SEBHA UNIVERSITY CONFERENCE PROCEEDINGS, VOL.04 NO. 3 2025
DOI: 10.51984/SUCP.V4I3.4177

وقائع مؤتمرات جامعة سبها
Sebha University Conference Proceedings

Confrence Proceeeding homepage: http://www.sebhau.edu.ly/journal/CAS

تحليل استكشافي لخصائص مرض الكلى المزمن: الجنوب الليبي كحالة دراسة

*نورا علي[1] حموده شفتر [1] سلوى علي [2] تجديدة امغاير [3]

[1]قسم علوم الحاسب، كلية تقنية المعلومات، جامعة سبها، ليبيا
[2] قسم الشبكات والاتصالات، كلية تقنية المعلومات، جامعة سبها، ليبيا
[3]قسم الباطنة، كلية الطب البشري، جامعة سبها، ليبيا

الملخص

يُعد مرض الكلى المزمن (CKD) تحدياً صحياً عالمياً يتطلب تحسين جودة التشخيص وإدارة المرض عبر استخدام تقنيات متقدمة في معالجة البيانات. تستعرض هذه الدراسة تجربة تحليل وتجهيز بيانات مرضى الكلى المزمن في الجنوب الليبي والموجودة بمراكز الكلى بمدينة سبها، حيث تم جمع بيانات من 1000 سجل طبي تشمل 30 متغيراً سريرياً ومخبرياً. أظهرت البيانات تبايناً في نسب القيم المفقودة، مما استلزم تطبيق منهجيات معالجة مسبقة لضمان جودة البيانات وصلاحيتها للتحليل. بعد استبعاد المتغيرات غير المناسبة، تم تطبيق تقنيات إحصائية وتحليلية لتحديد المؤشرات الحيوية الأكثر ارتباطاً بتطور المرض، وذلك تمهيداً لبناء نماذج تنبؤية تعتمد على الذكاء الاصطناعي لتحسين التشخيص المبكر ودعم اتخاذ القرار الطبي. تؤكد النتائج على أهمية تجهيز البيانات وتنقيتها كخطوة أساسية في الاستفادة من تقنيات التعلم الآلي في مجال الرعاية الصحية، مما يساهم في تحسين جودة الرعاية وتقليل العبء الصحي والاقتصادي لمرضى الكلى المزمن.

الكلمات المفتاحية:

نموذج تنبؤي
مرض الكلى المزمن
تحليل البيانات

Exploratory Analysis of Chronic Kidney Disease Features: Southern Libya as a Case Study

*Noura Alahwel[1], Hamouda Chantar[1], Salwa Ali[2], Tajdida Magayr[3]

[1]Department of computer scince, faculty of information technology, sebha university, Libya
[2]Department of Network and Communication, faculty of information technology, sebha university, Libya
[3]Department of Inner Medicine, faculty of Medicine, Sebha Unversity, Libya

A B S T R A C T

Chronic Kidney Disease (CKD) is a global health challenge that necessitates improved diagnostic quality and disease management through the use of advanced data processing techniques. This study presents an analysis and preparation of CKD patient data in southern Libya, specifically from kidney centers in the city of Sebha. Data were collected from 1,000 medical records, including 30 clinical and laboratory variables. The dataset revealed varying levels of missing values, necessitating the application of preprocessing methodologies to ensure data quality and suitability for analysis. After excluding irrelevant variables, statistical and analytical techniques were applied to identify the biomarkers most strongly associated with disease progression. This serves as a foundation for developing predictive models based on artificial intelligence to enhance early diagnosis and support medical decision-making. The findings emphasize the importance of data preparation and cleaning as a fundamental step in leveraging machine learning techniques in healthcare, contributing to improved care quality and a reduction in the health and economic burden of CKD patients.

Keywords:

Predictive Model
CKD
Data analysis

1. المقدمة

في ظل التطور السريع الذي يشهده العالم في المجال الصحي، يظل مرض الكلى المزمن (CKD) أحد أبرز التحديات الصحية العالمية التي تواجه أنظمة الرعاية الطبية .تشير الإحصاءات إلى أن ما يقرب من 10%من سكان العالم يعانون من هذا المرض المزمن، الذي يتميز بتدهور تدريجي غير قابل للعكس في وظائف الكلى، مما يهدد حياة المرضى ويستلزم في مراحله المتقدمة اللجوء إلى غسيل الكلى أو زراعة الكلى .ويزداد هذا العبء الطبي والاقتصادي مع ارتفاع معدلات الإصابة وتفاقم المرض نتيجة لتأخر التشخيص، إذ أن هذا المرض غالبًا ما

*Corresponding author:
E-mail addresses: sal.ali1@sebhau.edu.ly
Article History : Received 20 February 2025 - Received in revised form 01 September 2025 - Accepted 07 October 2025

يتطور بصورة صامتة، ولا تظهر أعراضه إلا في مراحل متقدمة يصعب معها تقديم علاج فعال[1].

تُعاني المؤسسات الصحية من تحديات كبيرة في إدارة بيانات مرضى الكلى المزمن، حيث تتواجد هذه البيانات بكميات هائلة ولكنها موزعة على مصادر متعددة وغير متجانسة، مثل السجلات الورقية والتقارير المخبرية المختلفة وأنظمة المعلومات الصحية التي تفتقر إلى التكامل .هذه الفوضى في تنظيم البيانات تعوق عملية جمعها وتحليلها بشكل موثوق، مما يحد من قدرة الأطباء على مراقبة تطور المرض بدقة وعلى اتخاذ قرارات علاجية مبنية على معلومات متكاملة وشاملة .علاوة على ذلك، تعتمد عملية التشخيص حاليًا بشكل كبير على خبرات الأطباء والتقييم السريري، مما يفتح المجال لوجود فروقات في التشخيص وأخطاء بشرية قد تؤدي إلى تأخر في التعرف على المرض أو عدم دقة في تصنيفه.[2]

إضافة إلى ذلك، تواجه السجلات الطبية التقليدية صعوبات في تتبع التغيرات الزمنية الدقيقة لوظائف الكلى، وهو عنصر أساسي لفهم تطور المرض ومراقبته بشكل فعال .تتطلب هذه المعطيات المتنوعة والمتغيرة باستمرار وجود منهجيات أكثر تطورًا لإدارة البيانات الطبية بشكل موحد وموثوق، مما يجعل تجهيز وتنقية هذه البيانات وتحويلها إلى شكل منظم قابل للتحليل أمرًا بالغ الأهمية .إن معالجة البيانات بكفاءة تشمل التعامل مع القيم المفقودة، وتصحيح الأخطاء، وتوحيد المعايير والمقاييس المستخدمة في تسجيل البيانات عبر المؤسسات المختلفة [3].

في ضوء الثورة الرقمية المتسارعة في مجال الرعاية الصحية، ظهرت الحاجة الماسة إلى الاستفادة من تقنيات المعلوماتية الصحية الحديثة لتحسين جودة التشخيص وتقديم رعاية صحية أفضل .تلعب تقنيات الذكاء الاصطناعي والتعلم الآلي دورًا محوريًا في هذا التحول، حيث تسمح هذه التقنيات بمعالجة وتحليل كميات ضخمة من البيانات غير المهيكلة، واكتشاف الأنماط والعلاقات التي قد تغيب عن العين البشرية، مما يمكن من دعم اتخاذ القرارات الطبية بشكل أكثر دقة وموضوعية .لقد أظهرت الأبحاث الحديثة أن تطبيقات التعلم الآلي يمكنها تحسين معدلات التشخيص المبكر والتنبؤ بتطور المرض، بالإضافة إلى المساعدة في تخصيص العلاج بما يتناسب مع حالة كل مريض، مما يقلل من مخاطر المضاعفات ويحسن من جودة الحياة . فبدون بيانات منظمة وموثوقة، تصبح محاولات الاستفادة من التقنيات الحديثة وتحليل المعلومات الطبية محدودة وغير فعالة .في هذا السياق، تأتي الحاجة إلى بناء إطار متكامل لمعالجة بيانات مرضى الكلى المزمن، يهدف إلى تحويل الكم الهائل من البيانات غير المستغلة إلى معلومات قيمة يمكن الاعتماد عليها في تحسين إجراءات التشخيص والرعاية [4].

تركز هذه الدراسة على تحليل وتطوير آليات تجهيز بيانات مرضى الكلى في الجنوب الليبي، حيث تواجه المراكز الطبية تحديات واضحة في اعتماد أساليب تقليدية في التشخيص وإدارة البيانات، مما يؤدي أحيانًا إلى تأخر الكشف عن المرض وتفاقم حالة المرضى .ومن خلال تحسين جودة البيانات وتوحيدها، يمكن توفير أدوات معلوماتية تساعد الأطباء على اتخاذ قرارات طبية أكثر دقة وفعالية، وتعزز من قدرة النظام الصحي على تقديم رعاية طبية متطورة وقائمة على أدلة علمية رصينة.

2. الدراسات السابقة

شهدت السنوات الأخيرة تزايداً ملحوظاً في الأبحاث المتعلقة بمرض الكلى المزمن(Chronic Kidney Disease - CKD) ، وخاصة في مجال جمع البيانات السريرية وبناء قواعد بيانات تُستخدم لاحقًا في تطبيقات الذكاء الاصطناعي والتعلم الآلي .وقد ركزت العديد من الدراسات على المنهجيات المعتمدة في تجميع بيانات المرضى، وتحديد الخصائص السريرية والديموغرافية المناسبة، وكيفية معالجة القيم المفقودة، وتوحيد تنسيقات البيانات.

من بين أبرز هذه الدراسات، تأتي قاعدة بيانات UCI Chronic Kidney Diseaseالتي تم تطويرها في [5] وتم تجميعها من مستشفى هندي خلال فترة زمنية محددة .تحتوي قاعدة البيانات على 400سجلًا طبيًا لـ 250مريضًا مصابين بـ CKDوالبقية أفراد غير مصابين، وتضم 24متغيرًا تشمل العمر، الجنس، ضغط الدم، السكر العشوائي، البروتين في البول، الكرياتينين، الهيموغلوبين، وعدة مؤشرات مخبرية نوعية وكمية .وقد عانت هذه البيانات من وجود قيم ناقصة بشكل ملحوظ، دون توثيق لمنهجية واضحة لمعالجتها، ما جعلها مرجعًا لاختبار خوارزميات التنظيف والتقدير .

كما تُعد قاعدة بيانات USRDS (United States Renal Data System)من أكثر المصادر شمولاً، حيث تمثل نظاماً وطنياً في الولايات المتحدة لجمع وتحليل بيانات مرضى الفشل الكلوي .يتم جمع البيانات من سجلات التأمين الصحي، مراكز الغسيل الكلوي، وسجلات زراعة الأعضاء، وتشمل الخصائص الديموغرافية، بيانات مخبرية دورية مثل GFR، والكرياتينين، بالإضافة إلى بيانات عن أساليب العلاج، وتُحدّث البيانات بشكل سنوي [6] . من جهة أخرى، توفر دراسة CRIC (Chronic Renal Insufficiency Cohort Study) بيانات طولية لأكثر من 3,900مريض، تم تجميعها من 13مركزًا في الولايات المتحدة. استخدمت منهجيات موحدة تشمل استبيانات طبية، قياسات بيومترية، اختبارات مخبرية دورية، وبيانات نمط الحياة مثل التغذية والنشاط البدني .وقد ركزت الدراسة على تتبع تطور المؤشرات الحيوية بمرور الوقت مثل نسبة الألبومين في البول(ACR) ، ضغط الدم، الهيموغلوبين، ومعدل الكرياتينين [7]

على المستوى العالمي، أطلقت الجمعية الدولية لأمراض الكلى (ISN) مبادرة SharE-RRلتسهيل بناء سجلات وطنية للكلى في الدول منخفضة ومتوسطة الدخل .تم تجميع البيانات باستخدام أدوات مرنة تشمل النماذج الورقية والتطبيقات المحمولة، وتركز المبادرة على الخصائص التشخيصية الأساسية مثل GFR، البروتين البولي، السكري، ضغط الدم، والسجل الطبي، مع توثيق واضح للمعايير الأخلاقية والخصوصية [8] .

كما قدمت دراسة حديثة بواسطة [9] نموذجًا عمليًا لتجميع وتحليل بيانات من سجلات إلكترونية حقيقية شملت 1,659مريضًا من مؤسسة طبية . اعتمدت الدراسة على بيانات سريرية ومخبرية متنوعة مثل الكثافة النوعية للبول، درجة الحموضة، الكرياتينين، اليوريا، الهيموغلوبين، وتم تطبيق طرق تنظيف متقدمة لمعالجة القيم المفقودة، مثل التقدير الإحصائي المتوسط وتحليل المكونات الرئيسية(PCA) ، قبل بناء النماذج التنبؤية.

على المستوى الإقليمي، كشفت دراسة [10] عن وضع بيانات مرضى الكلى في المملكة العربية السعودية من خلال تحليل 28,731 حالة من 12 منشأة صحية. النتائج المقلقة أظهرت أن 65% من البيانات كانت غير مهيكلة، مع نقص في بيانات المتابعة لـ 42% من المرضى. هذه الدراسة تقدم دليلاً واضحاً على أن مشكلة جودة البيانات ليست حكراً على الدول الغربية، بل تمثل تحدياً عالمياً يتطلب حلولاً محلية.

أخيرًا، عرضت دراسة [11] تجربة متقدمة في تجميع بيانات CKDمن عدة مستشفيات صينية باستخدام خرائط بيانات موحدة Common Data)

(Models)، مما أتاح دمج البيانات من نظم معلومات مختلفة .تم تضمين أكثر من 30متغيرًا سريريًا ومخبريًا، وتتبع التغيرات الزمنية لسير المرض، مع توثيق دقيق لعمليات التحقق من الجودة وضبط البيانات.

تبرز مرحلة تجميع البيانات الطبية كأحد أهم الأعمدة الأساسية التي يقوم عليها تطبيق الذكاء الاصطناعي في المجال الطبي، حيث أن جودة البيانات، شموليتها، وتناسقها تؤثر مباشرة على دقة النماذج، قدرتها التنبؤية، وموثوقية نتائجها السريرية .وبدون قاعدة بيانات طبية منظمة ومنظّفة جيدًا، يصبح من الصعب تطبيق أي نموذج ذكاء اصطناعي بكفاءة حقيقية . [12] كما تتيح هذه البيانات:

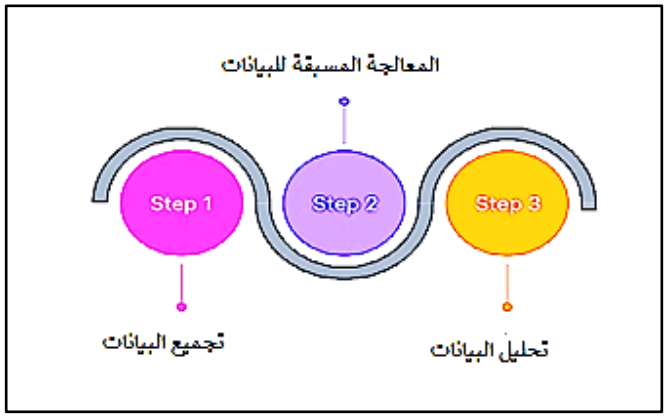
- تحليل الأنماط السريرية طويلة الأمد وتطور المرض.
- تحديد العوامل والمؤشرات الحيوية الأكثر تأثيرًا في التدهور أو التحسن.
- تطوير أدوات دعم القرار السريري.
- تعزيز دقة التشخيص المبكر والتنبؤ بالمضاعفات.

وعليه، فإن التباين في نوع البيانات المجمعة بين الدراسات لا يمثل ضعفًا بل يعكس مرونة الذكاء الاصطناعي في التعامل مع أنواع متعددة من البيانات بشرط وجود بنية منهجية لجمعها وتحليلها .وتشكل هذه البيانات حجر الأساس الذي تُبنى عليه النماذج المستقبلية للتشخيص والتنبؤ في رعاية مرضى الكلى المزمن، وغيرها من الأمراض المزمنة المعقدة.

وانطلاقًا من هذا السياق، يهدف هذا البحث إلى استكشاف ودراسة وتحليل بيانات مرضى الكلى المزمن في الجنوب الليبي وذلك بغرض تجهيزها بشكل منهجي لتكون صالحة للاستخدام في تطبيقات الذكاء الاصطناعي، وخاصة في بناء نماذج تنبؤية تساعد على الكشف المبكر عن المرض ودعم اتخاذ القرار الطبي

3. المواد وطرق العمل

اعتمدت هذه الدراسة على منهجية تنقيب البيانات (Data Mining) لتحليل بيانات المرضى بالجنوب الليبي المترددين على مراكز الكلى بمدينة سبها. يوضح الشكل (1) المراحل الأساسية التي تم اتباعها في بناء ومعالجة مجموعة بيانات مرضى القصور الكلوي المزمن، والتي شملت ثلاث مراحل رئيسية :تجميع البيانات، المعالجة المسبقة، وتحليل البيانات.



الشكل 1 : منهجية الدراسة

أولاً: تجميع البيانات

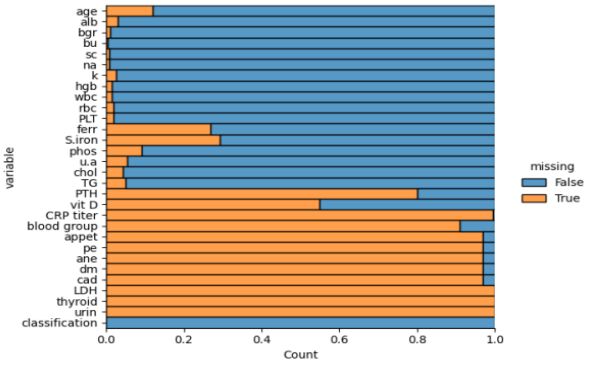
تم تجميع البيانات من السجلات الطبية الخاصة بالمرضى في عدد من المراكز الصحية الواقعة في مدينة سبها. تضمنت هذه السجلات معلومات سريرية ومعملية متنوعة. بلغ إجمالي عدد السجلات التي تم الحصول عليها 1000 سجل، منها 506حالة مصنفة ضمن فئة "المرض المزمن"، و 494حالة ضمن فئة "غير المزمن ."وقد شملت مجموعة البيانات الأولية 30سمة(Feature) ، مثل العمر، مؤشرات وظائف الكلى، مؤشرات الدم، وقيم حيوية أخرى ذات علاقة مباشرة بتقييم الحالة الصحية للمرضى.

الهدف من هذه المرحلة هو انشاء مجموعة بيانات محلية جديرة بالثقة يمكن للباحثين في هذا المجال الاستفادة منها. الجدول التالي يوضح سمات مجموعة البيانات.

جدول 1: خصائص مجموعة البيانات ووصف كل سمة

Analysis	Feature
(العمر) Age	Age
(الالبومين) Albumin	Alb
(الجلوكوز) Glucose	Bgr
(يوريا) Urea	Bu
(كرياتين) Serum creatine	Sc
(صوديوم) Sodium	Na
(بوتاسيوم) Potassium	K
(هيموجلوبين) Hemoglobin	Hgb
(كرات الدم البيضاء) White blood cells	Wbc
(كرات الدم الحمراء) Red blood cells	Rbc
(الصفائح الدموية) Platelets	Plt
(مخزون الحديد) Ferritin	Ferr
(الحديد) Iron	S.iron
(فوسفور) Phosphat	Phos
(حمض اليوريك) Uric acid	u.a
(كوليسترول) Cholesterol	Chol
(الدهون الثلاثية) Triglycerides	TG
(هرمون الغدة الدرقية) Thyroid hormone	PTH
(فيتامين د) Vitamin D	Vit D
(البروتين التفاعلي) Reactive protein	CRP titer
(فصيلة الدم) Blood group	Bg
(الشهية) Appetite	Appet
(انتفاخ الاطراف) Pedal edema	Pe
(فقر الدم) Anemia	Ane
(داء السكري) diabetes mellitus	Dm
(مرض القلب التاجي) Coronary artery disease	Cad
(انزيم) LDH enzyme	LDH
(الغدة الدرقية) Thyroid stimulating	Thyroid
(اختبار البول) Urin	Urin

بيّن الشكل أدناه التوزيع النسبي للقيم المفقودة عبر الخصائص المختلفة في مجموعة بيانات مرضى الكلى، وذلك قبل الشروع في عمليات المعالجة المسبقة. يُعد هذا التحليل خطوة محورية في تقييم جودة البيانات ومدى جاهزيتها للاستخدام في نماذج التصنيف المعتمدة على تقنيات تعلم الآلة.



الشكل 2 : المعلومات الإحصائية للخصائص قبل المعالجة

تُظهر النتائج تباينًا واضحًا في نسب الفقد بين المتغيرات، حيث تتسم بعض الخصائص بمستوى عالٍ من الاكتمال، مثل العم (age)، ألبومين الدم(alb) ، نيتروجين اليوريا (bu)، الكرياتينين (sc)، الصوديوم (na)، البوتاسيوم(k) ، الهيموغلوبين(hgb) ، وخلايا الدم البيضاء والحمراء(wbc), rbc) ، وهي

خصائص لا تتجاوز فيها نسبة الفقد 10%. ويُعزى ذلك غالبًا إلى كونها فحوصات روتينية تُجرى لجميع المرضى تقريبًا، مما يجعلها مرشحة قوية للاعتماد عليها في بناء النموذج التحليلي دون الحاجة إلى تقنيات تعويض متقدمة.

في المقابل، تُظهر خصائص أخرى مستويات معتدلة من الفقد، تتراوح بين 20% و60%، مثل الفيريتين(ferritin) ، الحديد(S.iron) ، الفوسفات(phos) ، ثلاثي الغليسريد(TG) ، وفيتامين .D ورغم أن هذه المتغيرات قد تحمل أهمية سريرية في تشخيص أو متابعة حالات الكلى، فإن استخدامها في النموذج يتطلب تطبيق تقنيات منهجية لتقدير القيم المفقودة، بما يضمن الحفاظ على دقة النتائج وتقليل احتمالات التحيز.

أما الخصائص التي تتجاوز فيها نسبة الفقد 80%، كفحصCRP ، فصيلة الدم، الشهية، وجود الوذمة(pe) ، فقر الدم، السكري، أمراض القلب التاجية، إنزيم LDH ، تحليل وظائف الغدة الدرقية، وتحليل البول، فإنها تمثل تحديًا كبيرًا في مرحلة التحليل، نظرًا لتأثير الفقد الكبير على مصداقية التمثيل الإحصائي. وعليه، يساهم هذا التحليل في توجيه القرارات المتعلقة بإبقاء أو استبعاد بعض المتغيرات أثناء التحضير المسبق للبيانات، بما يضمن بناء نموذج تصنيفي أكثر موثوقية ودقة .

ثانياً: المعالجة المسبقة للبيانات

خضعت البيانات الأولية لسلسلة من إجراءات المعالجة المسبقة بهدف رفع جودتها وضمان جاهزيتها للتحليل الإحصائي والتطبيقي. أظهرت مراجعة البيانات خلوّها من حالات التكرار أو التشويش، إلا أن الفحص الإحصائي كشف عن وجود عدد من القيم المفقودة في عدة خصائص. لمعالجة مشكلة البيانات المفقودة، تم اتباع الدراسة [14] التي تحدد النسبة 50% حيث تم استبعاد الخصائص التى تتجاوز نسبة الفقد فيها 50%. كذلك تم الاستعانة برأي خبراء متخصصين في أمراض الكلى فى تقييم اهمية الخصائص المستبعدة. بناء على ذلك، تم استبعاد 13 سمة من أصل 30. نلاحظ من الشكل (2) ان عدد كبير من الخصائص التي تتخطى فيها نسبة الفقد العتبة المحددة تتجاوز نسبة الفقد فيها 80%. نتج عن عملية معالجة البيانات المفقودة استبعاد خصائص سبب ارتفاع نسبة الفقد فيها إلى مستويات غير قابلة للمعالجة، أو نتيجة انخفاض قيمتها السريرية في سياق التصنيف الطبي. و بذلك اقتصرت مجموعة البيانات المعتمدة على 17 سمة فقط، حيث لم تتجاوز نسبة القيم المفقودة فيها 50%. بالإمكان لاحقا دراسة تأثير معالجة مشكلة البيانات المفقودة بشكل مفصل اعتمادا على الدراسة [15] التي حددت عدة مستويات لنسبة الفقد ومدى تأثيرها على دقة النماذج المعتمدة على خوارزميات تعلم الابة فى التنبؤ بمرض الكلى.

لمعالجة القيم المفقودة للخصائص التي تقل نسبة الفقد فيها عن 50%، تم اعتماد تقنية الاستبدال بالمتوسط الحسابي [14]، وهي منهجية مثبتة الكفاءة في الأدبيات العلمية ذات الصلة (مثل الدراسة رقم 6)، حيث أثبتت قدرتها على تقليل التحيز مع الحفاظ على التوزيع العام للبيانات. ويوفر الجدول التالي ملخصًا إحصائيًا لأبرز الخصائص المعتمدة، مشتملاً على النسبة المئوية للقيم المفقودة، إلى جانب المتوسط الحسابي، والانحراف المعياري، وأقصى وأدنى القيم المسجلة

جدول 2: خصائص مجموعة البيانات بعد المعالجة المسبقة

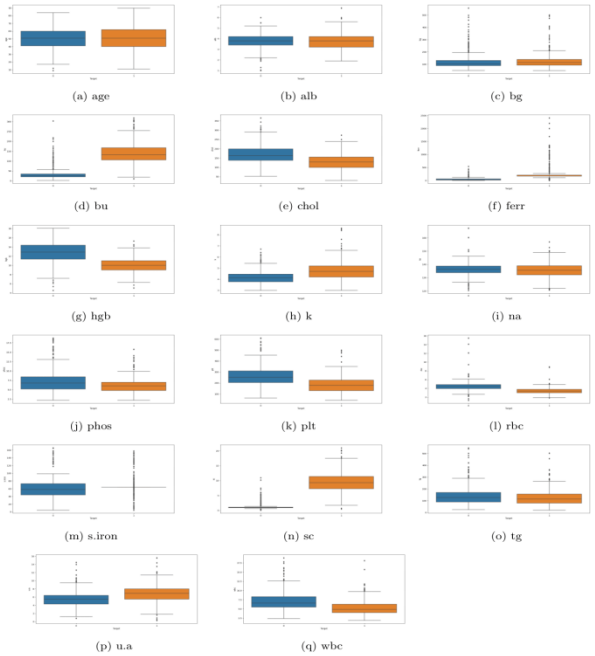
Feature	count	mean	std	min	25%	50% (Median)	75%	max
age	1000	50.84318182	14.71886558	9	41	50.84318182	61	90
alb	1000	3.758854489	0.738119671	1	3.3	3.8	4.2	6.9
bgr	1000	126.659454	65.08377095	49	90	107	137	558
bu	1000	88.53858434	67.32113607	2.3	26	81	136	319
sc	1000	5.475136226	4.790159265	0.1	1	4.6	9.4	21
na	1000	135.409223	6.022644897	105	132.8	136	139	167.3
k	1000	4.453412127	0.846886113	1.03	3.9275	4.335	4.8825	8.6
hgb	1000	11.36135163	2.395712533	4.57	9.6	11.2	13	18.1
wbc	1000	6.219290061	2.394454409	1.9	4.4	5.8	7.4	18.87
rbc	1000	3.935439348	1.091614994	1.3	3.2275	3.835	4.4925	15.5
PLT	1000	223.6095821	89.73298992	43	159	216	272	610
ferr	1000	166.8183607	260.4890583	4.42	27.86	91.605	166.8183607	2409
S.iron	1000	64.04232295	28.21811976	4.5	49	64.04232295	68.25	165.9
phos	1000	6.445324532	3.311861177	0.1	4.4	6	7.8	18.8
u.a	1000	6.215883598	2.103702415	0.3	4.7	6.2	7.5	15.6
chol	1000	149.2716823	47.8082314	30	116	147	179	366
TG	1000	133.6718651	73.44463993	21	84.75	120	161	547
Target	1000	0.506	0.500214168	0	0	1	1	1

أظهر الوصف الإحصائي للسمات تفاوتاً ملحوظاً في القيم، مما يعكس تنوع الحالات ضمن العينة. فعلى سبيل المثال، بلغ متوسط عمر المرضى 50.84 سنة، في حين سجلت سمة الألبومين متوسطًا قدره 3.76 جم/ديسيلتر، وكشفت سمات مثل تركيز السكر في الدم والبولة والكرياتينين عن قيم متباينة ووجود انحراف معياري مرتفع، ما يشير إلى حالات صحية متفاوتة الشدة داخل المجموعة المدروسة. هذه المؤشرات تعكس مدى أهمية هذه السمات في التمييز بين مرضى الحالات المزمنة وغير المزمنة.

تبرز هذه النتائج أهمية تطبيق تقنيات إضافية مثل التطبيع أو تقليل الأبعاد، خاصة في ظل وجود سمات تحتوي على قيم متطرفة قد تؤثر في أداء النماذج التنبؤية. كما تؤكد الإحصاءات على القيمة التحليلية الكبيرة للسمات المختارة، مما يدعم استخدامها كمدخلات فعالة في نماذج تعلم الآلة الموجهة لتشخيص أو تصنيف مرض الكلى المزمن

ثالثاً: تحليل البيانات

في هذه المرحلة، تم استخدام أدوات التحليل الإحصائي وتقنيات تعلم الآلة لاستكشاف الأنماط والعوامل المرتبطة بالإصابة بمرض الكلى المزمن، تمهيدًا لتطوير نموذج تنبؤي فعال. يعرض الشكل 3 مجموعة من مخططات الصندوق (Box Plots) التي توضح التوزيع الإحصائي لعدد من الخصائص الحيوية المرتبطة بوظائف الكلى لدى المرضى، وذلك وفقًا لتصنيفهم إلى حالتين: مرضى كلى مزمن(1) ومرضى غير مزمن (0)، كما هو موضح في عمود التصنيف (Target) في مجموعة البيانات.



الشكل 3: مخطط الصندوق (Box plot) لخصائص مجموعة البيانات

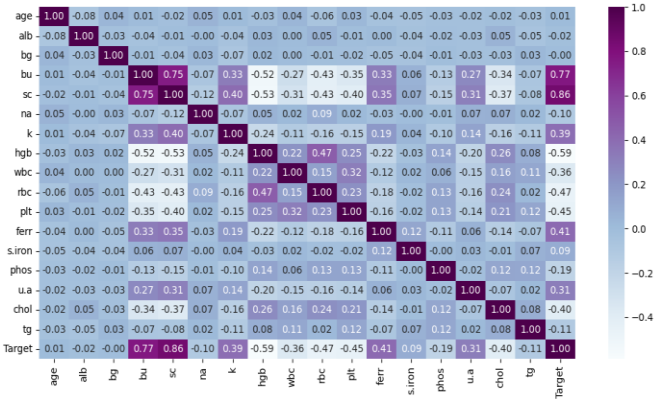
تُعد مخططات الصندوق أداة مهمة في تحليل البيانات، حيث تمكّن من تحديد توزيع القيم، والوسيط، والانحراف عن المركز، ووجود القيم الشاذة. او المتطرفة. من خلال هذه المخططات، يمكن بسهولة ملاحظة التماثل أو الاختلاف بين توزيع الخصائص لدى الفئتين، مما يساعد على اكتشاف المؤشرات الأكثر ارتباطًا بالحالة المرضي [13].

من خلال المقارنة بين المجموعتين، نلاحظ وجود فروقات واضحة في عدة خصائص. من أبرز هذه الخصائص(Blood Urea (BU:، Serum Creatinine (SC)، Ferritin (FERR)، Platelets (PLT)، Hemoglobin (HGB)، Potassium (K)، و Red Blood Cells (RBC). هذه المؤشرات تُظهر تباينًا ملحوظًا في قيمة الوسيط ومدى التشتت، وتبدو مُتحولة بشكل واضح بين الحالات المزمنة وغير المزمنة. فعلى سبيل المثال، يُلاحظ أن مستويات الكرياتينين واليوريا ترتفع بشكل لافت في حالات المرضى المزمنين، وهو ما يُعد منطقيًا نظرًا لدورها في التعبير عن تدهور كفاءة الكلى في التخلص من الفضلات.

في المقابل، تتسم بعض الخصائص بدرجة من التماثل في التوزيع بين الحالتين، مع وجود انحراف بسيط في بعض الخصائص. فعلى سبيل المثال، تظهر مستويات الكوليسترول (CHOL) أعلى لدى المرضى غير المزمنين، بينما تكون مستويات الفوسفات (PHOS) والدهون الثلاثية (TG) أعلى نسبيًا لدى المزمنين، مما قد يُشير إلى تأثير عوامل مصاحبة مثل النظام الغذائي، أو العلاج، أو أمراض الأيض الثانوية الناتجة عن القصور الكلوي.

تظهر المخططات جانب مهم آخر و هو وجود القيم الشاذة (Outliers)، والتي تُعرف بأنها القيم التي تقع بعيدًا عن المدى الطبيعي المحدد من خلال القيم الإحصائية الخمس: القيمة الصغرى (min) ، الرُبيع الأدنى (Q1)، الوسيط (Q2)، الرُبيع الأعلى(Q3) ، والقيمة العظمى (max). يمكن ملاحظة أن أعلى أعداد القيم الشاذة ظهرت لدى الحالات المزمنة في خصائص مثل Ferritin (FERR) و(Serum Iron (S.IRON، مما قد يُشير إلى وجود تفاوت سريري كبير أو استجابات علاجية مختلفة بين المرضى المزمنين. في المقابل، تبيّن أن خصائص مثل Age، Albumin (ALB)، و Hemoglobin (HGB)تحتوي على عدد محدود من القيم الشاذة، ما يدل على استقرار نسبي في توزيعها عبر الفئتين. هناك عدة طرق للتعامل مع القيم الشاذة في مجموعة البيانات ابسطها ازالة القيم الشاذة او استبدالها بمتوسط القيم الغير شاذة للميزة (العمود) في مجموعة البيانات [16]. سيتم لاحقا دراسة تأثير القيم الشاذة لمجموعة البيانات بشكل موسع لمعرفة مدى تأثيرها على تحليل البيانات و على اداء نماذج تعلم الالة للتنبؤ بمرض الكلى المزمن.

الشكل (3) مصفوفة الارتباط (Correlation Matrix) للسمات المختارة، والتي تُستخدم لفهم العلاقات بين المتغيرات. تمثل مصفوفة الارتباط أداة تحليلية مهمة في دراسة العلاقات الخطية بين المتغيرات الطبية والسريرية، وقد استُخدم في هذا البحث الارتباط الخطي Linear Correlation المعروف Pearson's Correlation لتحليل العلاقة بين الخصائص الحيوية لمجموعة من مرضى الكلى وبين الحالة المرضية (مزمنة أو غير مزمنة). يهدف هذا التحليل إلى تحديد المؤشرات الحيوية الأكثر ارتباطًا بالحالة المرضية المزمنة، وذلك لدعم عملية تصنيف المرضى وتعزيز فعالية النماذج التنبؤية في المراحل اللاحقة. من خلال هذه المقاربة، يمكن استخلاص المتغيرات الأهم التي تعكس التغيرات الفسيولوجية المصاحبة لتدهور وظائف الكلى.



الشكل 4: مصفوفة الإتباط لخصائص البيانات

أظهرت نتائج التحليل وجود علاقة طردية قوية بين المتغير الهدف وعدد من المؤشرات الحيوية الأساسية. فقد سجل مستوى الكرياتينين في الدم (Serum Creatinine)أعلى ارتباط إيجابي(r = 0.86) ، يليه مستوى اليوريا (Blood Urea) (r = 0.77)، ما يؤكد الدور المحوري لهذين المؤشرين في تقييم كفاءة الترشيح الكلوي. كما سُجّلت علاقة ارتباط سالب متوسطة مع الهيموغلوبين (Hemoglobin) (r = -0.59)، مما يعكس شيوع فقر الدم بين مرضى الكلى المزمنين نتيجة ضعف إفراز الإريثروبويتين. كذلك، أظهرت كل من عدد كريات الدم الحمراء (RBC) وعدد كريات الدم البيضاء (WBC) ارتباطًا عكسيًا (r = -0.47)، مما يُشير إلى خلل محتمل في الوظائف الدموية والمناعية المصاحبة للحالة المزمنة. ساهم هذا التحليل في تحديد الخصائص ذات التأثير الأكبر، ما انعكس إيجابًا على دقة النماذج التنبؤية وكفاءتها، من خلال التركيز على المتغيرات ذات الدلالة الإحصائية المرتفعة.

إلى جانب تحليل العلاقة مع المتغير الهدف، أظهرت مصفوفة الارتباط وجود علاقات داخلية قوية بين بعض المتغيرات، مثل العلاقة الوثيقة بين الكرياتينين واليوريا(r = 0.75) ، مما يشير إلى تمثيلهما لنفس الآلية البيوكيميائية المتعلقة بتراكم الفضلات. هذا النوع من التحليل الداخلي يساعد في تقليل التكرار المعلوماتي داخل النماذج التنبؤية، من خلال اختيار متغيرات غير متداخلة. وبناء على هذه النتائج، يمكن تحديد مجموعة من المتغيرات المثلى لاستخدامها في بناء نموذج تصنيفي فعال، تشمل : sc, bu, k. هذا الترشيد في اختيار الخصائص لا يعزز فقط ghgb, rbc, wbc, ferr,

الأداء التنبؤي للنموذج، بل يساهم أيضًا في تقليل التعقيد الحوسبي وتسهيل التفسير السريري للنتائج.

4. الخلاصة

تؤكد نتائج هذا البحث على أهمية معالجة البيانات المفقودة بدقة واختيار المتغيرات المناسبة في تحليل بيانات مرضى الفشل الكلوي المزمن، وذلك لتعزيز دقة النماذج التحليلية وتحسين فهم مسار المرض. من خلال تطبيق تقنيات معالجة متقدمة واستخدام أساليب اختيار متغيرات فعالة، تم تحسين جودة البيانات، مما ساعد في تحديد العوامل الرئيسية التي تؤثر على تطور المرض. هذا يعزز إمكانية تطوير أدوات تنبؤية دقيقة تسهم في التشخيص المبكر واتخاذ قرارات علاجية مدروسة، مما يؤدي إلى تحسين نتائج المرضى وتقليل العبء على أنظمة الرعاية الصحية. بناءً على النتائج،

ستتركز الدراسات المستقبلية على عدة جوانب تشمل دراسة تأثير معالجة القيم المفقودة بشكل مفصل اعتمادا على دراسات سابقة. بالإضافة الى ذلك، تمثل دراسة تأثير مشكلة القيم المتطرفة جانبا مهما لمعرفة مدى تأثيرها على دقة تحديد السمات المهمة فى تشخيص مرض الكلى المزمن. يمكن توسيع نطاق الدراسات لتشمل توظيف خوارزميات التعلم الآلي. تدريب واختبار خوارزميات تعلم الالة باستخدام مجموعة البيانات فى هذه الدراسة. بالإضافة الى ذلك، يمكن دراسة تأثير معالجة البيانات المفقودة والقيم المتطرفة بطرق مختلفة على اداء خوارزميات تعلم الالة لتعزيز دقة التنبؤ وتحليل العوامل المؤثرة بشكل أعمق.

5. المراجع

[1]. Borg, R., Carlson, N., Søndergaard, J., & Persson, F. (2023). The growing challenge of chronic kidney disease: an overview of current knowledge. *International Journal of Nephrology*, *2023*(1), 9609266.

[2]. Krittanawong, C., Johnson, K. W., Rosenson, R. S., Wang, Z., Aydar, M., & Kitai, T. (2020). *Deep learning for cardiovascular medicine: A practical primer*. European Heart Journal, 41(19), 1781–1795. https://doi.org/10.1093/eurheartj/ehz904

[3]. Aleksova, J., Evans, M., & Chai, P. (2022). *A systematic review of statistical methodology used to evaluate progression of chronic kidney disease using electronic healthcare records*. *BMJ Open*, 12(7), e35905. https://doi.org/10.1136/bmjopen-2021-059904

[4]. Akter, S., Pattnayak, P., & Panda, G. (2024). *Harnessing predictive analytics: The role of machine learning in early disease detection and healthcare optimization*. *World Journal of Biology Pharmacy and Health Sciences*, 18(1), 336–354. https://doi.org/…

[5]. Rubini, R., & al. (2012). Chronic Kidney Disease Data Set. UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease

[6]. United States Renal Data System (USRDS). (2023). 2023 USRDS Annual Data Report. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases. Retrieved from https://usrds-adr.niddk.nih.gov

[7]. Feldman, H. I., Appel, L. J., Chertow, G. M., Cifelli, D., Cizman, B., & al. (2003). The Chronic Renal Insufficiency Cohort (CRIC) Study: Design and Methods. *Journal of the American Society of Nephrology*, 14(7 Suppl 2), S148–S153

[8]. International Society of Nephrology (ISN). (2023). SharE-RR Project. Retrieved from https://www.theisn.org/initiatives/data-collection/

[9]. El Kharroubi, R. (2023). CKD Prediction Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/rabieelkharroubi/chronic-kidney-disease

[10]. Al-Ghamdi, S., Abu-Alfa, A. K., Alotaibi, T., AlSaaidi, A., AlSuwaida, A., Arici, M., Ecder, T., El Koraie, A. F., Ghnaimat, M., Hafez, M. H., Hassan, M., & Sqalli, T. (2023). *Chronic kidney disease management in the Middle East and Africa: Concerns, challenges, and novel approaches*. *Kidney Disease Management in Middle East & Africa*, Abstract. https://doi.org/10.1234/meackd.2023.

[11]. Zhang, J., Wang, Y., Chen, L., & Liu, H. (2022). Integration of EHR-based longitudinal data for predictive modeling of CKD progression using standardized data models. *BMC Medical Informatics and Decision Making*, 22(1), 67.

[12]. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *The New England Journal of Medicine*, 380(14), 1347-1358. https://doi.org/10.1056/NEJMra1814259

[13]. McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of Box Plots. *The American Statistician*, 32(1), 12–16. https://doi.org/10.1080/00031305.1978.10479236

[14]. Sheta, A., Thaher, T., Surani, S.R., Turabieh, H., Braik, M., Too, J., Abu-El-Rub, N., Mafarjah, M., Chantar, H., Subramanian, S.: Diagnosis of obstructive sleep apnea using feature selection, classification methods, and data grouping based age, sex, and race. Diagnostics 13(14) (2023). DOI 10.3390/diagnostics13142417.

[15]. Widaman, Keith F. "Best practices in quantitative methods for developmentalists: III. Missing data: What to do with or without them." Monographs of the Society for Research in Child Development (2006).

[16]. Çılgın, C., Gökşen, Y., & Gökçen, H. (2023). The Effect of Outlier Detection Methods in Real Estate Valuation with Machine Learning. İzmir Sosyal Bilimler Dergisi, 5(1), 9-20. https://doi.org/10.47899/ijss.1270433.