Abstract—Chronic Kidney Disease is one of the leading causes of death worldwide. An intelligent diagnostic system capable of early detection is therefore, becoming increasingly important. These would allow effective intervention to be delivered to patients, thus prolonging kidney function and reducing risk of mortality. The system should be non-invasive, convenient, accurate, and reliable in detecting the required attributes. This study compares between urine- and blood-based attributes in acute renal failure prediction using artificial neural network. A total of 400 sample data is obtained from UCI Machine Learning Repository. Multiple imputation is then implemented to generate synthetic data. These overcome the issue of missing datapoints and unbalanced sample distribution. Two artificial neural network models are trained. One using the urine-based attributes and the other, using blood-based attributes. Both models attained excellent classification accuracies of 96.0% and 98.0%, respectively. However, the ANN model developed based on urine-based attributes are recommended intelligent diagnostic systems due to lower computational requirements and the sample acquisition protocol is much convenient for patients and medical practitioners.

Index Terms—Acute renal failure, blood, urine, multiple imputation, artificial neural network

I. INTRODUCTION

IDENTIFYING and profiling important attributes of chronic diseases have led to the development intelligent diagnostic systems that can perform early detection. These allow medical practitioners to anticipate and deliver effective intervention to patients, thus prolonging the function of critical organs and reducing risk of death [1]. Chronic Kidney Disease (CKD) is the fastest growing cause of death with at least 2.4 million cases reported every year [2].

CKD is defined as kidney damage, or glomerular filtration rate of less than 60ml/min/1.73 m2 for duration of more than three months [3]. Excessive consumption of sugar can lead the causes of kidney damage. High volume of glucose will increase the function of kidney to filtering wastes and extra fluid from the blood [4]. Thus, the initial sign of CKD is existence protein in urine called albumin. A healthy kidney does not let albumin pass from blood into the urine [5]. CKD is divided into five stages with the worse being the end-stage renal disease. This requires the patient to undergo kidney replacement [6].

Based on the reports from American Society of Nephrology, International Society of Nephrology, European Renal Association-European Dialysis and Transplant Association in 2017, the total number of individuals with CKD, acute kidney injury, and those on renal replacement therapy exceeds 850 million. The figure doubles the number of diabetes cases. The Associations also predicted that CKD will be the top 5 causes of death worldwide [7]. The Institute of Health Metric and Evaluation reported 14% of women and 12% of men are affected by the diseases [8]. Based on the 24th Report of Malaysian Dialysis and Transplant Registry in 2016, the number of CKD cases steadily increased for the past 10 years [9].

Previous work related to CKD focuses on detection of the disease and its stages from various diagnostic methods. Among of the diagnosis method is feature extraction from the ultrasound kidney images [10]. In other method, CKD also can be predicted by the parameter of two main clinical tests, which are the urine and blood tests. Urine test or Urinalysis includes three parts of exam which are visual, microscopic and dipstick test. Several parameters can be extracted from urine test such that albuminuria, specific gravity, protein and glucose. Albumin is a protein contained in the blood, but if it is found too much in urine, the kidney would not function normally and...
urea nitrogen (BUN) is used to measure the amount of nitrogen. Another type of data is based on continuous value. For example, blood test is significant to determine kidney level because the amount of albumin in the urine is specific to chronic kidney disease. Specific gravity is a ratio of urine to water density. The range of this measurement is between 1.005 and 1.030. Meanwhile, there are several clinical blood tests for kidney diseases such as random blood glucose, blood urea nitrogen, serum creatinine, serum sodium and potassium lab test. Both blood test significant to determine kidney level because the type of data is based on continuous value. For example, blood urea nitrogen (BUN) is used to measure the amount of nitrogen in blood. The level of BUN is rising when kidney is not able to remove urea from the blood normally.

These tests can be complemented by machine learning techniques that enable generalization of attributes between normal subjects and those with CKD. In fact, there are already attempts to model clinical traits with different stages of the disease. The established machine learning methods include convolutional neural network, support vector machines, decision tree classifier and artificial neural network (ANN). The goals are to improve diagnostic systems and reduce cost of medical treatment through early interventions. Despite the continual effort by various researchers, the systems have yet to attain its optimum potential. Majority of previous works are built from open-source repositories that contains missing data. The unrecorded test attributes have been not only among the types of attributes, but also across individual. Studies have shown that these deficiencies have an adverse effect on prediction performance and needs to be solved prior to development of the computational model.

Among the widely used method for dealing with missing datapoints is multiple imputation. When compared to single imputation method, multiple imputation enables analysis of dataset and utilize it to obtain consistent synthetic data points. The method introduces statistical inference and illustrate references to the respective cases. These can be achieved through three to five repetitions by adding average values to the missing data. Furthermore, it can reduce errors caused by unbalanced distribution among the control groups, thus improving the model performance. Multiple imputation is, therefore, an important method to improve validity of the study and reduce waste of resources caused by missing data.

ANN is one of the established machine learning techniques. The approach mimics the biological functioning of neurons in the brain and learn from given examples. Although considered as one of the pioneering artificial intelligence paradigms, it remains relevant due to its ability to generalize solution to a problem. The method has been widely implemented in modelling physiological phenomena such as cognitive behaviour, dengue prediction, and heart diseases. Thus far, ANN has also been used in studies involving CKD. However, the findings are debatable as it lacks technical clarity on the pre-processing and modelling aspects.

Despite the growing body of literature related to CKD prediction using artificial intelligent technique, the following gaps have been identified. 1) The literature has presented a myriad of testing methods to obtain the desired attributes. However, none has emphasized on the importance of selecting the most convenient, non-invasive, and cost-effective attributes; yet maintaining the prediction accuracy and reliability. 2) In studies related to CKD, most of the work on handling missing data and unbalanced dataset involves the use of primitive techniques such as attribute removal, and correlation process to identify the significant parameters. However, methods such as multiple imputation is equally important for accurately generating reliable data points. 3) Although there are already studies on CKD involving ANN, the findings were not well presented as it lacks the technical descriptions on both data pre-processing and development of model, rendering the results debatable.

To solve these issues, the following objectives are proposed. First, the study aims to characterize the attributes based on urine and blood tests. Imaging techniques such as computed tomography has been ruled out as it relatively more expensive than clinical laboratory tests. Second, the study aims to implement multiple imputation method to solve missing data points within the dataset and balance out data distribution between those with CKD and healthy controls. Third, two prediction models are developed. One model uses the urine-based attributes, while the other model uses the blood-based attributes. A comparative analysis is then performed to reach a conclusive recommendation.

II. METHODS
The study comprises of data acquisition, characterization of urine- and blood-based attributes, implementation of multiple imputation method for generating synthetic data, development of CKD prediction for urine- and blood-based attributes, as well as comparative analysis. These are summarized in the flowchart shown in Figure 1.
Ghafar et.al.: Comparative Performance of Urine and Blood-based Attributes in Chronic Kidney Disease Prediction using Artificial Neural Network

Fig. 1. Overview of research methods.

A. Data Acquisition
A total of 400 data samples were acquired from the open-source UCI Machine Learning Repository [31]. The dataset is further segregated into two control groups: 250 samples in the CKD group and the remaining 150 samples as healthy controls. The database provides thirteen critical attributes. Table I shows the five urine-based and nine blood-based attributes.

B. Multiple Imputation
The missing data is generated by considering the average value from the overall data. Subsequently, five repetitions are made through the multiple imputation process: enhancing the total number of samples to 1000 for each control group. The mean values are only implemented for numerical attributes. Meanwhile, nominal attributes remain unchanged. The concept of imputation is based on the resulting model, a new regression model used to impute the missing value for each variable [32].

This can be expressed by (1). Y_j with missing value, a model is fitted with the non-missing observations. The fitted model has the regression parameters estimates (β_0, β_1, ..., β_{(j-1)}) and associated covariance matrix σ_jY_j, where Y_j is the usual matrix from the intercept and variable Y_1, Y_2, ..., Y_{(j-1)}.

Y_j = β_0 + β_1Y_1 + β_2Y_2 + ... + β_{(j-1)}Y_{(j-1)}                             (1)

For each imputation, new parameters (β_{*0}, 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 expression of (2)

β_{*0} + β_{*1}y_1 + β_{*2}y_2 + ... + β_{*_{(j-1)}}y_{(j-1)} + z_iσ_{*j}                           (2)

Where y_1, y_2, ..., y_{(j-1)} are the covariate values of the first (j-1) variables and z_i is a simulated normal deviate.

C. Artificial Neural Network
The modelling tasks in this study is performed using MATLAB. Generally, the ANN comprises of an input layer, several hidden layers, and an output layer [33]. However, a study had shown that a network with single hidden layer is sufficient to approximate and arbitrary function of to an acceptable level of accuracy [34]. The number of input nodes will depend on the number of attributes to the network. Meanwhile, single output node is used for the study. The optimum number of hidden nodes is determined based on a separate set of optimization experiment which considers unique traits observed from network training [35].

Vector of input variables x_i is transformed into a vector of hidden variables, h_j, via activation function, Γ_1. This can be expressed by (3). w_ij is the connecting weights from i-th input node to j-th hidden node, and θ_j represents the biases. Γ_1 employs the hyperbolic tangent function.

h_j = Γ_1\left(\sum_{i=1}^{M} w_{ij}x_i + \theta_j\right)                           (3)

<table>
<thead>
<tr>
<th>Urine-based</th>
<th>Blood-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specific gravity</td>
<td>Blood glucose random</td>
</tr>
<tr>
<td>Albumin</td>
<td>Blood urea</td>
</tr>
<tr>
<td>Sugar</td>
<td>Serum creatinine</td>
</tr>
<tr>
<td>Pus Cell</td>
<td>Sodium</td>
</tr>
<tr>
<td>Pus Cell Clumps</td>
<td>Potassium</td>
</tr>
<tr>
<td></td>
<td>Hemoglobin</td>
</tr>
<tr>
<td></td>
<td>Packed cell volume</td>
</tr>
<tr>
<td></td>
<td>White blood cell count</td>
</tr>
<tr>
<td></td>
<td>Red blood cell count</td>
</tr>
</tbody>
</table>

TABLE I

ATTRIBUTES
The vector of hidden variables, \( h_j \), is then transformed into the output, \( y_k \), through activation, \( \Gamma_2 \). This is expressed by (4). \( w_{jk} \) is the connecting weights between \( j \)th hidden node to \( k \)th output node, and \( \theta_k \) represents the bias. \( \Gamma_2 \) employs the pure linear function.

\[
y_k = \Gamma_2 \left( \sum_{j=1}^{N} w_{jk} h_j + \theta_k \right)
\]  
(4)

The general structure of ANN is illustrated by Fig. 2.

![General structure of ANN](image)

As shown by (5), the error, \( e \), is obtained as the difference between the output, \( y_k \), and the desired output, \( y_d \).

\[
e = y_k - y_d
\]  
(5)

The error is then used in the back-propagation weight-update procedure through suitable learning algorithms. In this study, Levenberg-Marquardt algorithm is adopted. Network training will continue in an iterative manner until the error is being minimized [36]. Data is initially randomized. 70% is used for training, 15% is used for validation and the remaining 15% is used for testing. The validation set is used to avoid the network from over-fitting. Throughout training, data from the validation set is periodically used to gauge the network’s generalization ability. If the validation error increases, network training is stopped and the set of network weights prior to that is used in the final structure. The performance of the model is then assessed using the testing sets [37].

Two models will be developed using ANN: one using the urine-based attributes and the other using blood-based attributes. The first model adopts a five-input structure, and the second model will take on a nine-input structure. The number of hidden nodes, however, relies on an experiment performed using the rule of thumb. The lower limit is set as 2/3 the size of input and output layers. Meanwhile, the upper limit should be less than twice the size of the input layer [38]. The experiment starts by training the ANN using the lower limit setting for 40 iterations and averages its performance. As shown in Fig. 3, the procedure is then repeated until it reaches the maximum limit.

![Flowchart to determine optimum number of hidden nodes](image)

The evaluation will consider the best average training accuracy and mean squared error (MSE). Accuracy (\( Acc \)) is expressed by (6), where \( TP \) is the true positives, \( TN \) is the true negatives, \( FP \) is the false positives, and \( FN \) is the false negatives.

\[
Acc = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%
\]  
(6)

Meanwhile, MSE is presented by (7), where \( e(t) \) is the error at iteration instant, \( t \), and \( N \) is the total number of iterations.

\[
MSE = \frac{1}{N} \sum_{t=1}^{N} e(t)
\]  
(7)
In this study, samples with CKD is given the output index of 1, while healthy samples are given the output index of 2. Once the optimum number of hidden nodes is finalized, the network is trained to obtain the best results. These are assessed not only in terms of accuracy, but also sensitivity ($Se$) and positive predictivity ($Pp$). Both parameters are each expressed by (8) and (9).

\[
Se = \frac{TP}{TP + FN} \times 100\% \quad (8)
\]

\[
Pp = \frac{TP}{TP + FP} \times 100\% \quad (9)
\]

III. RESULTS AND DISCUSSION

A. Data Acquisition

The five urine-based attributes are specific gravity, albumin, sugar, pus cell and pus cell clumps. These are comprised of both numerical and nominal data. The distribution of attributes is shown in Fig. 4 to Fig. 8.

Meanwhile, the blood-based attributes consist of blood glucose, blood urea, serum creatinine, sodium, potassium, haemoglobin, packed cell volume, white blood cell count and red blood cell count. Compared to the urine, blood-based attributes are all composed of numerical data. The distribution of attributes is shown in Fig. 9 to Fig. 17.
Fig. 9. Blood glucose random (N=400).

Fig. 10. Blood urea (N=400).

Fig. 11. Serum creatinine (N=400).

Fig. 12. Sodium (N=400).

Fig. 13. Potassium (N=400).

Fig. 14. Haemoglobin (N=400).
B. Multiple Imputation

There are missing data points from the original dataset. Hence, multiple imputation generates the synthetic version of them, and the overall dataset is enhanced to N=2000. The new distribution of attributes for urine-based attributes are shown in Fig. 18 to Fig. 22. The pattern of albumin, sugar, pus cell and pus cell clumps remain unchanged. Although specific gravity is takes on a nominal value, the pattern slightly differs from the original dataset.
Meanwhile, the new distribution of blood-based attributes is shown in Fig. 23 to Fig. 31. The pattern for these attributes slightly differs from the original dataset but contains more outlier due to the increasing volume of samples.
C. **Artificial Neural Network**

Initially, the experiment is conducted to determine the optimum number of hidden nodes for both urine-based attributes and blood-based attributes. The lower- and upper limits for urine-based attributes was set at four and nine hidden nodes. The average training accuracy and MSE is shown in Fig. 32 and Fig. 33. Based on the results obtained, the optimum number of hidden nodes for the ANN model is eight.

The network is then trained, validated, and tested with the optimum structure, yielding the results shown in Table II.
Meanwhile, the lower- and upper limits for urine-based attributes was set at seven and seventeen hidden nodes. The average training accuracy and MSE is shown in Fig. 34 and Fig. 35. Based on the results obtained, the optimum number of hidden nodes for the ANN model is seventeen.

The network is then trained, validated, and tested with the optimum structure, yielding the results shown in Table III.

D. Comparative Analysis

Both the urine-based and blood-based attributes has each attained excellent overall accuracy of 96.0% and 98.0%. The ANN prediction model developed using blood-based attributes has shown superior performance, although by a small margin of 2.0%. However, this study would recommend the model developed based on urine-based attributes based on the following reasons: 1) The model only requires five input and eight hidden nodes. These reduces the computational complexity of the model and is relatively more suitable to be implemented in intelligent diagnostic systems. 2) The use of urine-based attributes presents a more convenient and non-invasive method for both patients and medical practitioners, therefore rendering data collection and analysis more effective than acquisition of blood-based attributes.

IV. Conclusion

The study initially sets out with the following objectives. First, the study aims to characterize the attributes based on urine and blood tests. Second, the study aims to implement multiple imputation method to solve missing data points within the dataset and balance out data distribution between those with CKD and healthy controls. Third, the study aims to develop two ANN prediction models based on urine-based and blood-based attributes. A comparative analysis is then performed to reach a conclusive recommendation.

Generally, of those objectives have been achieved. The study was able to characterize the distribution of attributes between samples with CKD and healthy controls. The missing datapoints are dealt with using multiple imputation technique. The method was also able to generate synthetic samples, enhancing the dataset to N=2000. Prior to developing the prediction models, the ANN structure is optimized for urine-based and blood-based attributes. Two ANN prediction model...
is then trained, validated, and tested; yielding excellent performance. Despite attaining inferior accuracy, the model developed based on urine-based attributes is recommended for implementation in intelligent diagnostic systems due to the efficient computational requirements. Furthermore, the sample acquisition protocol is more convenient for both patients and medical practitioners.

V. ACKNOWLEDGEMENT

Authors express their appreciation to UniversitiTeknologi MARA and the Ministry of Higher Education, Malaysia for funding this work through the Fundamental Research Grant Scheme (600-RMI/FRGS 5/3 (153/2019).

VI. REFERENCES


[32] Y. Yuan, “Multiple imputation for missing data: Concepts and new


