 https://doi.org/10.24191/jeesr.v26i1.003

Chronic Kidney Disease Diagnostic Tools Based on Machine Learning Algorithms: A Review

Muhamad Huzaimi Abdul Ghafar, Syed Abdul Mutalib Al-Junid, Megat Syahirul Amin Megat Ali, Fathimah Mohamad, and Abdul Hadi Abdul Razak*

***Abstract*—Chronic kidney disease (CKD) is a global health crisis, responsible for approximately 60% of worldwide deaths. With a projected increase in CKD patients on dialysis exceeding 2 million by 2030, there is an urgent need for improved diagnostic methods. Current procedures, such as laborious and time-consuming blood tests, fail to differentiate between drug-resistant phases of CKD. This paper aims to explore the potential of Artificial Intelligence (AI) tools, specifically machine learning (ML), in revolutionizing CKD diagnosis. This work intends to enlighten the evolution of ML techniques in CKD diagnosis and their contemporary applications. We conducted an extensive literature review, identifying 70 papers pertaining to ML-based CKD diagnostic tools recently published. These papers were thoroughly examined to categorize the diverse AI methods utilized in medical diagnostics, particularly those aimed at CKD detection. The review identified a range of AI methods used in CKD diagnosis, signifying substantial progress in this domain over the last decade. These methods show promise in addressing the challenges associated with early CKD detection. This paper highlights the evolving landscape of ML applications in CKD diagnosis and their current relevance. This paper concludes with a discussion of prospects for future research on AI-based CKD diagnostic systems, including deep learning algorithms applied to an assortment of open problems and challenges.**

***Index Terms*—Chronic Kidney Disease, Blood, Urine, Multiple Imputations, Machine Learning**

I. INTRODUCTION

Explorations of artificial intelligence (AI) techniques to manage and predict [1] the condition of chronic disease patients have been published by a lot of researchers since 2012. The development of an intelligent diagnosis model for a concerning disease worldwide has great potential to improve chronic disease care [2]. However, most of the research works at that moment focused on personalising health trend patents [3], smart monitoring [4], and disease control [5] for critical issues of health care such as heart disease [6] and diabetes [7].

The term AI refers to the ability of a computer, robot, or another machine to do any human-like capabilities and intelligence [8]. The advancement of these methods has been divided into different subs of techniques such as machine learning [9], neural networks [10], and deep learning [11]. AI has gone through evolution by developing a hybrid prediction model [12], which combines several machine learning techniques to improve accuracy and efficiency.

Realising the promise and necessity of AI for medical applications, especially critical sickness, has boosted the demand for intelligence prediction models. Implementing this intelligent approach does not only focus on chronic disease but is widely used on other high-risk diseases such as cancer [13], cardiovascular [14], and diabetes mellitus [15] at an early stage.

Developing an intelligent prediction model with the utilisation of machine learning helps doctors provide more patient intervention alternatives. Early identification of chronic disease allows doctors to provide successful treatment. Thus, it will prolong critical organs' function and reduce the risk of death [16].

Chronic kidney disease (CKD) is one of the most serious medical issues worldwide. According to World Health Organisation (WHO) records, the mortality rates of these chronic diseases have grown faster in recent years. CKD has become among the top causes of death with at least 2.4 million cases reported annually [17]. For that reason, prediction models using AI classifiers have been studied for a decade by many researchers.

Theoretically, CKD is defined as kidney damage or a glomerular filtration rate of less than 60ml/min/1.73m2 for more than three months [18]. Several causes of CKD include loss of blood flow to the kidney, urinary infection, high blood pressure, and diabetes [19]. Heart disease, dehydration, and liver failure cause blood flow loss. Prostate, colon, and cervical malignancies would cause CKD due to urination problems.

On the other hand, diabetes and high blood pressure contribute to two-thirds of CKD cases [20]. Excessive sugar consumption can lead to kidney damage, causing an increase in the kidney's filtration function to remove waste and excess fluid from the bloodstream [21]. One of the initial indicators of CKD is the presence of a protein called albumin in the urine. In a healthy kidney, albumin should not pass from the blood into the urine [22]. CKD is typically categorized into five stages, with the most severe being end-stage renal disease, which necessitates kidney replacement therapy [23].

The prevalence of CKD has been on the rise over the past

This manuscript is submitted on August 15, 2024, revised on September 24, 2024, accepted on October 15, 2024. Muhamad Huzaimi Abdul Ghafar, Syed Abdul Mutalib Al-Junid and Abdul Hadi Abdul Razak are from the Electronics Architecture and Applications (EArA), School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia.

Megat Syahirul Amin Megat Ali is from Microwave Research Institute (MRI), Universiti Teknologi MARA, Shah Alam, Malaysia.

Fathimah Mohamad is from Department of Pathology, Faculty of Medicine, Universiti Teknologi MARA, Cawangan Sungai Buloh, Malaysia.

*Corresponding author
Email address: hadi@ieee.org

decade, as highlighted in the 24th Report of the Malaysian Dialysis and Transplant Registry in 2016. This report indicated that 14 percent of women and 12 percent of men were affected by the disease [24,25]. Meanwhile, over 850 million people worldwide have CKD, acute kidney damage, or are receiving renal replacement treatment, according to the American Society of Nephrology. This amount increases the total number of people with diabetes by 100 percent. In addition, the association projected that CKD would rank among the top five global killers [26].

Therefore, this paper reviews CKD diagnostic tools based on machine learning (ML). The rest of this paper is organized as follows. The common ML terminology and artificial intelligence process in medical diagnosis are presented in this section. The next section discusses related works on ML-applied methods for CKD diagnostic systems. Then, challenges in CKD prediction using ML are presented. Next, the way forward in CKD prediction would be proposed whereas the final section concludes this paper.

II. Machine Learning in Disease Diagnosis

This section presents common ML terminology, the basic type of AI sub-field, and the medical diagnosis process using ML approaches.

2.1 Commonly Used ML Terminology

Artificial Neural Network (ANN) is the most popular ML technique used for intelligent diagnostics. There are eight components commonly applied in ANN terminology: neuron; input, output, and hidden nodes layer; weights; threshold value; training classifier; and learning parameters.

ANN consists of the main building block called neuron. It contains three major layers; input, hidden nodes and output for receiving information performing a complex mathematical operation and classifying the data points [27]. It is inspired by the human brain, where all the mathematical operation in this method is likely a brain that performs a particular task [28].

Therefore, to perform computing diagnostics similar to the human brain, data needs to be multiplied by the 'weights' in each layer. Value weights can be considered as the strength of the connection between two neurons [29]. The threshold value of neuron output can be determined as 0 or 1. If the class's data is more than two, the threshold value will be considered 1 until the nth data class. Thus, it is the parameter of neurons and must be determined as an integer.

The ANN terminology also includes training a neural network to reduce the error of intelligent diagnostic models [28]. The value of weight in the training phase is based on the number of inputs implemented in the model [30]. Another ANN terminology is learning to evaluate the changes in its input or output. All processes in the learning phase depend on the number of parameters.

Generally, a patient will meet a doctor and undergo a urinalysis, as well as a blood test. Patients will be diagnosed at critical stages – stages 3, 4 and 5, where at these stages, the kidney condition can no longer be saved. Therefore, it is very important to monitor health conditions twice a year to make sure that the kidney functions well.

2.2 Basic Types of ML

There are three types of neural networks for ML techniques which is single-layered feed-forward, multi-layered feed-forward, and recurrent neural network. Both single and multi-layered feed-forward neural network consists of input and output layers, however, it is not reversible.

The difference between these two neural networks lies in the hidden layer or hidden neurons. Multi-layered neural networks can consist of more than one hidden neuron that works by linking external input with other neurons in the network. For example, the output of the second layer becomes the input for the third layer, and it continues until the output layer.

2.3 Medical Diagnosis Process Using ML

The ML approach gives new abilities to perform diagnosis without using human judgment. This technique consists of seven major steps, which are data collection, data preparation, selection, training and model evaluation, parameter tuning, as well as predictions.

Data collection and preparation are the first two important processes for the overall performance of the prediction model. Each process aims to gather and clean raw data into useful data. Then, ML will select, train, and evaluate the prediction model.

Selection is key to proposing a task-appropriate algorithm. The model must then be trained to improve forecast accuracy. In the evaluation step, model output is compared to input parameters.

Finally, parameter tuning includes training iterations, performance, outcome, learning rate, beginning values, and distribution. The model must be evaluated with the actual dataset to make sure that the prediction outcome matches the domain expert. The overall steps of ML are shown in Figure 1.

III. Potential Machine Learning or Deep Learning Method for CKD

This section discusses the findings of several research groups on the proposed classifier for the CKD prediction model. This section also examines the issue of missing data and solutions proposed by researchers to increase the model's accuracy.

3.1 Existing Method for CKD Diagnostic

Several tests and experiments have been employed for CKD prediction, such as image processing from ultrasound, capturing disease trends using a signal, missing and imbalanced data, feature selection and a comparison of effectiveness between various classifiers. Although each of the studies obtains high accuracy for several classifiers, there is a major drawback to identifying CKD and non-CKD. In addition, this identification requires a great deal of time for result interpretation and a complicated method to execute. The primary CKD diagnostic tests and procedures are detailed in Table 1.

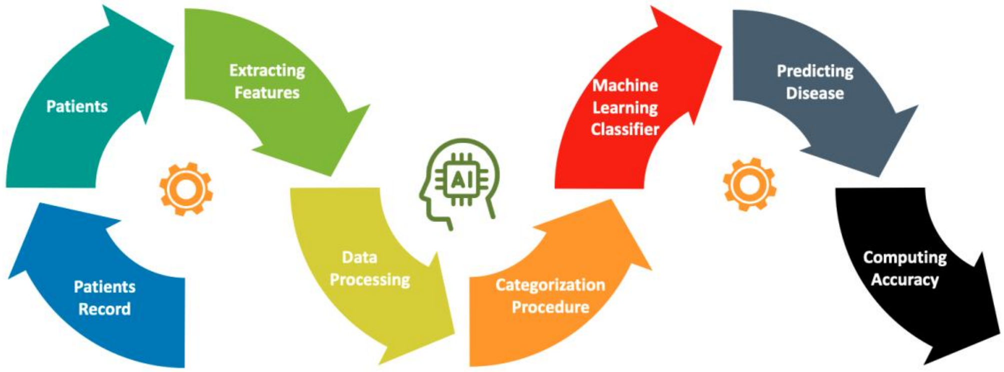

Fig. 1. Machine Learning System

Current prediction techniques, such as ultrasonography and signal waves, perform accurate diagnostic results and are non-invasive, but the cost of clinical tests is expensive. Each medical data needs to go through the pre-processing phase before training the model with various classifiers due to missing data and imbalanced classes. This has spawned a vast area of research for the creation of better, faster, and more effective diagnostic instruments and methodologies to achieve greater sensitivity and specificity, as well as to control the disease and minimise mortality rates. Consequently, the desire to identify new and improved ways has led to the development of ML.

TABLE 1. MAIN SHORTCOMINGS OF CKD DIAGNOSTIC AND PREDICTION TESTS.

Test	Methodology	Interpretations	Shortcomings	References
Kidney condition based on ultrasound image.	Imaging processing techniques and ML approach for the diagnosis of different CKD stages.	Diagnosis through non-invasive ultrasonographic imaging techniques.	Expensive	[11]
Diagnose CKD using four classifier model.	Applied Correlation Based attribute selection (CBAS) method and Fuzzy Rough Set Based attribute selection (FRSBAS) method.	Predict CKD using four ML classifiers k-Nearest Neighbour, naïve Bayes, Random Forest and Logistic Regression.	Need to apply different feature selections.	[31]
Handling missing data for CKD dataset.	Applied features selection and imputation method using K-neighbouring algorithm.	Perform CKD prediction using Random Forest and Decision Tree classifier.	Both classifiers perform 87% of accuracy.	[32]
Design CKD prediction model on imbalance dataset.	Proposed sampling method Synthetic Minority Oversampling (SMOTE) and Randomly Under Sampling (RUS) to solve the class imbalance.	Applied Linear Regression, SVM, Multi-Layer Perceptron and K-NN for the classifier.	Complexity and time-consuming.	[33]
Capture the long-term trends in the CKD data, while effectively handling the noise in the signal.	Proposed the Time-Aware Long Short-Term Memory (LSTM) for CKD prediction.	Extracted from two larger clinical datasets: DARTNet and MIMIC-III dataset.	Expensive and limited to stage 3 patients.	[34]
Feature selection of CKD attributes	Approach Ant Lion Optimization (ALO) technique to choose optimal features for the classification process.	Utilising DNN method as a classifier for the prediction model.	Low number of attributes in the prediction model.	[35]
Compare the efficiency of different classifiers.	Applied various ML algorithms to predict the CKD and analysed their efficiency.	Proposed KNN, Decision Tree, ANN and SVM as the classifier.	Non-specific classifier.	[36]

3.2 Prediction using Machine Learning Approach

Previous research in the field of CKD prediction models has primarily concentrated on detecting the disease and its different stages using various diagnostic approaches. CKD prediction often relies on two key clinical tests: urine tests and blood tests. Urinalysis involves three fundamental steps: visual inspection, chemical analysis, and microscopic examination. Albuminuria [37], specific gravity [38], protein, and glucose are among the urinalysis parameters. Several clinical blood tests for kidney diseases include fasting blood sugar, urea, serum creatinine, serum sodium, and potassium lab tests.

Both urinalysis and blood tests play a crucial role in assessing kidney function as all their parameters are interrelated and contribute to the determination of kidney health. For instance, the measurement of urea assesses the nitrogen content in the blood [35]. Elevated urea levels indicate that the kidneys are unable to efficiently eliminate urea from the bloodstream.

ML techniques enhance these tests by enabling the generalization of attributes across both individuals with normal kidney function and those with CKD [39]. Efforts have already been made to model clinical traits at different stages of the disease [40]. Well-established ML methods for this purpose encompass convolutional neural networks [41], support vector machines [42], decision tree classifiers [43], and artificial neural networks (ANN) [44].

3.3 CKD Prediction using ANN

Previous research in the field of CKD prediction models has primarily concentrated on detecting the disease and its different stages using various diagnostic approaches. CKD prediction often relies on two key clinical tests: urine tests and blood tests. Urinalysis involves three fundamental steps: visual inspection, chemical analysis, and microscopic examination. Albuminuria [37], specific gravity [38], protein, and glucose are among the urinalysis parameters. Several clinical blood tests for kidney diseases include fasting blood sugar, urea, serum creatinine, serum sodium, and potassium lab tests.

Both urinalysis and blood tests play a crucial role in assessing kidney function as all their parameters are interrelated and contribute to the determination of kidney health. For instance, the measurement of urea assesses the nitrogen content in the blood [35]. Elevated urea levels indicate that the kidneys are unable to efficiently eliminate urea from the bloodstream.

One of the most well-known ML approaches is ANN. This method learns from examples and mimics the biological workings of neurons in the brain. Despite being one of the first AI paradigms, it is still significant due to its capacity to generalise a problem's solution [45]. Modelling of physiological phenomena such as cognitive development [46], dengue prediction [47], and cardiovascular issues [48], has been frequently used with this approach. ANN has also been employed in CKD investigations in the past.

Despite the expanding quantity of studies on applying AI to predict CKD, the following gaps have been discovered: 1) Numerous testing methods have been investigated in the literature to attain specific qualities. However, there has been a lack of emphasis on prioritizing the selection of convenient and cost-effective attributes while simultaneously maintaining prediction accuracy and reliability; 2) In the realm of CKD studies, a significant portion of research addressing missing data and unbalanced datasets has predominantly relied on rudimentary strategies like attribute elimination and correlation-based processes to identify relevant parameters. In contrast, methods such as multiple imputations play a pivotal role in accurately generating credible data points; 3) Previous research on CKD employing Artificial Neural Networks (ANN) has encountered issues related to ineffective reporting of findings. This was primarily attributed to insufficient technical details regarding both data pre-processing and model development, thereby casting doubt on the credibility of the results. Figure 2 illustrates a flowchart used to determine the optimal number of hidden nodes.

In this paper, MATLAB is employed for modelling tasks. Generally, an Artificial Neural Network (ANN) comprises an input layer, multiple hidden layers, and an output layer [49]. Nevertheless, previous research [50] has revealed that a network with just one hidden layer can accurately approximate any function. The number of input nodes is determined by the number of network attributes, while this paper employs a single output node. The optimal number of hidden nodes is ascertained through a distinct set of optimization experiments, which take into account the network's unique training characteristics [51].

Subsequently, the error is utilized in the back-propagation weight-update process, incorporating appropriate learning techniques. Levenberg-Marquardt algorithm is used in this research. Iterative network training would continue until the error is reduced to a minimum [52]. The data is first randomised. Afterwards, 70 percent is utilised for training, 15 percent for validation, and 15 percent is used for testing. The validation set serves the purpose of preventing the network from overfitting. Occasionally, data from the validation set is incorporated during training to evaluate the network's ability to generalize. If the validation error surpasses a certain threshold, the training process is stopped, and the previous set of network weights is adopted for the final structure. Subsequently, the model's performance is assessed using the testing set [53].

Two distinct models are developed using Artificial Neural Networks (ANN): one incorporating urine-based attributes with a five-input structure, and the other incorporating blood-based attributes with a nine-input structure. The determination of the number of hidden nodes, however, follows a rule-of-thumb experiment. The lower limit is set to 2/3 of the combined size of the input and output layers, while the upper limit is chosen to be less than twice the size of the input layer [54]. The experimentation process begins by training the ANN for 40 iterations using the lower limit configuration and then averaging the results. This process is iterated until the maximum limit is reached, as depicted in Figure 2.

Table 2 shows the CKD prediction model's performance using several ML and hybrid classifiers. Random Forest and Neural Networks with a hybrid approach achieve a higher accuracy percentage. However, other classifiers such as Support Vector Machine (SVM), Deep Neural Networks (DNN), and ANN also produced excellent results with more than 80 pecent accuracy.

Findings from [31], [55], [56], and [57] used the same CKD dataset with different approaches to the AI classifier. In [31], the number of attributes was reduced using the correlation-based attribute selection (CBAS) method and fuzzy rule set-based attribute selection (FRSBAS). Both approaches then reduced the number of attributes to 13 and 15 features, respectively. However, the selected attributes were not listed in this paper.

Although [55] experimented with a basis for the ANN approach, the study did not indicate the total number of attributes being used. The research used a correlation-based feature subset (CFS) to reduce the number of attributes. Since there is no proper number of attributes and the input of the prediction model, the performance could not be comparable.

Meanwhile, findings from [56] proposed four different algorithms, which are PNN, Multilayer Perceptron (MLP), SVM, and Radial Basis Function (RBF). The accuracy of each algorithm has been compared. PNN produced the highest percentage of accuracy. All proposed prediction model performs up to 96 percent, however, the attribute for the study was doubted. One of the significant outputs did not tally with the original CKD dataset of UCI ML Respiratory.

Results from [57] established a hybrid modified Neural Network approach in CKD classification. Multilayer

Perceptron Feed-forward Networks (MLP-FFN) and neural networks based on Particle Swarm Optimization (PSO-NN) were used in this experiment. Although the accuracy is slightly high, the reports did not mention any feature selection and imputation process to reduce the number of attributes and missing data.

In other studies, for example [58], [59], [60], and [61], researchers used different types of datasets to develop a CKD prediction model. In [58], [59], and [60], CT scans and ultrasound images of CKD were used to produce classifiers. Both experiments were performed with high accuracy. However, it might be costly in actual diagnosis, and patients must undergo an expensive clinical test to determine CKD stages.

In [61], 35,332 Electronic Health Records of hypertension patients were used for analysis. In this experiment, a hybrid classifier was implemented to predict CKD. The performance of this prediction model is more than 85 percent; however, the imbalanced dataset between CKD and non-CKD patients was not observed. Only 3.11 percent of data represent CKD, with the rest being non-CKD. The proper method to handle imbalanced data in this experiment was also not mentioned.

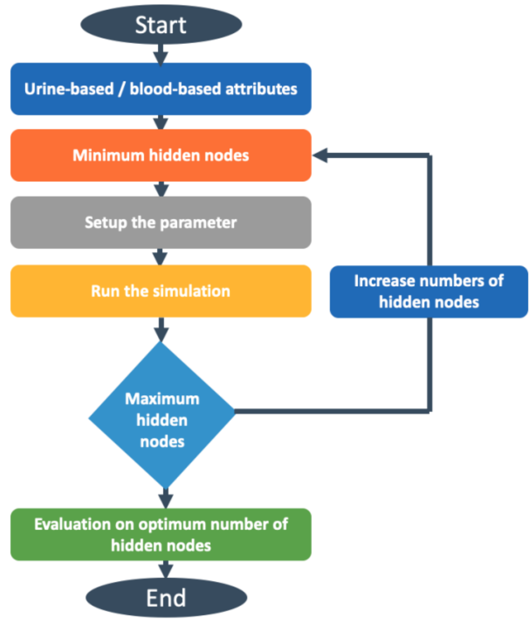

Fig. 2 Flowchart to determine the optimum number of hidden nodes.

TABLE 2. CKD/NON-CKD CLASSIFICATION RESULTS.

Class	Subject	Dataset	Method	Acc. (%)	References
CKD/ Non-CKD	250 CKD 150 non-CKD	UCI ML [62]	FRSBAS with Random Forest Classifier	99.5%	[31]
Risk of CKD on radiation therapy	CT images 29 CKD 21 non-CKD	50 adult patients with abdominal cancers	Random Forest Classifier	94%	[58]
CKD/ Non-CKD	4,940 augmented images of kidney	Radiology department, Hospital, Chennai, India [63][64]	SVM + RLN	87.31%	[59]
CKD/ Non-CKD	4,940 augmented images of kidney	Radiology department, Hospital, Chennai, India [63][64]	DNN	96.54%	[60]
CKD/ Non-CKD	250 CKD 150 non-CKD	UCI ML [62]	ANN	90.28%	[55]
CKD/ Non-CKD	250 CKD 150 non-CKD	UCI ML [62]	Probabilistic Neural Networks (PNN)	96.7%	[56]
CKD/ Non-CKD	250 CKD 150 non-CKD	UCI ML [62]	Hybrid Modified Cuckoo Search-Neural Network	99.6%	[57]
CKD/ Non-CKD	1,100 CKD 34,232 non-CKD	Electronic Health Records	Hybrid Neural Network	89.7%	[61]

[1]Class.: classification; bac. pneu.: bacterial pneumonia; Sens.: sensitivity; Spec.: specificity; Prec.: precision; Acc.: accuracy; Ref.: reference

IV. CHALLENGES IN CKD PREDICTION USING ML

This section reviews the data type, feature selection, and method of handling missing particular data. This chapter also discusses the difference between online repositories and real data from clinical practice.

4.1 Data Type and Form

This study comprises data collection and characterisation of urine and blood-based attributes, employing multiple imputation methods to generate synthetic data. The primary objective is the development of a CKD prediction model for both urine and blood-based attributes, followed by a comparative analysis.

An experiment to predict CKD utilized a dataset from a previous researcher, comprising 400 data samples obtained from the publicly available UCI Machine Learning Repository [62]. The dataset was divided into two distinct control groups: 250 samples representing the CKD group and the remaining 150 samples serving as the healthy control group. This database encompasses thirteen essential attributes for analysis.

The original dataset contains some missing data points. To address this issue, multiple imputations were performed, resulting in the creation of synthetic data points. Consequently, the overall dataset size was increased to N=2000. Notably, attributes such as Albumin, Sugar, Pus Cell, and Pus Cell Clump patterns were retained without alteration. However, it's worth noting that while Specific Gravity adopts a nominal value, its pattern deviates slightly from that of the original dataset.

The total accuracy of both urine-based and blood-based properties was 96.0 percent and 98.0 percent, respectively. Although only by a slight margin of 2.0 percent, the ANN prediction model constructed using blood-based variables outperformed the others. The model is constructed using urine-based features. However, this model is recommended for this study due to the following reasons: 1) The model requires only five input variables and utilizes eight hidden nodes. This reduction in the number of nodes contributes to a decrease in the model's computational complexity, rendering it more suitable for integration into intelligent diagnostic systems. 2) Urine-based attributes offer greater convenience and are non-invasive for both patients and medical practitioners, resulting in more efficient data collection and processing compared to blood-based attributes.

4.2 Feature Selection

The primary objectives of this analysis are to enhance diagnostic systems and decrease medical treatment expenses by facilitating early interventions [65]. Despite the ongoing efforts of numerous researchers, these systems have not yet reached their maximum potential. Many prior studies were constructed using open-source repositories that include incomplete data. Consequently, the process of feature selection, focusing on the most influential attributes in each dataset, is of paramount importance to optimize the predictive model. For example, in [66], Common Spatial Patterns (CSP) and Linear Discriminant Analysis (LDA) were proposed to filter and identify the dominant attributes in detecting CKD.

Furthermore, a study in 2018 implemented feature selection to provide an optimal solution before training the data with a different classifier [67]. Another two groups of researchers also proved that feature selection gave a significant method to improve the performance of the CKD prediction model. Findings from [68] implemented feature selection to enhance CKD diagnosis. Meanwhile, findings from [69] reduced the number of attributes from 25 to 14 using the feature selection approach. As a result, both studies were studied to gain the percentage of accuracy of the CKD detection model up to 98 percent.

Findings from [70] proposed feature selection adaptive probabilistic divergence-based feature selection (APDFS) with a combination of ML classifier hyper-parameterized logistic regression model (HLRM) in performing early prediction of CKD. The experiment showed only 19 of 52 attributes are used to identify chronic diseases with 91.6 percent accuracy. It can be concluded that feature selection will help to save money for patients by reducing clinical tests and increasing the ML classifier efficiency.

4.3 Missing Data

Several tests and experiments have been employed for CKD prediction, such as image processing from ultrasound, capturing disease trends using a signal, missing and imbalanced data, feature selection and a comparison of effectiveness between various classifiers.

Multiple imputations are one of the most common methods for dealing with missing data points. Multiple imputations, in contrast to a single imputation approach, enable a comprehensive examination of a dataset and the generation of coherent synthetic data points. This method incorporates statistical inference and includes illustrative case examples [71]. Through the iterative process of adding average values to the missing data, reliable imputed values can be obtained through three to five iterations [72]. It can also improve model performance by reducing errors caused by unequal distribution among the control groups [73]. As a result, multiple imputations become a significant tool for increasing the study's validity and reducing resource loss caused by missing data [36].

The average value from the whole data is used to generate missing data. Following that, the multiple imputation process is repeated five times, bringing the total number of samples per control group to 1000. Only numerical properties have their mean values implemented. Nominal qualities, on the other hand, are unaffected. The imputation process relies on a newly derived regression model, which is employed to estimate missing values for each variable [74].

This relationship can be represented by equation (1). When dealing with a variable Y_j that has missing values, a model is constructed using the available non-missing observations. This fitted model provides estimates for the regression parameters $(\beta_o, \beta_1, \dots, \beta_{(j-1)})$ and associated covariance matrix $\sigma_j Y_j$, where Y_j is the usual matrix from the intercept and variable $Y_1, Y_2, \dots, Y_{(j-1)}$.

$$Y_j = \beta_o + \beta_1 Y_1 + \beta_2 Y_2 + \cdots + \beta_{(j-1)} Y_{(j-1)} \tag{1}$$

For each imputation, new parameters (β_(*o),β_(*1),…,β_(*(j-1))) and σ_(*j) are drawn from the posterior predictive distribution of the missing data. The missing values of the original data are then replaced by an expression of (2).

$$\beta_o + \beta_1 y_1 + \beta_2 y_2 + \cdots + \beta_{(j-1)} y_{(j-1)} + z_i \sigma_j \tag{2}$$

Where $y_1, y_2, \dots, y_{(j-1)}$ are the covariate values of the first (*j-1*) variables and zi is a simulated normal deviation.

4.4 Online Repository versus Real Data from Clinical Practise

Online repositories and real data from clinical practice are significant for evaluating the safety and efficacy of the chronic disease prediction model. Both types of data have come out from the sources that were collected with real patients and scenarios. However, online repositories might contain outdated data that are not relevant to the latest experiment. Therefore, fresh data within a population is very important to get an accurate result.

Findings from [75] mentioned that the impact of utilisation of real clinical data can shed light on how well medications work over the long term in big, diverse groups of people. However, the management of clinical datasets must follow the standard and adhere to the full protocol to avoid data leakage that can lead to concerns regarding data security. Therefore, a group of researchers in China [76] proposed improvements in handling electronic medical records (EMR) in their country. The researcher recommended modernising hospital information systems, pushing for data standards, and setting up a separate platform for clinical research. The suggestion may be to fix the current issue of clinical data in China.

V. WAY FORWARD IN CKD PREDICTION

Many research works have introduced a lot of prediction models for CKD. Among the studies implement ML, neural network, deep learning, as well a hybrid prediction model, consisting of two major ML approaches. Generally, previous studies always implement datasets from online repositories in fact, several studies conduct experiments from real data collection.

In future work, CKD prediction should be based on invasive and non-invasive methods. Therefore, data on CKD should be separated into two different groups, which are blood-based and urine-based analytes [77]. The CKD dataset also includes with GFR value that has been calculated based on age, gender, and race to perform stage prediction.

Currently, this paper focuses on developing a CKD prediction model using the ANN approach. In this paper, we also apply multiple imputations to prevent and reduce the missing value of CKD datasets [78]. In the future, Deep Learning techniques like Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), or Recurrent Neural Networks (RNN) will be employed for predicting the likelihood of CKD in patients.

The ultimate goal of all these prediction models should be actionable information for either patients or clinicians. This paper thus concurrently developed an Internet of Things-based kidney home monitoring system using urine samples [79]. The system does a daily analysis of urine concentration in the bathroom. Due to the asymptotic nature of CKD, the basic idea is that individuals might identify early indicators of the disease at home. The technology is non-invasive, and users may use it without help from medical professionals. The technology connects to users' mobile apps, collects data from their urine samples, and sends alerts based on the results. The app will then alert the user, whether a doctor's diagnosis is necessary for their health condition.

VI. CONCLUSION

Recent breakthroughs in AI methodologies have led to the development of successful AI applications in the field of CKD. Even the possibility that AI expert systems would one day replace human physicians has been a heated issue of debate. Despite this, we explore the possibility that an AI expert system can assist a human physician in making better decisions, and in certain cases even replace human judgment. Various AI systems can help extract vital information from vast quantities of clinical data. Recently, ML has demonstrated considerable promise for early CKD diagnosis.

In addition, ML systems are trained to be capable of self-learning, error correction, and generating accurate results. This paper investigated the application of ML techniques in diagnosing CKD. In this paper, the effect of ML algorithms and their consistency on CKD diagnosis was evaluated to reduce misdiagnosis mistakes. To achieve the primary objective, this paper devised a search strategy. In this prospect, different scientific journals, including Google Scholar, IEEE, ScienceDirect, Web of Science, Wiley Online Library, and Elsevier, were chosen to fetch published scientific papers.

All the retrieved papers were organised according to their authors, publication years, ML approaches used for various diseases, results, and, finally, the future of AI-based disease detection methods. The data indicate that the number of papers published in the medical profession has increased rapidly. Another aim of this paper was to reveal the standard methodologies for applying ML in disease detection. Next, following most researchers, this paper investigated which ML method was most effective for CKD diagnosis. Based on the findings, this paper concluded that ML-assisted diagnosis enhanced the diagnostic process and identified CKD in its early stages, enabling the selection of the most effective treatment approach. In addition, the impact of each ML technique was evaluated based on the accuracy of the CKD diagnosis described in the literature.

Aside from that, it was observed that many publications had reported a combination of ML approaches to detect any disease or to enhance the diagnostic process. Afterwards, this paper summarised the challenges in CKD prediction using ML. As such, it was concluded that data analytics confronted several issues in utilising ML for CKD prediction, namely the online respiratory. The main issues revolved around different data types, feature selection, and dealing with missing data in the respiratory itself. Future researchers should gather first-hand real data from clinical practices for better accuracy.

AI is not confined to recognising any specific disease. The results of this paper may therefore be useful for research in the future. Furthermore, we found that over 91 percent of AI approaches had a favourable influence on illness detection in this paper. The effectiveness of AI in detecting CKD cannot be overlooked. Using AI methods, medical journal papers published within a specific decade were observed. In future investigations, more comprehensive studies should be planned to determine if deep learning methods may indeed enhance CKD diagnosis. Furthermore, the data type for deep learning approaches should be studied which is more effective. Finally, future research should include a thorough evaluation of AI's economic impact on healthcare in general.

ACKNOWLEDGEMENT

The authors would like to express their deepest appreciation to the Ministry of Higher Education Malaysia (MOHE) through the Fundamental Research Grant Scheme (FRGS/1/2018/TK04/UITM/02/21) for funding this research. This research is also supported by Universiti Teknologi MARA (UiTM).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of the paper.

AUTHOR CONTRIBUTION

The authors confirm their contribution to the paper as follows: **Writing–Original Draft:** MHAG; **Conceptualization:** AHAR; **Formal Analysis:** AHAR and SAMAJ; **Investigation:** SAMAJ; **Writing–Review & Editing:** AHAR, MSAMA and MF. All authors reviewed and approved the final version of the manuscript.

REFERENCES

[1] A. Nasir, K. Shaukat, I. A. Hameed, S. Luo, T. M. Alam and F. Iqbal, "A Bibliometric Analysis of Corona Pandemic in Social Sciences: A Review of Influential Aspects and Conceptual Structure," in IEEE Access, vol. 8, pp. 133377-133402, 2020, doi: 10.1109/ACCESS.2020.3008733.

[2] X. Y. Zhou, W. S. Chan, S. Chen and W. J. Feng, "Broadband Highly Efficient Doherty Power Amplifiers," in IEEE Circuits and Systems Magazine, vol. 20, no. 4, pp. 47-64, Fourthquarter 2020, doi: 10.1109/MCAS.2020.3027221.

[3] V. Camarchia, M. Pirola, R. Quaglia, S. Jee, Y. Cho and B. Kim, "The Doherty Power Amplifier: Review of Recent Solutions and Trends," in IEEE Transactions on Microwave Theory and Techniques, vol. 63, no. 2, pp. 559-571, Feb. 2015, doi: 10.1109/TMTT.2014.2387061.

[4] Y. Yin et al., "A Broadband Switched-Transformer Digital Power Amplifier for Deep Back-Off Efficiency Enhancement," in IEEE Journal of Solid-State Circuits, vol. 55, no. 11, pp. 2997-3008, Nov. 2020, doi: 10.1109/JSSC.2020.3005798.

[5] Y. Yin, L. Xiong, Y. Zhu, B. Chen, H. Min and H. Xu, "A Compact Dual-Band Digital Polar Doherty Power Amplifier Using Parallel-Combining Transformer," in IEEE Journal of Solid-State Circuits, vol. 54, no. 6, pp. 1575-1585, June 2019, doi: 10.1109/JSSC.2019.2896407.

[6] I. A. Alimi, R. K. Patel, N. J. Muga, A. N. Pinto, A. L. Teixeira, and P. P. Monteiro, "Towards Enhanced Mobile Broadband Communications: A Tutorial on Enabling Technologies, Design Considerations, and Prospects of 5G and beyond Fixed Wireless Access Networks," Applied Sciences, vol. 11, no. 21, p. 10427, Nov. 2021, doi: https://doi.org/10.3390/app112110427.

[7] G. Nikandish, R. B. Staszewski and A. Zhu, "Breaking the Bandwidth Limit: A Review of Broadband Doherty Power Amplifier Design for 5G," in IEEE Microwave Magazine, vol. 21, no. 4, pp. 57-75, April 2020, doi: 10.1109/MMM.2019.2963607.

[8] A. Tunç and M. B. Yelten, "A Theoretical Efficiency Analysis of Doherty Power Amplifiers built with Different Amplifier Classes," 2021 29th Telecommunications Forum (TELFOR), Belgrade, Serbia, 2021, pp. 1-4, doi: 10.1109/TELFOR52709.2021.9653262.

[9] G. Nikandish, R. B. Staszewski and A. Zhu, "Breaking the Bandwidth Limit: A Review of Broadband Doherty Power Amplifier Design for 5G," in IEEE Microwave Magazine, vol. 21, no. 4, pp. 57-75, April 2020, doi: 10.1109/MMM.2019.2963607.

[10] V. Diddi, H. Gheidi, J. Buckwalter and P. Asbeck, "High-power, high-efficiency digital polar doherty power amplifier for cellular applications in SOI CMOS," 2016 IEEE Topical Conference on Power Amplifiers for Wireless and Radio Applications (PAWR), Austin, TX, USA, 2016, pp. 18-20, doi: 10.1109/PAWR.2016.7440131.

[11] J. Singh and D. Kedia, "Efficient Technique for PAPR and BER Reduction in Downlink Large Scale MU-MIMO-OFDM System," International Journal of Sensors Wireless Communications and Control, vol. 11, no. 8, pp. 845–857, Nov. 2020, doi: https://doi.org/10.2174/2210327910999201124210039.

[12] A. Giry et al., "A 2.5GHz LTE Doherty Power Amplifier in SOI-CMOS Technology," 2019 17th IEEE International New Circuits and Systems Conference (NEWCAS), Munich, Germany, 2019, pp. 1-4, doi: 10.1109/NEWCAS44328.2019.8961306.

[13] V. E. Athira and K. J. Nelson, "Bridging the Broadband Gap and Power Up To 5G," 2020 International Conference on Power, Instrumentation, Control and Computing (PICC), Thrissur, India, 2020, pp. 1-4, doi: 10.1109/PICC51425.2020.9362495.

[14] M. Sajedin, I. Elfergani, J. Rodriguez, R. Abd Alhameed, N. O. Parchin and Y. I. A. Al-Yasir, "Advancement of a Highly Efficient Class-F power Amplifier for 5G Doherty Architectures," 2019 IEEE 2nd 5G World Forum (5GWF), Dresden, Germany, 2019, pp. 86-90, doi: 10.1109/5GWF.2019.8911678.

[15] S. Bhardwaj and J. Kitchen, "Broadband Parallel Doherty Power Amplifier in GaN for 5G Applications," 2019 IEEE Topical Conference on RF/Microwave Power Amplifiers for Radio and Wireless Applications (PAWR), Orlando, FL, USA, 2019, pp. 1-3, doi: 10.1109/PAWR.2019.8708727.

[16] V. Vorapipat, C. S. Levy and P. M. AsbeckIEEE, "A Class-G Voltage-Mode Doherty Power Amplifier," in IEEE Journal of Solid-State Circuits, vol. 52, no. 12, pp. 3348-3360, Dec. 2017, doi: 10.1109/JSSC.2017.2748283.

[17] M. S. Mugisho, S. K. Maity, C. Friesicke and R. Quay, "Harmonic-Injection Doherty Power Amplifiers with a High Small-Signal Gain," 2022 International Workshop on Integrated Nonlinear Microwave and Millimetre-Wave Circuits (INMMiC), Cardiff, United Kingdom, 2022, pp. 1-3, doi: 10.1109/INMMiC54248.2022.9762221.

[18] W. Dong, W. Kong and J. Xia, "A Highly Efficient Doherty Power Amplifier Based on Harmonic-Tuned Technique for 5G Application," 2021 Cross Strait Radio Science and Wireless Technology Conference (CSRSWTC), Shenzhen, China, 2021, pp. 157-159, doi: 10.1109/CSRSWTC52801.2021.9631586.

[19] G. Mu and H. Hayashi, "Enhancing Power Added Efficiency of Doherty Amplifier by Changing Power Ratio of Carrier Amplifier and Peak Amplifier," 2021 3rd International Conference on Computer Communication and the Internet (ICCCI), Nagoya, Japan, 2021, pp. 176-179, doi: 10.1109/ICCCI51764.2021.9486833.

[20] H. Chen, J. -X. Xu, W. Chen and X. Y. Zhang, "High-Efficiency Dual-Band Filtering Doherty Power Amplifier Based on Multi-Function Circuit," in IEEE Transactions on Microwave Theory and Techniques, vol. 70, no. 5, pp. 2697-2709, May 2022, doi: 10.1109/TMTT.2022.3154756.

[21] C. Liang, J. I. Martinez-Lopez, P. Roblin, Y. Hahn, D. Mikrut and V. Chen, "Wideband Two-Way Hybrid Doherty Outphasing Power Amplifier," in IEEE Transactions on Microwave Theory and Techniques, vol. 69, no. 2, pp. 1415-1428, Feb. 2021, doi: 10.1109/TMTT.2020.3019430.

[22] X. Y. Zhou, W. S. Chan, T. Sharma, W. J. Jie and J. Xia, "Harnessing Harmonics in Doherty Power Amplifiers [Application Notes]," in IEEE Microwave Magazine, vol. 22, no. 8, pp. 16-31, Aug. 2021, doi: 10.1109/MMM.2021.3078042.

[23] R. -Z. Zhan, Y. C. Li and H. Zhang, "A Highly Efficient 60 GHz CMOS Doherty Power Amplifier With Adaptive Gate Biases," 2020 IEEE MTT-S International Microwave Workshop Series on Advanced Materials and Processes for RF and THz Applications (IMWS-AMP), Suzhou, China, 2020, pp. 1-3, doi: 10.1109/IMWS-AMP49156.2020.9199696.

[24] N. Elsayed, H. Saleh, B. Mohammad, M. Ismail, and Mihai Sanduleanu, High Efficiency Power Amplifier Design for 28 GHz 5G Transmitters. Springer Nature (Netherlands), 2022. doi: https://doi.org/10.1007/978-3-030-92746-2.

[25] A. Jarndal and K. Husna Hamza, "On the Design of Broadband Doherty Power Amplifier for 5G Applications," 2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA), Ras Al Khaimah, United Arab Emirates, 2022, pp. 370-373, doi: 10.1109/ICECTA57148.2022.9990305.

[26] Adish, C. Jayaram and S. P, "Design of GaN HEMT based Doherty Power Amplifier for 5G Applications," 2021 IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkur, Karnataka, India, 2021, pp. 1-5, doi: 10.1109/ICMNWC52512.2021.9688386.

[27] S. Kim, H. -C. Park, D. Kang, D. Minn and S. -G. Yang, "A 24.5–29.5GHz Broadband Parallel-to-Series Combined Compact Doherty Power Amplifier in 28-nm Bulk CMOS for 5G Applications," 2021 IEEE Radio Frequency Integrated Circuits Symposium (RFIC), Atlanta, GA, USA, 2021, pp. 171-174, doi: 10.1109/RFIC51843.2021.9490410.

[28] J. Gu et al., "A 28-GHz Series-Parallel Combined Doherty Power Amplifier with PBO Efficiency Enhancement in 40nm Bulk CMOS," 2022 IEEE MTT-S International Wireless Symposium (IWS), Harbin, China, 2022, pp. 1-3, doi: 10.1109/IWS55252.2022.9977668.

[29] T. Götthans, R. Maršálek, J. Král and T. Urbanec, "Linearity and Efficiency Enhancement Techniques for Satellite Communications," 2022 IEEE Topical Conference on RF/Microwave Power Amplifiers for Radio and Wireless Applications (PAWR), Las Vegas, NV, USA, 2022, pp. 54-57, doi: 10.1109/PAWR53092.2022.9719692.

[30] K. P. Jung, O. S. Choi and C. S. Park, "A Broadband CMOS Doherty Power Amplifier with Asymmetric-Transformer-based Output Matching Network," 2022 14th Global Symposium on Millimeter-Waves &

Terahertz (GSMM), Seoul, Korea, Republic of, 2022, pp. 110-112, doi: 10.1109/GSMM53818.2022.9792333.

[31] C. Liang and P. Roblin, "The Analytic Doherty-Outphasing Power Amplifiers Continuum Theory : (Invited Paper)," 2022 IEEE 22nd Annual Wireless and Microwave Technology Conference (WAMICON), Clearwater, FL, USA, 2022, pp. 1-4, doi: 10.1109/WAMICON53991.2022.9786088.

[32] Z. Zhang, Z. Cheng, and G. Liu, "A Power Amplifier with Large High-Efficiency Range for 5G Communication," Sensors, vol. 20, no. 19, pp. 5581–5581, Sep. 2020, doi: https://doi.org/10.3390/s20195581.

[33] A. Tunç and M. B. Yelten, "The Design Overview of the Hybrid Doherty-Outphasing Technique," 2021 13th International Conference on Electrical and Electronics Engineering (ELECO), Bursa, Turkey, 2021, pp. 1-4, doi: 10.23919/ELECO54474.2021.9677816.

[34] C. Yu, J. Feng and D. Zhao, "A 28-GHz Doherty Power Amplifier With a Compact Transformer-Based Quadrature Hybrid in 65-nm CMOS," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 68, no. 8, pp. 2790-2794, Aug. 2021, doi: 10.1109/TCSII.2021.3068473.

[35] J. Lee, H. -C. Park, S. Baek, J. Jung, T. Kim and S. -G. Yang, "A 45-nm RFSOI CMOS Based Compact Doherty Power Amplifier For mm-wave 5G Applications," 2022 14th Global Symposium on Millimeter-Waves & Terahertz (GSMM), Seoul, Korea, Republic of, 2022, pp. 1-3, doi: 10.1109/GSMM53818.2022.9792352.

[36] Azliyana Azizan and A. Zaki, "Educational Insights from Bibliometric Patterns: Examining Depression Research in Malaysia," Asian Journal of Research in Education and Social Sciences, vol. 5, no. 3, pp. 33–47, 2023, Accessed: Sep. 26, 2023. [Online]. Available: https://myjms.mohe.gov.my/index.php/ajress/article/view/23192

[37] Elsevier, "Data | Curated. Connected. Complete.," Elsevier.com, 2019. https://www.elsevier.com/solutions/scopus

[38] P. Mongeon and A. Paul-Hus, "The Journal Coverage of Web of Science and Scopus: a Comparative Analysis," Scientometrics, vol. 106, no. 1, pp. 213–228, Oct. 2016, doi: https://doi.org/10.1007/s11192-015-1765-5.

[39] Clarivate, "Web of Science - Web of Science Group," Web of Science Group, 2019. https://clarivate.com/webofsciencegroup/solutions/web-of-science/

[40] C.-P. Hu, J.-M. Hu, S.-L. Deng, and Y. Liu, "A co-word analysis of library and information science in China," Scientometrics, vol. 97, no. 2, pp. 369–382, Jul. 2013, doi: https://doi.org/10.1007/s11192-013-1076-7.

[41] N. J. van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," Scientometrics, vol. 84, no. 2, pp. 523–538, Dec. 2010, doi: https://doi.org/10.1007/s11192-009-0146-3.

[42] B. Jarneving, "Bibliographic coupling and its application to research-front and other core documents," Journal of Informetrics, vol. 1, no. 4, pp. 287–307, Jan. 2007, doi: https://doi.org/10.1016/j.joi.2007.07.004.

[43] X. Zhang, S. Li, D. Huang and T. Chi, "A Millimeter-Wave Three-Way Doherty Power Amplifier for 5G NR OFDM," in IEEE Journal of Solid-State Circuits, vol. 58, no. 5, pp. 1256-1270, May 2023, doi: 10.1109/JSSC.2023.3238766.

[44] J. Du, Z. Li and Z. Li, "A 2.4 GHz Doherty Power Amplifier Based on Voltage Combining in 22 nm CMOS," 2023 12th International Conference on Modern Circuits and Systems Technologies (MOCAST), Athens, Greece, 2023, pp. 1-4, doi: 10.1109/MOCAST57943.2023.10176649.

[45] K. Koca, H. Ş. Savcı and S. Kent, "A Wideband High-efficiency Doherty Power Amplifier for LTE," 2023 33rd International Conference Radioelektronika (RADIOELEKTRONIKA), Pardubice, Czech Republic, 2023, pp. 1-5, doi: 10.1109/RADIOELEKTRONIKA57919.2023.10109057.

[46] J. R. Zhang, S. Y. Zheng and N. Yang, "An Efficient Broadband Symmetrical Doherty Power Amplifier With Extended Back-Off Range," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 70, no. 4, pp. 1316-1320, April 2023, doi: 10.1109/TCSII.2022.3227045.

[47] H. Hou, H. Kretschmer, and Z. Liu, "The structure of scientific collaboration networks in Scientometrics," Scientometrics, vol. 75, no. 2, pp. 189–202, Dec. 2007, doi: https://doi.org/10.1007/s11192-007-1771-3.