6% and 9.6% in men and women, respectively, in high-income countries, and 10.6% and 12.5% in men and women, respectively, in low- and middle-income countries. which are least equipped to deal with its consequences. [55]
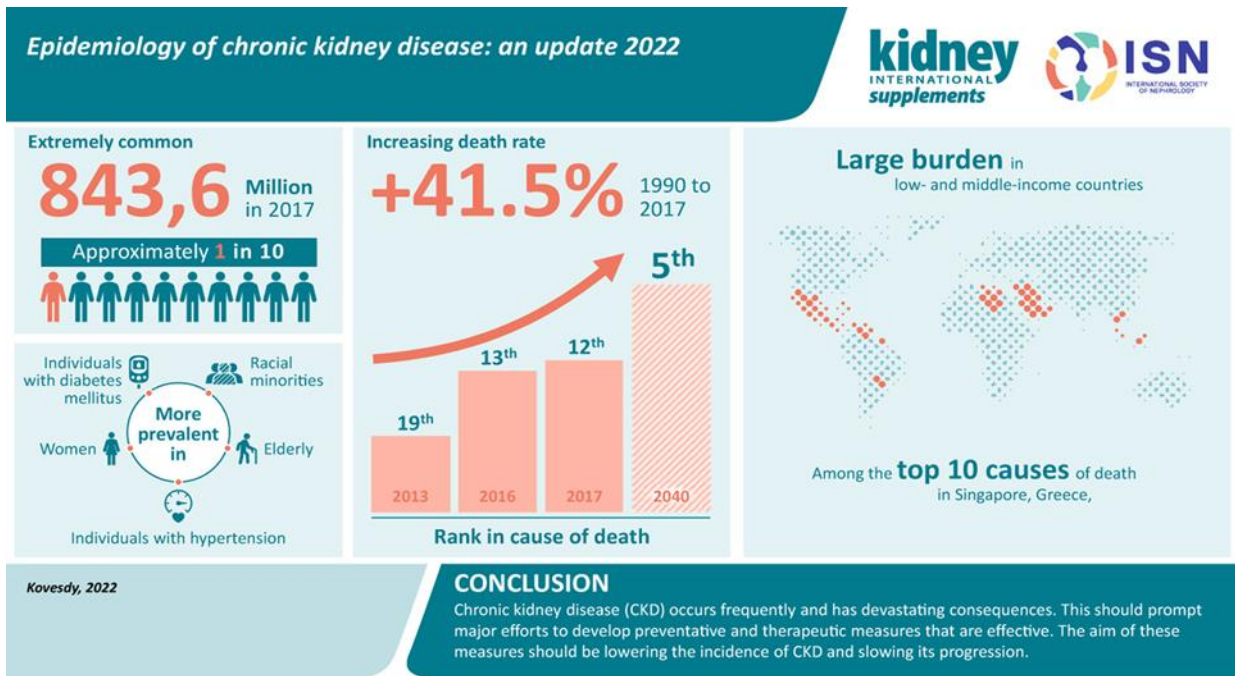
Figure (8). A summary of chronic kidney disease epidemiology. [55]

1.9. **Most Important Renal Function Tests.**

1.9.1. **Glomerular filtration rate (GFR).**

The best overall indicator of the glomerular function is the glomerular filtration rate (GFR). GFR is the rate in milliliters at which substances in plasma [57] in other words, the clearance of a substance from the blood, [56] are filtered through the glomerular capillaries and into the Bowman's capsule per unit of time (minute). [57] The normal GFR is 90 to 120 mL per minute. [56] The clearance rate for a given substance equals the GFR when it is neither secreted nor reabsorbed by the kidneys. For such a given substance, the urine concentration multiplied by the urine flow equals the mass of the substance excreted during urine collection. This mass divided by the plasma concentration is equivalent to the volume of plasma from which the mass was originally filtered. [57] Below is the equation used to determine GFR,

typically recorded in volume per time (e.g., mL/min): **GFR = [UrineX (mg/mL)] * urine flow (mL/min)/ [PlasmaX (mg/mL)], where X is a substance that is completely excreted.** [57]

1.9.2. **Serum Creatinine (SC).**

The most commonly used endogenous marker for the assessment of glomerular function is creatinine. [56] Creatinine is the by-product of creatine phosphate in muscle, and it is produced at a constant rate by the body. For the most part, creatinine is cleared from the blood entirely by the kidney. Decreased clearance by the kidney results in increased blood creatinine. [56] Elevated levels indicate impaired kidney function. [57] Serum creatinine is also utilized in GFR estimating equations such as the CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration) equation. These eGFR equations are superior to serum creatinine alone since they include race, age, and gender variables. [56] [57]

1.9.3. **Blood Urea Nitrogen (BUN).**

Urea or BUN is a nitrogen-containing compound formed in the liver as the end product of protein metabolism and the urea cycle. About 85% of urea is eliminated via kidneys; the rest is excreted via the gastrointestinal (GI) tract. Serum urea levels increase in conditions where renal clearance decreases (in acute and chronic renal failure/impairment). Urea is increased earlier in renal disease. [56]

1.9.4. **Random blood glucose.**

Random blood glucose test measures the glucose concentration in the blood at any given time, regardless of when the patient last ate. Chronic Kidney Disease (CKD) patients, especially those with diabetes, require careful monitoring of blood glucose levels. Random blood glucose testing provides an essential snapshot of glycemic control, helping to guide treatment adjustments and prevent complications. [58]

1.9.5. Packed cell volume (PCV).

Packed Cell Volume (PCV), also known as hematocrit (Hct), is the volume percentage of red blood cells (RBCs) in whole blood. It is a crucial measurement in evaluating the overall health and function of the blood and the body's ability to transport oxygen. [59] It is a critical parameter in evaluating the health of CKD patients. It reflects the proportion of blood volume occupied by red blood cells and provides insights into anemia and overall blood health, both of which are significantly impacted in CKD. [60] **Low PCV (Anemia)** Common in CKD, characterized by fatigue, weakness, and pallor. Causes include EPO deficiency, iron deficiency, and chronic inflammation. [60]

1.9.6. White blood cell (WBC) count.

White blood cell (WBC) count is a critical parameter in evaluating the immune status and overall health of CKD patients. CKD can significantly impact the immune system, leading to alterations in WBC count and function. Elevated WBC counts can indicate underlying inflammation, which is common in CKD and contributes to cardiovascular disease and other complications. [61] Patient on

29

hemodialysis or peritoneal dialysis are particularly prone to infections, which can affect WBC counts. [52] WBC count should be monitored regularly, typically every 1-3 months, depending on the patient's condition and treatment regimen. [61]

1.9.7. **Hemoglobin (Hb).**

Hemoglobin (Hb) is a key indicator of anemia, a common complication in CKD patients. Monitoring hemoglobin levels helps in diagnosing, managing, and treating anemia, thus improving patient outcomes and quality of life. The kidneys produce EPO, which stimulates red blood cell production in the bone marrow. CKD leads to reduced EPO production, causing decreased red blood cell production and anemia. In addition, CKD patients often have iron deficiency due to reduced dietary absorption, blood loss during dialysis, and increased iron requirements for erythropoiesis. General Target in CKD patients aim for hemoglobin levels of 10-11.5 g/dL to balance the risk of cardiovascular events with the benefits of reducing anemia symptoms. [60]

1.9.8. **Uric Acid.**

Uric acid is a waste product formed from the breakdown of purines, which are found in certain foods and are also produced by the body. In CKD patients, the kidneys' ability to excrete uric acid is impaired, leading to elevated blood levels. [26] CKD impairs the kidneys' ability to filter and excrete uric acid, leading to hyperuricemia (high blood uric acid levels. High uric acid levels can contribute to the progression of CKD, making it crucial to monitor and manage these levels

30

effectively. High uric acid levels can contribute to the progression of CKD, making it crucial to monitor and manage these levels effectively. [62]

1.9.9. **Albumin.**

The Albumin-to-Creatinine Ratio (ACR) urine test is a key diagnostic tool for detecting and monitoring kidney damage in CKD patients. Elevated urinary albumin levels are an early marker of kidney disease, often preceding reductions in glomerular filtration rate (GFR).[63] ACR is a sensitive test for early detection of kidney damage, often used in screening patients with risk factors like diabetes and hypertension. Persistent elevation in ACR over 3-6 months confirms CKD. ACR helps in staging CKD and determining the severity of kidney damage. [26]

Albuminuria: Albuminuria refers to abnormal loss of albumin in the urine. Albumin is one type of plasma protein found in the urine in normal subjects and in larger quantity in patients with kidney disease. Albuminuria is the earliest marker of glomerular diseases (kidney damage), where it generally appears before the reduction in GFR. [64] A summary of potential mechanisms of albuminuria as a result of diabetic complications is shown in Fig (8). Early stages of CKD often show microalbuminuria (30-300 mg/g), which can progress to macroalbuminuria (>300 mg/g) with worsening kidney function. [63]
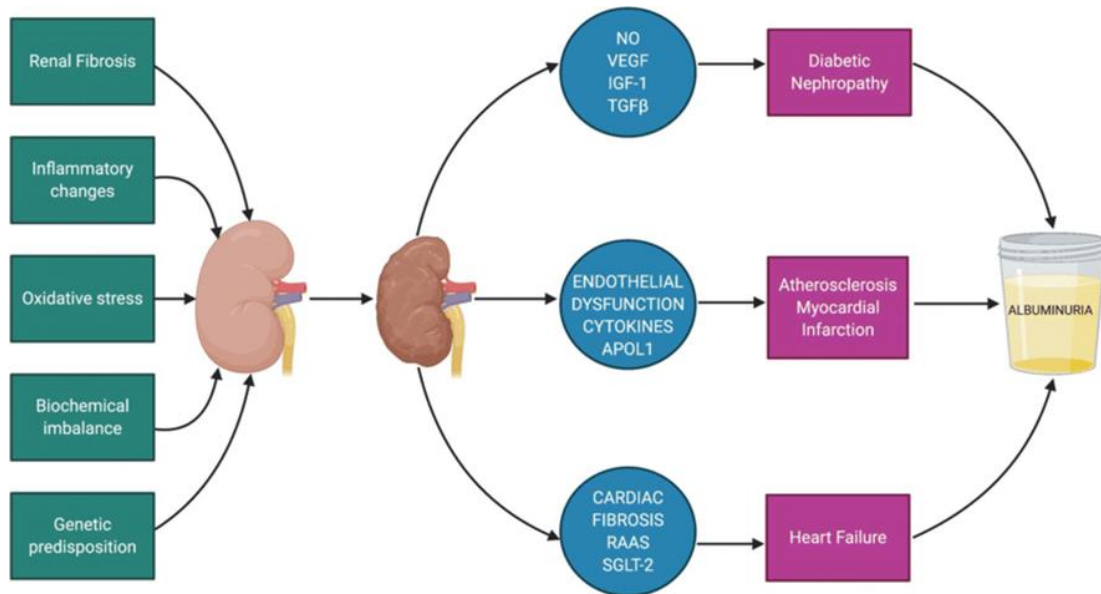
Figure (9). A summary of potential mechanisms of albuminuria as a result of micro and macrovascular diabetic complications. [64]

1.9.10. **Specific Gravity.**

The specific gravity (SG) test measures the concentration of solutes in the urine.[65] In CKD patients, it helps assess the kidney's ability to concentrate urine (depends on the functioning of the renal tubules), providing insights into kidney function and detecting potential abnormalities. CKD often impairs this function, leading to diluted urine even when dehydration should prompt concentrated urine. [26] High solute load in the urine may indicate kidney damage or impaired filtration.[65] SG testing can help detect early changes in kidney function before significant reductions in GFR occur and regular SG tests can monitor the progression of CKD. [26] For example, **Isosthenuria (SG ~1.010),** Indicates impaired tubular function, common in advanced CKD.[65]

1.10. **Importance of Early Prediction of Chronic Kidney Disease (CKD)**

- **Prevention of Disease Progression:** Early prediction allows for timely intervention strategies that can slow the progression of CKD. Lifestyle modifications, strict blood pressure control, and blood sugar management in diabetic patients can significantly delay the advancement to more severe stages of CKD. [42] Interventions initiated at early stages can help preserve remaining kidney function, thereby preventing or delaying the onset of end-stage renal disease (ESRD) which requires dialysis or kidney transplantation. [55]

- **Reduction in Morbidity:** Early identification and management of CKD can reduce complications associated with the disease, such as cardiovascular disease, which is the leading cause of death in CKD patients. Effective management of CKD risk factors contributes to lower morbidity and mortality rates.[26]

- **Enhanced Quality of Life:** Patients diagnosed early can avoid the severe symptoms associated with advanced CKD, such as fatigue, fluid retention, and cognitive impairment, thus maintaining a better quality of life. [55]

- **Cost-Effectiveness**: Early prediction and management of CKD can result in significant cost savings for healthcare systems. Treating advanced CKD and ESRD is expensive due to the high costs of dialysis and kidney transplantation. Preventative measures and early treatments are far less costly. [55

33

Chapter 2 – Machine Learning (ML) Role in (CKD).

2.1. Role of Bioinformatics in Chronic Kidney Disease (CKD).

Bioinformatics is an interdisciplinary field that combines biology, computer science, information technology, and statistics to analyze and interpret biological data. It involves the development and application of computational tools and techniques to understand biological processes and relationships. Bioinformatics is used extensively for the management and analysis of large sets of biological data, such as genomic sequences, protein structures, and metabolic pathways. [66] Bioinformatics plays a crucial role in advancing the understanding, early diagnosis, and treatment of CKD through various applications, including genomic analysis, biomarker discovery, data integration, building machine learning models for early diagnosis and personalized medicine. For example, bioinformatics tools are used in "Genomic and Proteomic Analysis" to identify genetic variants associated with CKD by analyzing large genomic datasets. [67] Proteomic analysis involves studying the protein composition of kidney tissues or urine samples. Bioinformatics tools can process mass spectrometry data to identify protein biomarkers that may indicate CKD or its progression.[68] Another example, bioinformatics tools help in **"**Drug Discovery and Development" by analyzing molecular data to understand the underlying mechanisms of CKD. Potential targets can include proteins, genes, or pathways involved in disease progression.,

such as specific kinases or cytokines, involved in CKD inflammation and fibrosis. [69]

2.2. AI and ML in Chronic Kidney Disease (CKD)

2.2.1. Artificial Intelligence (AI).

Artificial Intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think and learn. These systems can perform tasks that typically require human intelligence. [70] The term AI was coined by John McCarthy in 1956 during a conference held on this subject. However, the possibility of machines being able to simulate human behavior and actually think was raised earlier by Alan Turing who developed the Turing test in order to differentiate humans from machines. [71] AI encompasses a variety of technologies, including machine learning (ML), natural language processing (NLP), robotics, and neural networks. [70] More recently, AI has also begun to be incorporated into medicine to improve patient care by speeding up processes and achieving greater accuracy, opening the path to providing better healthcare overall. [71]

2.1.2. Machine Learning (ML).

Machine Learning (ML) is a subset of AI that involves the development of algorithms and statistical models that enable computers to perform specific tasks without using explicit instructions. [69] [72] The main purpose of ML is to introduce algorithms that ingest input data, use computer analysis to predict output values

within an acceptable range of accuracy, identify patterns and trends within the data, and learn from previous experience. [69] Instead, these systems rely on patterns and inference derived from data. ML techniques can be broadly classified into supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. [72] Algorithms are rules that precisely define a sequence of operations and examples include random forest algorithm (RF), support vector machine (SVM), and eXtreme gradient boosting (XGBoost). Because ML is highly effective in detecting hidden patterns in large datasets, its use in medicine can significantly improve the accuracy of diagnostic algorithms and personalize patient treatment. [69]

2.3. Role of AI and ML in Chronic Kidney Disease (CKD).

AI and ML has significant potential to transform the management and treatment of chronic kidney disease (CKD) through various applications, including early diagnosis, personalized treatment, predictive analytics, patient monitoring, [69] [73] as well as AI systems can detect anomalies in the performance of dialysis machines, alerting technicians to potential issues before they lead to equipment failure or patient harm. [73]
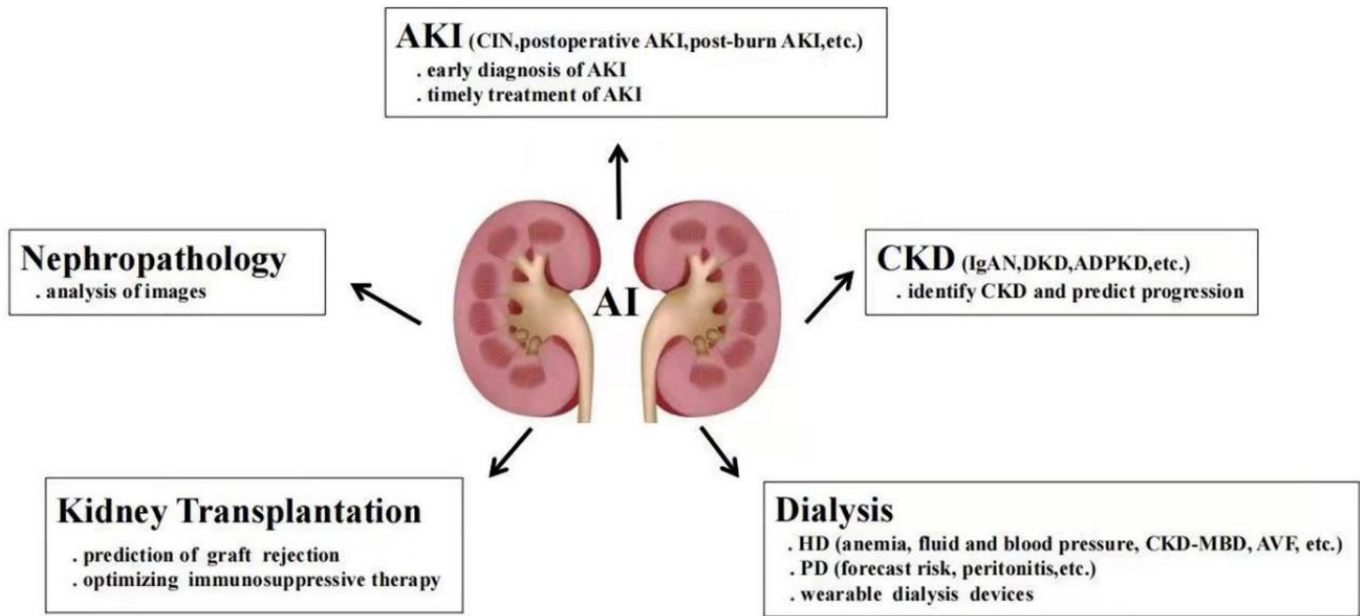
Figure (10): Clinical Applications of AI in Renal Disease. [69]

2.3.1**. Early Detection and Diagnosis**

- **Pattern Recognition:** ML algorithms can analyze large datasets from electronic health records (EHRs) to identify patterns and risk factors associated with CKD. This helps in the early detection of the disease, even before symptoms appear. [74] For example, supervised learning models can classify patients into different stages of CKD by analyzing clinical data, lab results, and imaging studies.[75]

- **Feature Selection:** ML techniques can identify the most relevant features (e.g., biomarkers, clinical parameters) that contribute to CKD, improving the accuracy and efficiency of diagnostic models.[76]

- **Predictive Analytics:** ML algorithms can predict the likelihood of CKD progression by analyzing longitudinal patient data. Predictive models can help

healthcare providers identify patients at high risk of rapid progression and intervene early to slow disease progression. [73]

- ML algorithms can analyze medical imaging (e.g., kidney ultrasounds, CT scans) to detect structural abnormalities and early signs of kidney damage, often with greater accuracy and speed than human radiologists. [74]

2.3.2. Personalized Treatment Plans.

ML algorithms can analyze individual patient data to recommend personalized treatment plans. This includes optimizing medication dosages, dietary recommendations, and lifestyle changes based on the patient's specific characteristics and disease stage. [78] ML-powered decision support systems provide clinicians with evidence-based recommendations for managing CKD. These systems can suggest the best treatment options, monitor patient progress, and alert providers to potential complications.[69] Patients can use wearable devices to track vital signs, blood pressure, and other health metrics, which are then analyzed by AI algorithms to detect any concerning trends. [79]

2.3.3. Enhancing Research and Clinical Trials.

ML can analyze large datasets from clinical trials and research studies to identify new insights into CKD pathophysiology, treatment efficacy, and patient outcomes. [75] For example, ML models can identify potential biomarkers for CKD, aiding in the development of new diagnostic tests and therapies. [80]

2.4. Developing a ML model for early prediction of CKD.

By following these steps, you can develop an effective machine learning model for early prediction of chronic kidney disease using Python as shown in Fig(11). First, **"Data Collection":** Gather a dataset containing relevant information about patients with chronic kidney disease. Second, **"Data Preprocessing":** This step is essential to ensure the quality of the data for training the model. Third, **"Feature Selection":** Identify the most important features that are likely to contribute to the prediction of chronic kidney disease. Forth, **"Model Selection and Training":** Choose an appropriate machine learning algorithm for building the prediction model then split the dataset into training and testing sets to train the model. Fifth, **"Model Evaluation":** Assess the performance of the trained model using evaluation metrics such as accuracy.
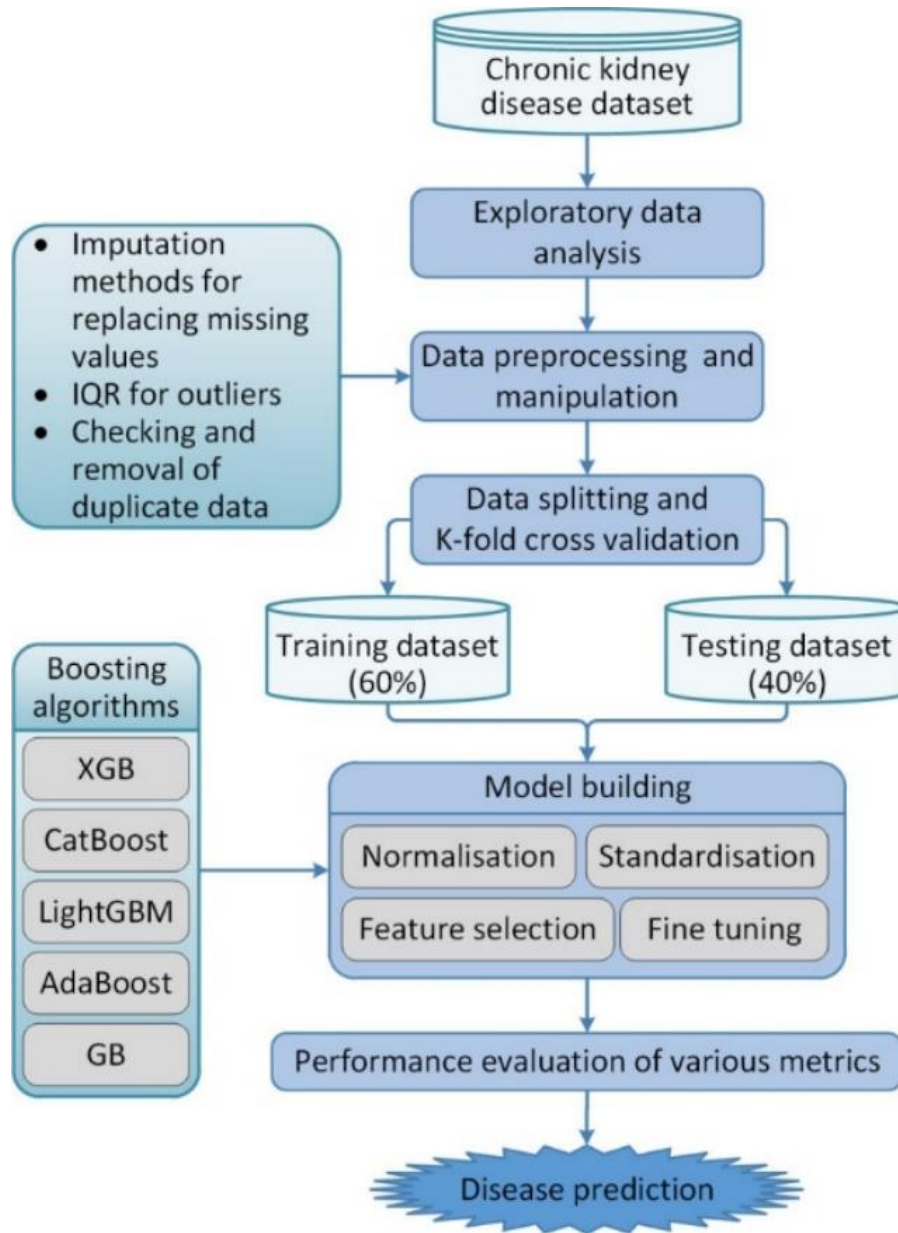
Figure (11). The workflow of the proposed ensemble learning based CKD prediction. [81]

Chapter 3– Methods and Materials.

3.1. Data Collection.

- Gather a dataset containing relevant information about patients with chronic kidney disease, including test results, and other relevant features.

- We get dataset from Kaggle.com which is publicly available at the UCI machine learning repository, for the experiment (https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease).

- The dataset contains 400 instances and 25 features in relation to Chronis kidney Disease. The first 24 features are predicate/independent, and the last one is a dependent/target attribute. Among the attributes, 11 are numeric, and 14 are categorical. The features are described in Table 3.

Table 3. features information of the dataset.

	Feature	Abbreviation	Type	description
1	Age	Age	numerical	Ages of participants taken in years
2	Blood Pressure	bp	numerical	Blood Pressure of participants taken in mm/Hg
3	Specific Gravity	sg	nominal	Urine specific gravity of the participant results (1.005,1.010,1.015,1.020,1.025)
4	Albumin	al	nominal	Albumin blood volume of the participant results (0,1,2,3,4,5)
5	Sugar	su	nominal	Participant's sugar level in the blood results (0,1,2,3,4,5)
6	Red Blood Cells	rbc	nominal	Red Blood Cells of urinalysis results (normal, abnormal)
7	Pus Cell	pc	nominal	Pus Cell of urinalysis results (normal, abnormal)

8	Pus Cell clumps	pcc	nominal	Presence of pus cell clumps in the participant's urine (present, not present)
9	Bacteria	ba	nominal	Presence of bacteria in the participant's urine (present, not present)
10	Blood Glucose Random	bgr	numerical	Blood Glucose Random in results in mgs/dl
11	Blood Urea	bu	numerical	Blood Urea results in mgs/dl
12	Serum Creatinine	sc	numerical	Serum Creatinine results in mgs/dl
13	Sodium	sod	numerical	Sodium level in the participant's blood results in mEq/L
14	Potassium	pot	numerical	Potassium level in the participant's blood results in mEq/L
15	Hemoglobin	hb	numerical	Hemoglobin measure in the participant's blood results in gms
16	Packed Cell Volume	Packed Cell Volume	numerical	Measure and size of RBCs in the participant's blood
17	White Blood Cell Count	wbc	numerical	WBCs count in the participant's blood results in cells/cmm
18	Red Blood Cell Count	rbc count	numerical	RBCs count in the participant's blood results in millions/cmm
19	Hypertension	htn	nominal	Dose the participant has Hypertension? (yes, no)
20	Diabetes Mellitus	dm	nominal	Dose the participant has Diabetes Mellitus? (yes, no)
21	Coronary Artery Disease	cad	nominal	Dose the patient has coronary artery disease? (yes, no)
22	Appetite	ppet	nominal	Participant's desire or need for something to eat (good, poor)
23	Pedal Edema	pe	nominal	Dose the participant has participant has swelling in the ankles and feet? (yes, no)
24	Anemia	ane	nominal	Dose the participant has Anemia? (yes, no)
25	Class	class	nominal	Class of the participant current kidney disease (CKD, not CKD)

3.2. Data Analysis and Preprocessing.

- This step is essential to ensure the quality of the data for training the model.

- We performed some preprocessing on the considered CKD dataset to make the dataset most usable. The purpose was to transform the available raw data into a format easily understood by the ensemble learning algorithms. The most important steps in this stage are, handling missing data, and encoding categorical features.

- We employed "Anaconda" for writing scientific codes in Python, to develop our model.

- Before working we imported many libraries using the following codes:

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

3.2.1. Handling Missing Values.

It involves various techniques depending on the type of data and the specific requirements of the analysis. [82] There were numerous missing values in the collected data. Therefore, the data must be clean of noise and complete in order to have reliable predictions for future decision making. The missing data were:

```
red blood cells            152
red blood cell count       131
white blood cell count     106
potassium                   88
sodium                      87
packed cell volume          71
 pus cell                   65
hemoglobin                  52
sugar                       49
specific gravity            47
albumin                     46
blood glucose random        44
blood urea                  19
serum creatinine            17
blood pressure              12
age                          9
bacteria                     4
pus cell clumps              4
hypertension                 2
diabetes mellitus            2
 coronary artery disease     2
anemia                       1
appetite                     1
pedal edema                  1
class                        0
dtype: int64
```

So, here we used "**median**" to fill missing values of nominal data and "**random values**"

in categorical data.

```
age                         0
blood pressure              0
specific gravity            0
albumin                     0
sugar                       0
red blood cells             0
 pus cell                   0
pus cell clumps             0
bacteria                    0
blood glucose random        0
blood urea                  0
serum creatinine            0
sodium                      0
potassium                   0
hemoglobin                  0
packed cell volume          0
white blood cell count      0
red blood cell count        0
hypertension                0
diabetes mellitus           0
coronary artery disease     0
appetite                    0
pedal edema                 0
anemia                      0
class                       0
dtype: int64
```

44

3.2.2. Label distribution of Nominal Data. We check the nominal data by drawing histograms describe the nominal features of the dataset and we got this plot:

Figure (12): Histogram of the nominal dataset features.

3.2.3. Label distribution of categorical Data: we did it by drawing bar charts describing

the categorical features of the dataset by and we got this output:
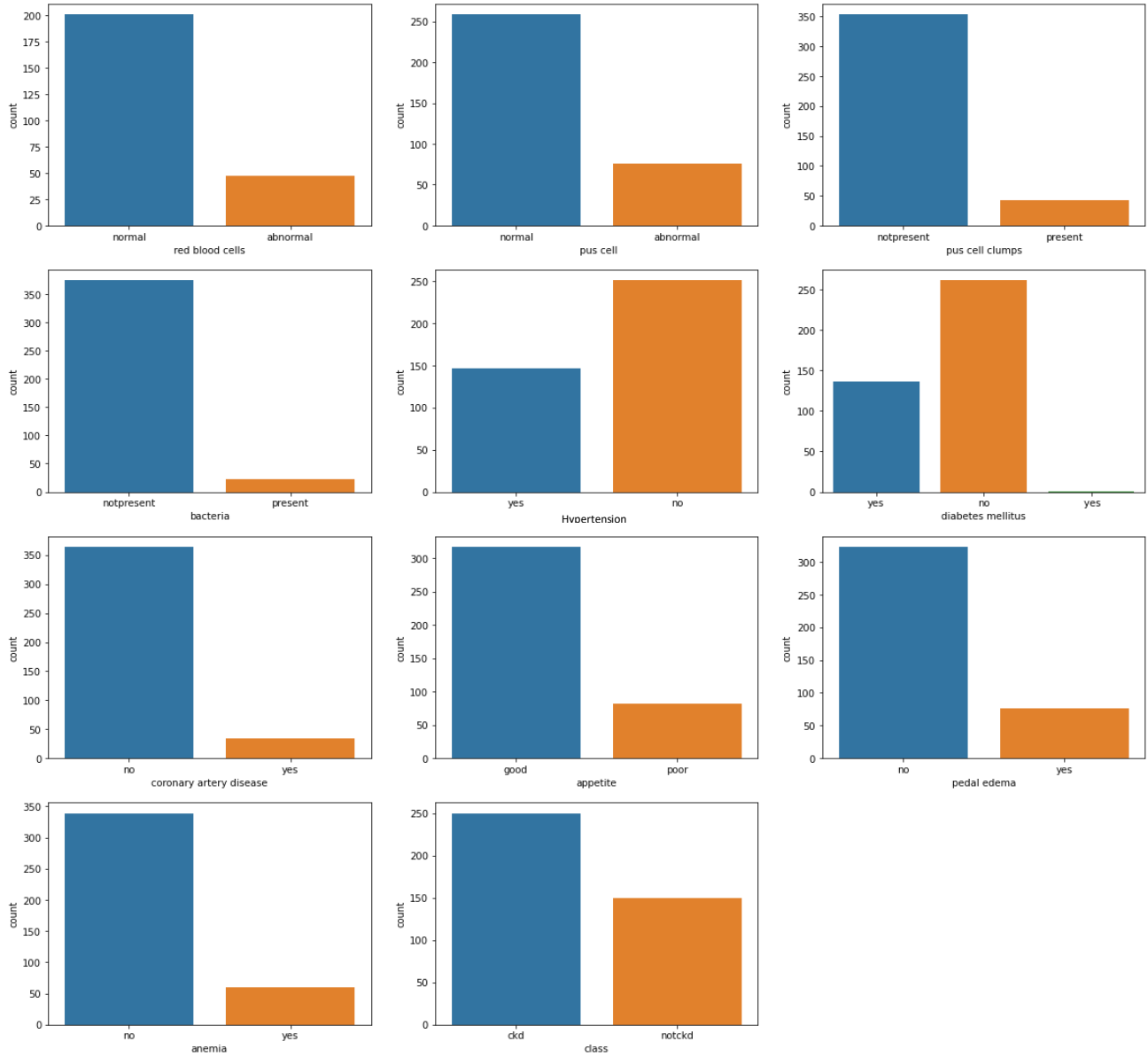
Figure (13): Charts of the categorical features.

3.2.4. CKD distribution

we did it by drawing the following plot Fig. (14), which shows that the largest number of participants in the dataset are CKD patients among 250 participants.
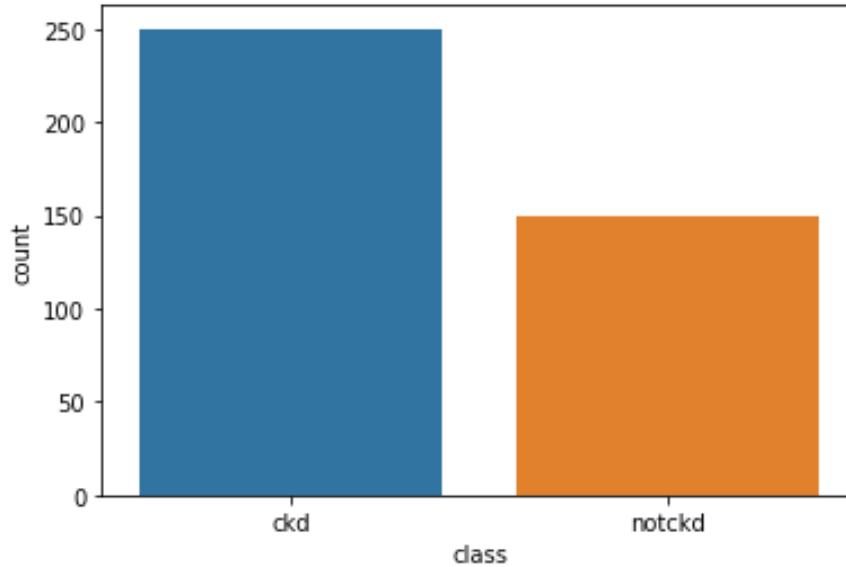
Figure (14): CKD distribution in the study sample

3.2.5. Encoding categorical features:

Machine learning models can only work with numerical values. For this reason, it is necessary to transform the categorical values of the relevant features into numerical ones. In this step we used label Encoding, because there are categories in each column. **"LabelEncoder"** can be used to normalize labels. It can also be used to transform non-numerical labels to numerical labels. [83]

3.3.   Feature Selection.

Identify the most important features that are likely to contribute to the prediction of chronic kidney disease.

3.3.1.   Data distribution.

We start studying by drawing the following plot for data features classification and distribution.
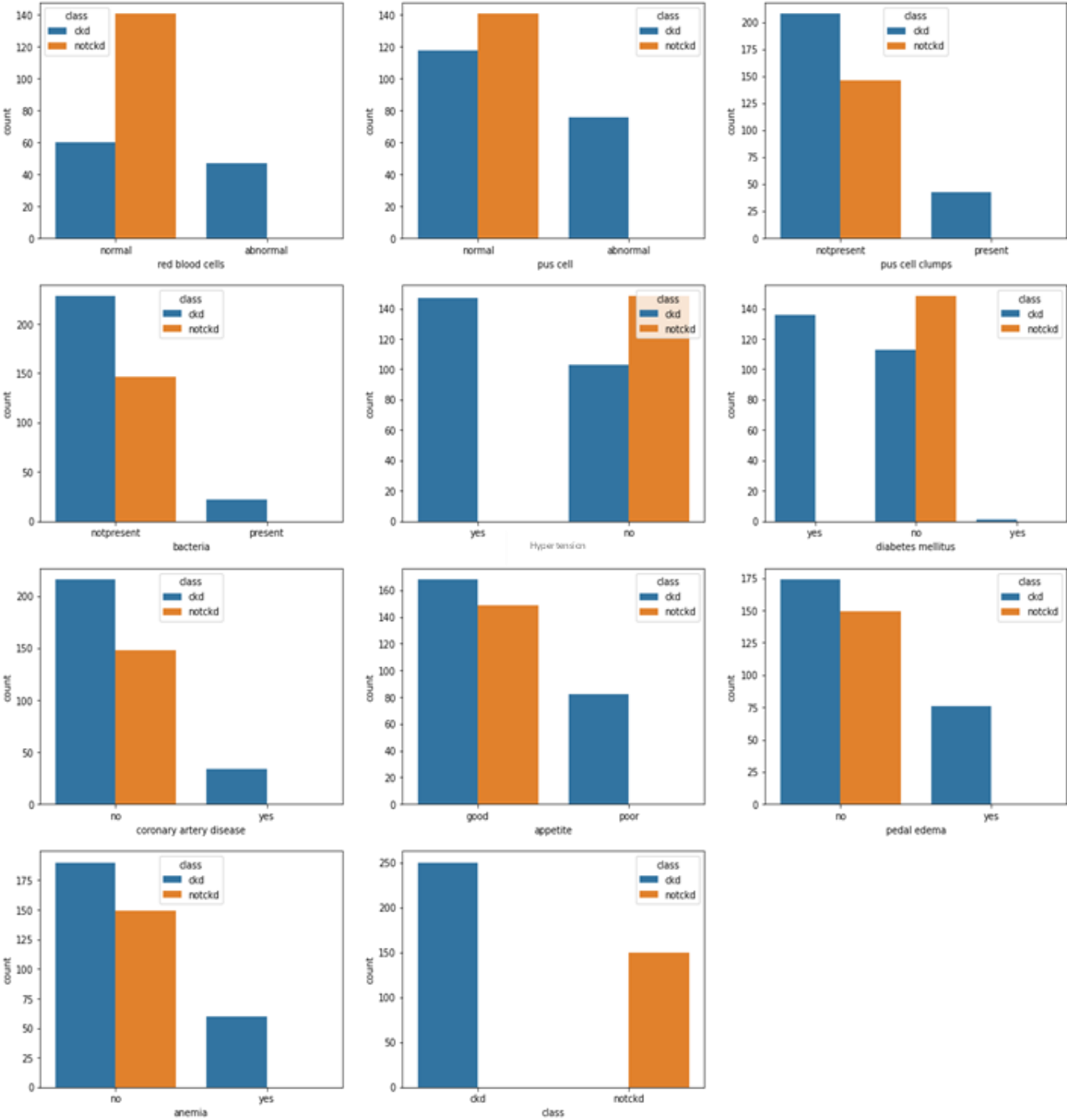

Figure (15): Features classification and distribution.

The output in Fig (14) shows that patients who have CKD are labeled with blue color and patients who haven't CKD are labeled with orange color. For example:

- "Red blood cell" plot shows that majority of patients have normal RBC urinalysis test results while most of CKD patients have abnormal results.

- "Hypertension" plot shows that majority of CKD patients have hypertension.

- "Diabetes mellites" plot shows that majority of CKD patients have diabetes mellites.

3.3.2. Correlation between features:

- Finding the correlation between features in a dataset is important for several reasons in data analysis and machine learning tasks.

- Identifying relationships: Correlation analysis helps in understanding the relationships between different features in a dataset. It helps to identify which features are positively correlated, negatively correlated, or not correlated at all. [84]

- Feature selection and Model performance: Correlation analysis can be used as a feature selection technique to identify redundant features. Understanding the correlation between features can help improve the performance of machine learning models and provides insights into how features are related to each other which can help in making informed decisions during data preprocessing. [85]

- strong relationship between the set of independent and dependent features indicates a good-quality dataset. In Fig (15) we draw a heatmap to illustrate correlation between features presents the CCA of the dataset features used in the

experiment. The relationship range lies between +1 to -1 along the X- and Y-axes. Fig (16) shows that:

✓ RBC count is positively correlated with specific gravity, hemoglobin, packed cell volume.

✓ RBC count is negatively correlated with albumin, blood urea.

✓ Packed cell volume and hemoglobin are highly positive correlated.

✓ Packed cell volume is negatively correlated with albumin and blood urea.

✓ Hemoglobin and albumin are negatively correlated.
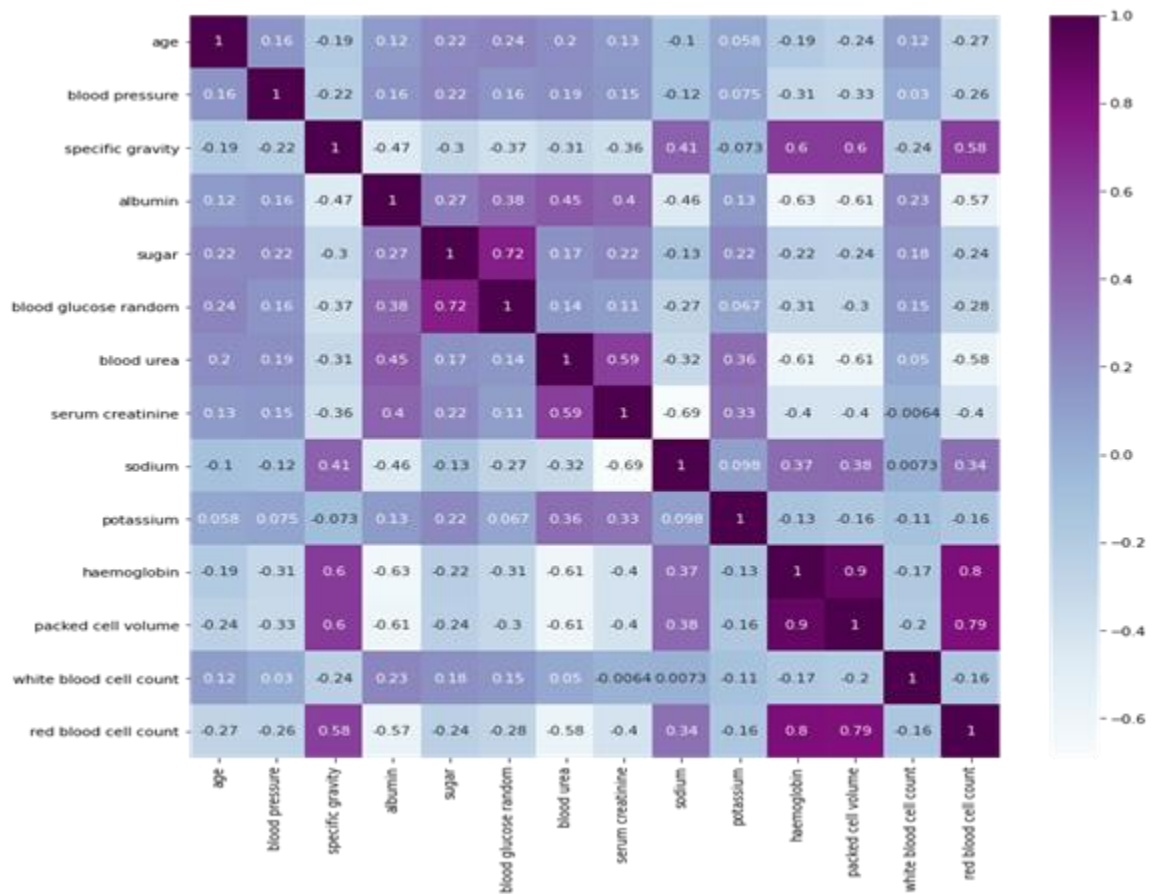
Figure (16): Correlation coefficient analysis of the dependent and independent features in the dataset.

3.3.3. Relationship between hemoglobin and red blood cell count.

For more intense, we draw scattered plot between hemoglobin and red blood cell count in CKD patients and participant who doesn't CKD, which gave some kind of linearity in all the relationships, whenever hemoglobin is below 13-14, he is positive for CKD, whenever hemoglobin is near 18, he is negative.
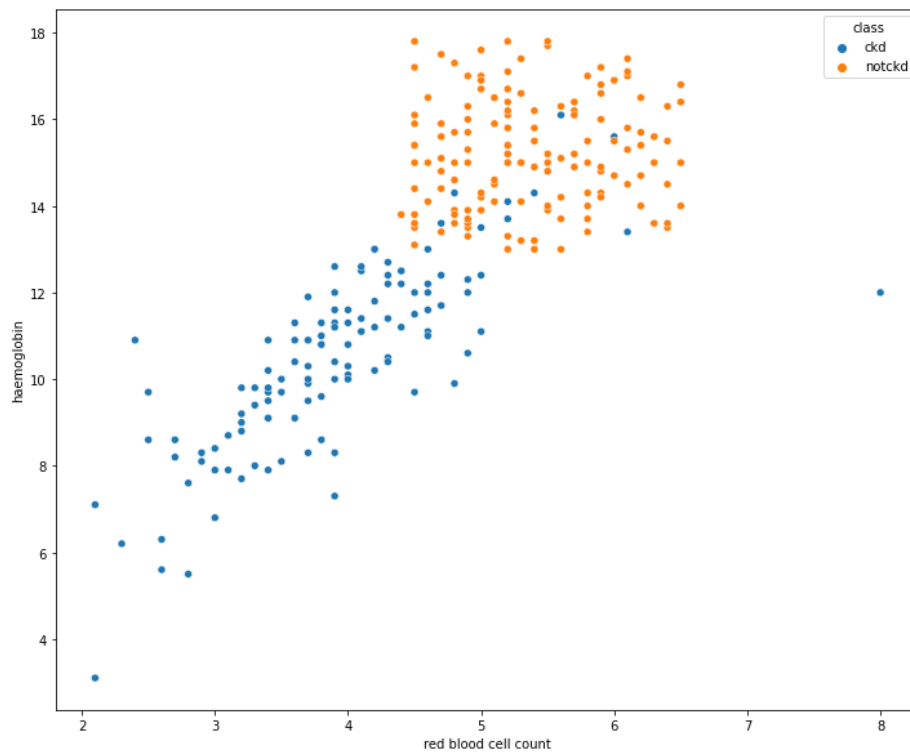
Figure (17): Relationship between hemoglobin and red blood cell count.

3.3.4. Analyze distribution of red blood cell count.

Here, by drawing distribution plot of red blood cell count distribution in CKD patients and participant who doesn't CKD, we can say that person with lower RBC count has high chances of having chronic kidney disease.
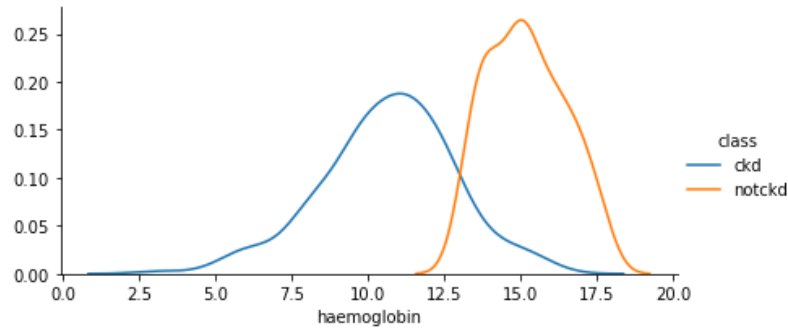
Figure (18): Red blood cell count distribution in study sample.

3.3.5. Analyze distribution of Hemoglobin.

Here, by drawing distribution plot of Hb distribution in CKD patients and participant who doesn't CKD, we can say that person with lower Hemoglobin has high chances of having chronic kidney disease.
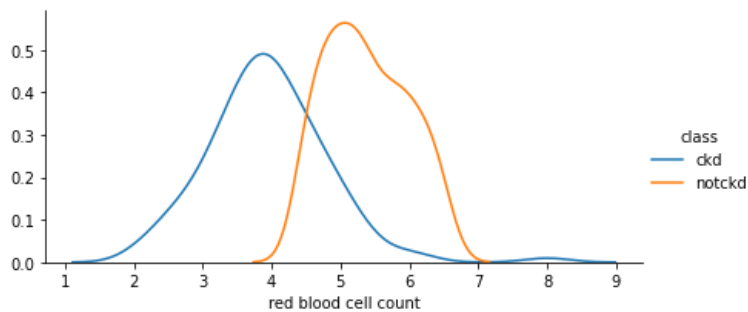
Figure (19): Hemoglobin distribution in study sample.

3.3.6. Selecting important features.

We used **"SelectKBest"** method which is a feature selection method in Python's scikit-learn library that selects the top k features based on univariate statistical tests such as "The chi-square" statistic which compares the size any discrepancies between the expected results and the actual results, given the size of the sample and the number of variables in the relationship. **SelectKBest** is commonly used to

	features	score
16	white blood cell count	9701.050391
10	blood urea	2343.097145
9	blood glucose random	2241.651289
11	serum creatinine	357.792101
15	packed cell volume	308.181415
3	albumin	216.000000
14	hemoglobin	123.856342
0	age	115.859940
4	sugar	94.800000
18	hypertension	88.200000

improve model performance by reducing the number of features, thereby helping to prevent overfitting and improving computational efficiency. [86] Here, we used the highest scored features which were:

3.4. Model selection and training.

In this step we should choose an appropriate machine learning algorithm for building the prediction model then split the dataset into training and testing sets to train the model on the training data and evaluate its performance on the testing data. The dataset was split into training (60%) and test (40%) subsets. In the

context of CKD prediction, machine learning has the potential to improve accuracy and reduce costs by identifying early signs of disease progression and predicting the risk of developing CKD in at-risk populations.[75]

3.4.1. Traditional machine learning techniques.

Traditional machine learning techniques suffer from some crucial limitations, including:[87]

- Overfitting, where the algorithm becomes too specialized to the training data and fails to generalize to new data.

- Large, high-quality datasets are needed to train and validate the algorithms, which can be challenging to obtain in some clinical settings.

- Training and evaluating machine learning algorithms may require considerable computational time and resources, especially for large datasets.

- High dependency on the quality and quantity of data available for training. If the data is incomplete, biased, or otherwise of poor quality, the resulting algorithm will be inaccurate or may not work at all.

- The machine learning algorithms can inadvertently incorporate biases present in the training data, leading to unfair or discriminatory outcomes.

3.4.2. Ensemble learning approaches.
The ensemble learning approaches are gaining attention for disease prediction with higher accuracy. Among the ensemble learning techniques, boosting algorithm is one of the effective approaches in the ensemble learning family.[88]

3.4.3. XGBoost.

XGBoost (eXtreme gradient boosting) is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way. [89]

[90] It works by combining different kinds of decision trees (weak learners) to calculate the similarity scores independently. [91] It helps to overcome the problem of overfitting during the training phase by adapting the gradient descent and regularization process. The mathematical formula for the XGBoost algorithm is shown in Eq:

$$f_\theta(x) = \sum_{m=1}^{T} \gamma_m h_m(x; \theta_m) = \sum_{m=1}^{T} \gamma_m l\left(x \in R_{jm}\right)$$

where $f_\theta(x)$ is XGBoost model with parameters θ, h_m is the mth weak decision tree with parameters θ_m, and γ_m is the weight associated with m^{th} tree. T denotes the number of decision trees, l denotes the loss function, and R_{jm} is an indicator function that returns 1 if x is in region R_{jm}, otherwise 0. [89]
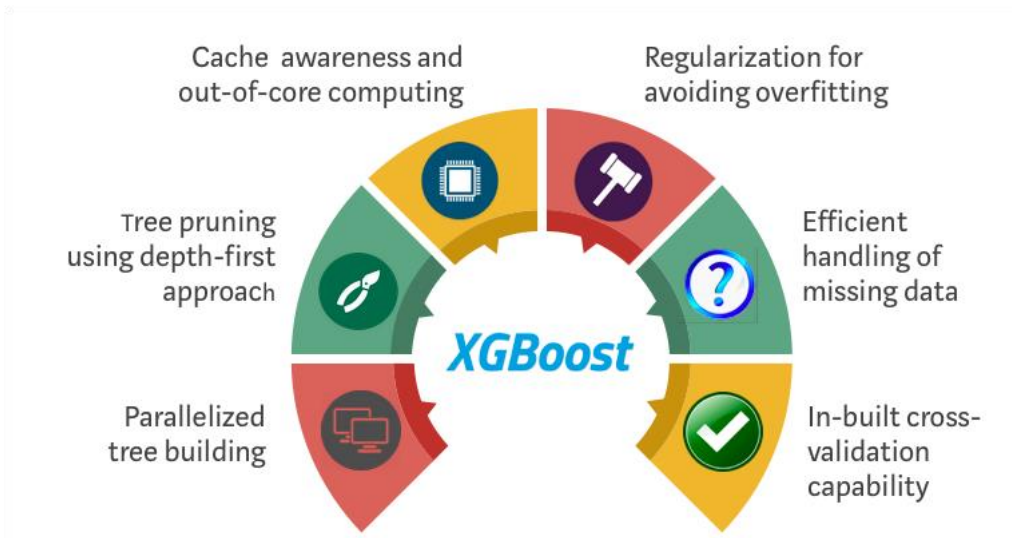
Figure (20): How XGBoost optimizes standard GBM algorithm. [92]

3.4.4. Model training.

3.4.4.1. XGBoost parameters.

XGBoost have parameters we used the following code to shows them:

Table.4 Description of some of XGBoost key parameters: [89]

Parameter	Description	Range
"eta"(or learning_rate):	Step size shrinkage used in update to prevent overfitting.	[0,0.5,0.20,0.25]
"max_depth"	Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit.	[5,8,10],
"min_child_weight"	Minimum sum of instance weight (hessian) needed in a child. Used to control overfitting.	[1,3,5,7],
"gamma"	Minimum loss reduction required to make a further	[0.0,0.1,0.2,0.4]

	partition on a leaf node. A larger value leads to more conservative models.	
"colsample_bytree"	Control the subsampling of features (columns) when constructing each tree in the ensemble. Specifically, it determines the fraction of features that will be randomly selected for building each individual tree. This can help to reduce overfitting and improve the model's generalization ability.	[0.3,0.4,0.7]

3.4.4.2.   Random search techniques:

The best parameters can vary based on your specific dataset and problem, so it's essential to experiment and validate your choices. This is done by the use of grid search or random search techniques to automate the hyperparameter tuning process. [93] In our case, data is very large so using random search techniques is the best choice. As its name suggests, they use random combinations of hyperparameters. This means that not all of the parameter values are tried, instead, parameters will be sampled with fixed numbers of iterations. [93] [94] **Random search techniques give us the best parameters;** we got this output:

```
{'min_child_weight': 3,
 'max_depth': 8,
 'learning-rate': 0.2,
 'gamma': 0.4,
 'colsample_bytree': 0.3}
```
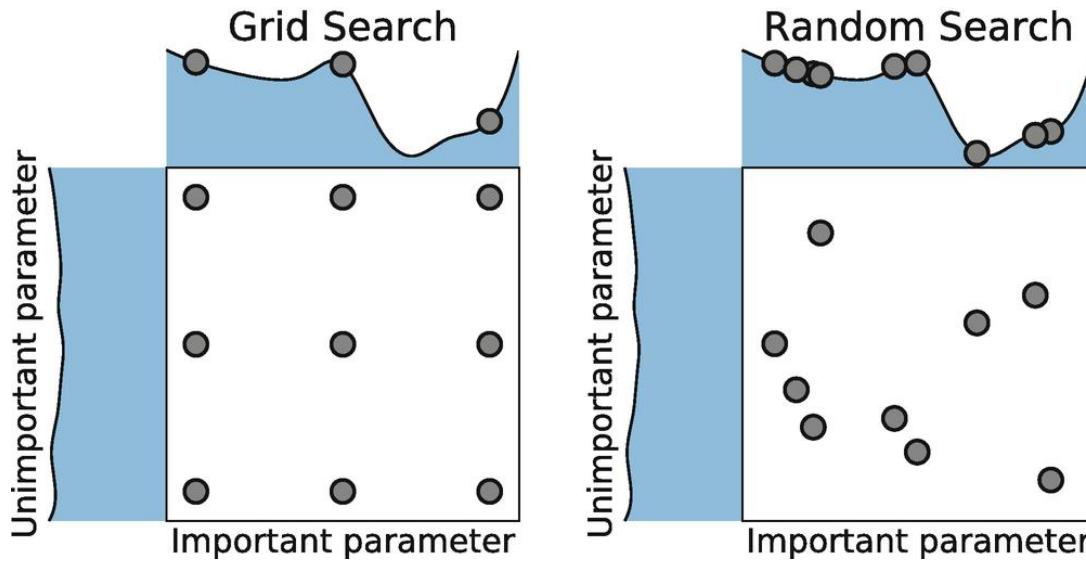
Figure (21): Grid search or random search techniques to automate the hyperparameter tuning process. [94]

3.5. Model evaluation:

Assess the performance of the trained model using evaluation metrics such as accuracy. In this section, the performance of the proposed prediction model for the XGBoost algorithms is discussed in terms of different performance metrics. Assess the performance of the trained model using evaluation metrics such as accuracy. Here's the code using "sklearn" for accuracy calculation:

```
from sklearn.metrics import confusion_matrix,accuracy_score
```

```
0.9833333333333333
```

As we performed all the methods and trained our model using different techniques and methods, we have got 98% prediction accuracy, which conceded a high accuracy score.

Conclusion, limitations, and future directions

Diagnosis and prevention of chronic kidney disease have become challenging for healthcare professionals and other concerned authorities. It can be mitigated to some extent if it can be pre-diagnosed in well advance. In this thesis, we attempted to predict CKD using one of ensemble learning approach the XGBoost. We employed different preprocessing techniques like the handling missing values, label distribution of nominal and categorical data, and encoding the categorical features. In addition, hyperparameter techniques like random search techniques were used to find the optimal parameter values. XGBoost emerged a high-performance accuracy (98%). Though the proposed model performed relatively well, it has some obvious limitations. The size of the considered dataset is small, which may limit the prediction model's performance in generic situations. It is observed that most of the features are having least contribution towards CKD. A more balanced dataset would lead to a better prediction model. As an extension of this work, other ensemble learning techniques, like bagging, stacking, etc., can be explored to improve the results. Additionally, deep learning techniques can also be experimented with the exercised dataset. To validate the effectivity of the proposed model, additional and larger datasets are needed in future. We expect more powerful disease prediction models to be developed and implemented in medical diagnosis and treatment.

References:

1. Johns Hopkins Medicine. "Anatomy of the Urinary System." *Johns Hopkins Medicine Health Library*, 2019, www.hopkinsmedicine.org/health/wellness-and-prevention/anatomy-of-the-urinary-system.

2. Smith, Ariana. "The Urinary Tract & How It Works | NIDDK." National Institute of Diabetes and Digestive and Kidney Diseases, 10 Apr. 2019, www.niddk.nih.gov/health-information/urologic-diseases/urinary-tract-how-it-works.

3. National Institute of Diabetes and Digestive and Kidney Diseases. "Your Kidneys & How They Work." National Institute of Diabetes and Digestive and Kidney Diseases, 2018, www.niddk.nih.gov/health-information/kidney-disease/kidneys-how-they-work.

4. McMahon, Robert S., et al. "Anatomy, Abdomen and Pelvis, Kidney Collecting Ducts." PubMed, StatPearls Publishing, 2020, www.ncbi.nlm.nih.gov/books/NBK549766/.

5. Ogobuiro, Ifeanyichukwu, and Faiz Tuma. "Physiology, Renal." Nih.gov, StatPearls Publishing, 24 July 2023, www.ncbi.nlm.nih.gov/books/NBK538339/.

6. Hopkins, Erin, et al. "Physiology, Acid Base Balance." National Library of Medicine, StatPearls Publishing, 12 Sept. 2022, www.ncbi.nlm.nih.gov/books/NBK507807/.

7. Mustaqeem, Ruqqiya, and Ali Arif. "Renal Tubular Acidosis." PubMed, StatPearls Publishing, 2020, www.ncbi.nlm.nih.gov/books/NBK519044/.

8. Roy, Ankita, et al. "Collecting Duct Intercalated Cell Function and Regulation." Clinical Journal of the American Society of Nephrology, vol. 10, no. 2, 28 Jan. 2015, pp. 305–324.

9. Fountain, John H., and Sarah L. Lappin. "Physiology, Renin Angiotensin System." PubMed, StatPearls Publishing, 2020, pubmed.ncbi.nlm.nih.gov/29261862/.

10. Patel, Vaibhav B., et al. "Role of the ACE2/Angiotensin 1–7 Axis of the Renin–Angiotensin System in Heart Failure." Circulation Research, vol. 118, no. 8, 15 Apr. 2016, pp.

11. Scott, Jonathan H., and Roberta J. Dunn. "Physiology, Aldosterone." PubMed, StatPearls Publishing, 2023, www.ncbi.nlm.nih.gov/books/NBK470339/.

12. Grillo, Andrea, et al. "Sodium Intake and Hypertension." Nutrients, vol. 11, no. 9, 21 Aug. 2019.

13. Cuzzo, Brian, and Sarah L Lappin. "Vasopressin (Antidiuretic Hormone, ADH)." National Library of Medicine, StatPearls Publishing, 2019, www.ncbi.nlm.nih.gov/books/NBK526069/.

14. Hering, Lydia, et al. "Role of α2-Adrenoceptors in Hypertension: Focus on Renal Sympathetic Neurotransmitter Release, Inflammation, and Sodium Homeostasis." Frontiers in Physiology, vol. 11, 2020, p. 566871.

15. Bhoopalan, Senthil Velan, et al. "Erythropoietin Regulation of Red Blood Cell Production: From Bench to Bedside and Back." F1000Research, vol. 9, no. 1, 18 Sept. 2020, p. 1153.

16. Martin, Kevin J, and Esther A González. "Vitamin D and the Kidney." Missouri Medicine, vol. 109, no. 2, 2012, pp. 124–6.

17. Makris, Konstantinos, and Loukia Spanou. "Acute Kidney Injury: Definition, Pathophysiology and Clinical Phenotypes." The Clinical Biochemist. Reviews, vol. 37, no. 2, May 2016, pp. 85–98.

18. Goyal, Abhinav, et al. "Acute Kidney Injury (Acute Renal Failure)." PubMed, StatPearls Publishing, 25 Nov. 2023, www.ncbi.nlm.nih.gov/books/NBK441896/.

19. Khwaja, Arif. "KDIGO Clinical Practice Guidelines for Acute Kidney Injury." Nephron, vol. 120, no. 4, 7 Aug. 2012, pp. c179–c184.

20. Webster, Angela C, et al. "Chronic Kidney Disease." The Lancet, vol. 389, no. 10075, 25 Mar. 2017, pp. 1238–1252.

21. Charles, Cornelia, and Allison H. Ferris. "Chronic Kidney Disease." Primary Care: Clinics in Office Practice, vol. 47, no. 4, Sept. 2020.

22. NIDDK. "What Is Polycystic Kidney Disease? | NIDDK." National Institute of Diabetes and Digestive and Kidney Diseases, 10 May 2019, www.niddk.nih.gov/health-information/kidney-disease/polycystic-kidney-disease/what-is-pkd.

23. Sung, Bong Mo, et al. "Chronic Kidney Disease in Neurogenic Bladder." Nephrology, vol. 23, no. 3, 20 Feb. 2018, pp. 231–236.

24. . Lucero, Claudia M., et al. "Hypertensive Nephropathy: Unveiling the Possible Involvement of Hemichannels and Pannexons." International Journal of Molecular Sciences, vol. 23, no. 24, 14 Dec. 2022, p. 15936.

25. "Chapter 1: Definition and Classification of CKD." Kidney International Supplements, vol. 3, no. 1, Jan. 2013, pp. 19–62.

26. Vaidya, Satyanarayana R., and Narothama R. Aeddula. "Chronic Renal Failure." PubMed, StatPearls Publishing, 2020, pubmed.ncbi.nlm.nih.gov/30571025/.

27. "Chapter 2: Definition, Identification, and Prediction of CKD Progression." Kidney International Supplements, vol. 3, no. 1, Jan. 2013, pp. 63–72.

28. Arieff, Allen I. "Diabetic Nephropathy and Treatment of Hypertension." PubMed, MDText.com, Inc., 2000, www.ncbi.nlm.nih.gov/books/NBK279103/.

29. Samsu, Nur. "Diabetic Nephropathy: Challenges in Pathogenesis, Diagnosis, and Treatment." BioMed Research International, vol. 2021, 8 July 2021, pp. 1–17.

30. Moarij Qazi, 1, et al. "Pathophysiology of Diabetic Kidney Disease." EMJ Nephrol Nephrology 10.1 2022, vol. 10, no. 1, 30 June 2022, pp. 102–113.

31. Kumar, Mukesh, et al. "The Bidirectional Link between Diabetes and Kidney Disease: Mechanisms and Management." Cureus, vol. 15, no. 9, 20 Sept. 2023.

32. Yang, Yang, and Gaosi Xu. "Update on Pathogenesis of Glomerular Hyperfiltration in Early Diabetic Kidney Disease." Frontiers in Endocrinology, vol. 13, 19 May 2022.

33. Twarda-Clapa, Aleksandra, et al. "Advanced Glycation End-Products (AGEs): Formation, Chemistry, Classification, Receptors, and Diseases Related to AGEs." Cells, vol. 11, no. 8, 1 Jan. 2022, p. 1312.

34. Ameer, Omar Z. "Hypertension in Chronic Kidney Disease: What Lies behind the Scene." Frontiers in Pharmacology, vol. 13, no. 949260, 11 Oct. 2022.

35. Gupta, Ankur, et al. "Hypertension in Chronic Kidney Disease: An Update on Diagnosis and Management." Southern Medical Journal, vol. 116, no. 2, Feb. 2023, pp. 237–244.

36. Freedman, Barry I., and Arthur H. Cohen. "Hypertension-Attributed Nephropathy: What's in a Name?" Nature Reviews Nephrology, vol. 12, no. 1, 10 Nov. 2015, pp. 27–36.

37. Kaur, Jasdeep, et al. "Sympathetic Overactivity in Chronic Kidney Disease: Consequences and Mechanisms." International Journal of Molecular Sciences, vol. 18, no. 8, 2 Aug. 2017.

38. Hamrahian, Seyed Mehrdad, and Bonita Falkner. "Hypertension in Chronic Kidney Disease." Advances in Experimental Medicine and Biology, vol. 956, 2017, pp. 307–325.

39. Gallo, Giovanna, et al. "Endothelial Dysfunction in Hypertension: Current Concepts and Clinical Implications." Frontiers in Medicine, vol. 8, no. 1, 20 Jan. 2022.

40. Wen, Yi, and Steven D. Crowley. "Renal Effects of Cytokines in Hypertension." Current Opinion in Nephrology and Hypertension, vol. 27, no. 2, 1 Mar. 2018, pp. 70–76.

41. Panizo, Sara, et al. "Fibrosis in Chronic Kidney Disease: Pathogenesis and Consequences." International Journal of Molecular Sciences, vol. 22, no. 1, 2 Jan. 2021, p. 408.

42. Stevens, Paul E, et al. "KDIGO 2024 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease." Kidney International, vol. 105, no. 4, 1 Apr. 2024, pp. S117–S314.

43. Zhang, Yaru, et al. "ACE Inhibitor Benefit to Kidney and Cardiovascular Outcomes for Patients with Non-Dialysis Chronic Kidney Disease Stages 3–5: A Network Meta-Analysis of Randomised Clinical Trials." Drugs, vol. 80, no. 8, 24 Apr. 2020, pp. 797–811.

44. Ali, Sehrish, et al. "Revisiting Diuretic Choice in Chronic Kidney Disease." Current Opinion in Nephrology & Hypertension, vol. 31, no. 5, 11 July 2022, pp. 406–413.

45. Grabie, M., et al. "Effects of Methyldopa on Renal Hemodynamics and Tubular Function." Clinical Pharmacology and Therapeutics, vol. 27, no. 4, 1 Apr. 1980, pp. 522–527.

46. Hahr, Allison J., and Mark E. Molitch. "Management of Diabetes Mellitus in Patients with Chronic Kidney Disease." Clinical Diabetes and Endocrinology, vol. 1, no. 1, 4 June 2015.

47. Schoener, Benjamin, and Judith Borger. "Erythropoietin Stimulating Agents." PubMed, StatPearls Publishing, 11 Mar. 2023, www.ncbi.nlm.nih.gov/books/NBK536997/.

48. Gutiérrez, Orlando M. "Treatment of Iron Deficiency Anemia in CKD and End-Stage Kidney Disease." Kidney International Reports, vol. 6, no. 9, Sept. 2021, pp. 2261–2269.

49. Goyal, Rajeev, and Ishwarlal Jialal. "Hyperphosphatemia." PubMed, StatPearls Publishing, 2020, www.ncbi.nlm.nih.gov/books/NBK551586/.

50. Muppidi, Vijayadershan, et al. "Secondary Hyperparathyroidism." PubMed, StatPearls Publishing, 2020, www.ncbi.nlm.nih.gov/books/NBK557822/.

51. "Dialysis in Chronic Kidney Disease." Nih.gov, Institute for Quality and Efficiency in Health Care (IQWiG), 8 Mar. 2018, www.ncbi.nlm.nih.gov/books/NBK492979/.

52. Murdeshwar, Himani N., and Fatima Anjum. "Hemodialysis." PubMed, StatPearls Publishing, 2023, www.ncbi.nlm.nih.gov/books/NBK563296/.

53. Francis, Anna, et al. "Chonic Kidney Disease and the Global Public Health Agenda: An International Consensus." Nature Reviews Nephrology, 3 Apr. 2024, pp. 1–13.

54. Kovesdy, Csaba P. "Epidemiology of Chronic Kidney Disease: An Update 2022." Kidney International Supplements, vol. 12, no. 1, Apr. 2022, pp. 7–11.

55. Benjamin, Onecia, and Sarah L Lappin. "End-Stage Renal Disease." National Library of Medicine, StatPearls Publishing, 2023, www.ncbi.nlm.nih.gov/books/NBK499861/.

56. Gounden, Verena, and Ishwarlal Jialal. "Renal Function Tests." National Library of Medicine, StatPearls Publishing, 2023, www.ncbi.nlm.nih.gov/books/NBK507821/.

57. Shahbaz, Hassan, and Mohit Gupta. "Creatinine Clearance." PubMed, StatPearls Publishing, 20 May 2023.

58. Galindo, Rodolfo J, et al. "Glycemic Monitoring and Management in Advanced Chronic Kidney Disease." Endocrine Reviews, vol. 41, no. 5, 26 May 2020.

59. Mondal, Himel, and Saran Lotfollahzadeh. "Hematocrit (HCT)." Nih.gov, StatPearls Publishing, 2 Jan. 2023.

60. Shaikh, Hira, and Narothama R. Aeddula. "Anemia of Chronic Renal Disease." PubMed, StatPearls Publishing, 24 Feb. 2023.

61. Yen, Chiung-Hui, et al. "The Prognostic Value of Peripheral Total and Differential Leukocyte Count in Renal Progression: A Community-Based Study." PloS One, vol. 16, no. 10, 28 Oct. 2021, pp. e0258210–e0258210.

62. Srivastava, Anand, et al. "Uric Acid and the Risks of Kidney Failure and Death in Individuals with CKD." American Journal of Kidney Diseases, vol. 71, no. 3, Mar. 2018, pp. 362–370.

63. Christofides, Elena A., and Niraj Desai. "Optimal Early Diagnosis and Monitoring of Diabetic Kidney Disease in Type 2 Diabetes Mellitus: Addressing the Barriers to Albuminuria Testing." Journal of Primary Care & Community Health, vol. 12, Jan. 2021, p. 215013272110036.

64. Raja, Pappitha, et al. "The Potential of Albuminuria as a Biomarker of Diabetic Complications." Cardiovascular Drugs and Therapy, vol. 35, no. 3, 17 July 2020.

65. Queremel Milani, Daniel A., and Ishwarlal Jialal. "Urinalysis." PubMed, StatPearls Publishing, 2021, www.ncbi.nlm.nih.gov/books/NBK557685.

66. National Human Genome Research Institute. "Bioinformatics." Genome.gov, 2022, www.genome.gov/genetics-glossary/Bioinformatics.

67. Tin, Adrienne, and Anna Köttgen. "Genome-Wide Association Studies of CKD and Related Traits." Clinical Journal of the American Society of Nephrology, vol. 15, no. 11, 14 May 2020, pp.

68. . Øvrehus, Marius A, et al. "Urinary Proteomics in Chronic Kidney Disease: Diagnosis and Risk of Progression beyond Albuminuria." Clinical Proteomics, vol. 12, no. 1, 7 Aug. 2015.

69. Yao, Lijing, et al. "Application of Artificial Intelligence in Renal Disease." Clinical EHealth, vol. 4, 2021, pp. 54–61.

70. Collins, Christopher, et al. "Artificial Intelligence in Information Systems Research: A Systematic Literature Review and Research Agenda." International Journal of Information Management, vol. 60, no. 102383, 2021.

71. Mintz, Yoav, and Ronit Brodie. "Introduction to Artificial Intelligence in Medicine." Minimally Invasive Therapy & Allied Technologies, vol. 28, no. 2, 27 Feb. 2019, pp. 73–81.

72. Pugliese, Raffaele, et al. "Machine Learning-Based Approach: Global Trends, Research Directions, and Regulatory Standpoints." Data Science and Management, vol. 4, Dec. 2021, pp. 19–29.

73. Chaudhuri, Sheetal, et al. "Artificial Intelligence Enabled Applications in Kidney Disease." Seminars in Dialysis, vol. 34, no. 1, 13 Sept. 2020, pp. 5–16.

74. Krisanapan, Pajaree, et al. "Revolutionizing Chronic Kidney Disease Management with Machine Learning and Artificial Intelligence." Journal of Clinical Medicine, vol. 12, no. 8, 21 Apr. 2023, p. 3018.

75. Ghosh, Samit Kumar, and Ahsan H. Khandoker. "Investigation on Explainable Machine Learning Models to Predict Chronic Kidney Diseases." Scientific Reports, vol. 14, no. 1, 14 Feb. 2024, p. 3687.

76. Rashed-Al-Mahfuz, Md., et al. "Clinically Applicable Machine Learning Approaches to Identify Attributes of Chronic Kidney Disease (CKD) for Use in Low-Cost Diagnostic Screening." IEEE Journal of Translational Engineering in Health and Medicine, vol. 9, 2021, pp. 1–11,

77. Pinto-Coelho, Luís. "How Artificial Intelligence Is Shaping Medical Imagin Technology: A Survey of Innovations and Applications." Bioengineering, vol. 10, no. 12, 18 Dec. 2023, p. 1435.

78. Johnson, Kevin B., et al. "Precision Medicine, AI, and the Future of Personalized Health Care." Clinical and Translational Science, vol. 14, no. 1, 12 Oct. 2020.

79. Shaghayegh Shajari, et al. "The Emergence of AI-Based Wearable Sensors for Digital Health Technology: A Review." Sensors, vol. 23, no. 23, 29 Nov. 2023, pp. 9498–9498.

80. Delrue, Charlotte, et al. "Application of Machine Learning in Chronic Kidney Disease: Current Status and Future Prospects." Biomedicines, vol. 12, no. 3, 1 Mar. 2024, p. 568.

81. Shahid Mohammad Ganie, et al. "Chronic Kidney Disease Prediction Using Boosting Techniques Based on Clinical Parameters." PloS One, vol. 18, no. 12, 1 Dec. 2023, pp. e0295234–e0295234.

82. Ullah, Zahid, and Mona Jamjoom. "Early Detection and Diagnosis of Chronic Kidney Disease Based on Selected Predominant Features." Journal of Healthcare Engineering, vol. 2023, 30 Jan. 2023, p. 3553216.

83. Deniz Gunay. "Feature Encoding - Deniz Gunay - Medium." Medium, Medium, 17 Aug. 2023, medium.com/@denizgunay/feature-encoding-f099a6c1abe8.

84. Kumar, Sunil, and Ilyoung Chong. "Correlation Analysis to Identify the Effective Data in Machine Learning: Prediction of Depressive Disorder and Emotion States." International Journal of Environmental Research and Public Health, vol. 15, no. 12, 19 Dec. 2018, p. 2907.

85. Ashraf, Abdallah. "Correlation in Machine Learning — All You Need to Know." Medium, 22 Sept. 2023, medium.com/@abdallahashraf90x/all-you-need-to-know-about-correlation-for-machine-learning-e249fec292e9.

86. D, Kavya. "Optimizing Performance: SelectKBest for Efficient Feature Selection in Machine Learning." Medium, 16 Feb. 2023, medium.com/@Kavya2099/optimizing-performance-selectkbest-for-efficient-feature-selection-in-machine-learning-3b635905ed48.

87. Onyema E. M., Almuzaini K. K., Onu F. U., Verma D., Gregory U. S., Puttaramaiah M. and Afriyie R. K.,"Prospects and Challenges of Using Machine Learning for Academic Forecasting," Computational Intelligence and Neuroscience, vol. 2022.

88. Mahajan, Prerna, et al. "Ensemble Learning for Disease Prediction: A Review." Healthcare, vol. 11, no. 12, 20 June 2023, pp. 1808–1808.

89. xgboost developers. "XGBoost Documentation — Xgboost 1.5.1 Documentation." Xgboost.readthedocs.io, 2022, xgboost.readthedocs.io/en/stable/.

90. Morde, Vishal. "XGBoost Algorithm: Long May She Reign!" Medium, 8 Apr. 2019, towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d?gi=f88bde1bac09.

91. Santhanam R., Uzir N., Raman S. and Banerjee S., "Experimenting XGBoost algorithm for prediction and classification of different datasets," International Journal of Control Theory and Applications, vol. 9,no. 40, pp. 651–662, 2016.

92. Memon, Nimrabanu, et al. "Comparative Analysis of Artificial Neural Network and XGBoost Algorithm for PolSAR Image Classification." Lecture Notes in Computer Science, 1 Jan. 2019, pp. 452–460.

93. Hestisholihah. "Hyperparameter Tuning Showdown: Grid Search vs. Random Search — Which Is the Ultimate Winner?" Medium, 19 June 2023, medium.com/@hestisholihah01/hyperparameter-tuning-showdown-grid-search-vs-random-search-which-is-the-ultimate-winner-5927b322e54d.

94. Lee, Elyse. "An Intro to Hyper-Parameter Optimization Using Grid Search and Random Search." Medium, 5 June 2019, medium.com/@cjl2fv/an-intro-to-hyper-parameter-optimization-using-grid-search-and-random-search-d73b9834ca0a.