A NEURAL NETWORK BASED DATA MINING APPROACH FOR RECOGNITION OF CHRONIC KIDNEY DISEASE

RAVINDRA BV¹, NARENDRA VG², SHIVAPRASAD G³

¹Associate Professor, Manipal School of Information Sciences, Manipal Academy of Higher Education, Manipal, India-576104
²Associate Professor Senior Scale, Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India-576104
³Assistant Professor Selection Grade, Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India-576104

E-mail: ¹ravindra.bv@manipal.edu, ²narendra.vg@manipal.edu, ³shiva.prasad@manipal.edu

ABSTRACT

Chronic kidney failure occurs when the regular kidney filtration functionalities fail which leads to accumulation of electrolytes, wastes and other fluids in the body. One has to go appropriate dialysis procedure for their survival. It is very critical to recognize the level of chronic kidney disease (CKD) for the nephrologist and further dialysis period cannot be predicted appropriately for individuals. Data mining approaches have shown a promising path in the last decade to develop automated decision making tool for clinical diagnosis. This specific research suggests the application of neural network as critical qualitative indicator to mine the kidney dialysis attributes for classification of CKD from non-chronic kidney disease (NCKD). Two datasets one from open source UCI machine learning repository CKD database and other local hospital were considered for this study. Initially clustering was applied to remove the inconsistency from the datasets. Numerical and nominal normalized data was employed to multilayer perceptron neural network (MLPNN) to perform the classification of CKD and NCKD. MLPNN was configured optimally by appropriate network parameters and was evaluated in terms of Specificity, Sensitivity and classification accuracy. Further other classifier performance metrics, such as, position and negative predictions, error rate, F-Score, MCC and Kappa test were also evaluated. Experimental simulation shows that the proposed pattern classifier yields a classification accuracy of 93.22% and 92.78% respectively for the two different data sets.

Keywords: Chronic Kidney Disease, Non-Chronic Kidney Disease, Multilayer Perceptron Neural Network, Classification

1. INTRODUCTION

Kidney health is a top priority around the world. This reflects the kidneys' critical role in maintaining fluid and electrolyte balance as well as removing waste (including medicine processing), releasing hormones to control blood pressure (BP) and stimulate red blood cell production (and thus reduce the risk of cardiovascular disease and anemia), and activating vitamin D to maintain bone health [1].

Chronic kidney disease (CKD) is a serious and widespread non-communicable disease that affects people all over the world. CKD is defined and staged according to national and international recommendations based on measures of kidney function that allow for a degree of risk stratification using generally available data in its early stages, and early detection is critical to reducing future risk. For most people with CKD, the risk of cardiovascular consequences is greater than the chance of progression to end-stage kidney disease [1].

The demand for non-invasive, early, real-time, and pain-free alternative medical diagnostics of the human kidney or renal has increased in recent years, and some reasons for this include its non-invasive, early, real-time, and pain-free mechanism. One of the major kidney issues that requires early identification is chronic renal disease [2].

Chronic kidney disease (CKD) recognition and classification is a difficult research topic for the biomedical engineering and clinical communities [8] [15] [27]. Typical blood and urine tests reveal the subject's health status, and the severity of CKD leads...
to additional symptoms such as anemia, bone
disease, high blood pressure, and so on [10] [36].

In general, kidney dialysis is conducted owing to
renal failure, and depending on the severity of the
clinical situation, either hemodialysis, peritoneal
dialysis, or full transformation is required. The
pathological test provides a clear indicator of
whether CKD is present.

The rest of this paper is organized as follows:
Section 2 reviews the literature related to data
mining for kidney disease. In Section 3, we present
the materials and methods. In Section 4, results and
discussion. Finally, Section 5 presents the
conclusions and future work.

2. RELATED WORK

Peritoneal dialysis (PD) was first developed in
India in 1962, followed by hemodialysis (HD), and
then transplantation. Currently, over 1,30,000 people
are undergoing dialysis, with the number of patients
growing by approximately 232 cases per one million
people in the population. Only a few patients with
severe CKD were given hepatitis B immunization
[3], and only a few of those who received the vaccine
had protection against hepatitis B surface antibodies
[4]. Those who were referred late to therapy had a
higher risk of suffering from anemia, having a bad
prognosis, beginning dialysis without an
arteriovenous fistula, having a lower likelihood of
receiving Hepatitis B vaccination, and having a
higher overall death rate [5]. Protein energy waste is
observed in 68%–93% of dialysis patients from low
and moderate socioeconomic groups.

HD is used to filter water and waste from the
blood in the same way that the kidneys did when they
were healthy. HD helps to keep blood pressure under
control and balance minerals like calcium,
kidney dialysis patients has been reported [8]. Two
data mining approaches were used, and the
classification accuracy was reported to be 97.78
percent. The same research group suggested an
improved version of the data mining approach [38].
For the proposed study, a database of 188 patients
was developed based on 707 visits, with two
categories: "living" and "diseased," to determine the
patient's survival span. On the basis of two data
mining techniques, the rough set theory and decision
tree, a three-day dialysis time was explored. The
complete database was separated into eight trails,
each of which was classified. The rules were created
using lower and upper approximations. The survival
rate may be predicted using sixteen distinct
classifiers based on data mining algorithms. Overall,
a classifier accuracy range of 75% to 85% was
achieved. The researchers came to the conclusion
that the renal dialysis parameters chosen have a
major influence in predicting survival. A decision
tree approach was reported for Bacille Calmette-
Guerin (BCG) naive and prior BCG linked
subpopulations.

In [22], used an integrated intelligent fuzzy
expert system to analyze the evolution of renal
failure in chronic kidney disease. The experimental
investigation made use of clinical data spanning ten
years and involving ten vital indicators. In this study,
the effect of GFR was accurately predicted, with a
normalized mean absolute error of 4.88 percent. The
findings of the study were promising, and they
should be carried forward into clinical practice.

In [29], proposed a comparison analysis report
for predicting CKD at an early stage. Three models
were used: multilayer perceptron neural network,
radius basis neural network, and logistic regression,
with characteristics like F-score, kappa, and
accuracy evaluated. For the experimental study, an
open source UCI [20] machine learning repository
was employed.

There have been several attempts to distinguish
between chronic kidney disease (CKD) and non-
CKD [21] [27] [29] [37]. CKD and NCKD were
predicted and classified by [13], who used a wrapper
subset attribute evaluator and best first search
to distinguish between the two conditions.

In order to learn more about kidney diseases and
how they are classified, you can look at the reports
by [9] [18] [19]. We looked at the age-based
classification of CKD in [16]. In [23], have written
about clinical practice guidelines for CKD
classification rules that can help people with CKD
get better. A method for predicting the survival of
kidney dialysis patients has been reported [8]. Two
data mining approaches were used, and the
classification accuracy was reported to be 97.78
percent. The same research group suggested an
improved version of the data mining approach [38].
For the proposed study, a database of 188 patients
was developed based on 707 visits, with two
categories: "living" and "diseased," to determine the
patient's survival span. On the basis of two data
mining techniques, the rough set theory and decision
tree, a three-day dialysis time was explored. The
complete database was separated into eight trails,
each of which was classified. The rules were created
using lower and upper approximations. The survival
rate may be predicted using sixteen distinct
classifiers based on data mining algorithms. Overall,
a classifier accuracy range of 75% to 85% was
achieved. The researchers came to the conclusion
that the renal dialysis parameters chosen have a
major influence in predicting survival. A decision
tree approach was reported for Bacille Calmette-
Guerin (BCG) naive and prior BCG linked
subpopulations.
ultrasound imaging for stage categorization in chronic renal disease, In [17], have demonstrated that it is possible. This was the first attempt to calculate the creatinine index without using the creatinine index. A random forest data mining technique for the diagnosis of chronic kidney disease (CKD) has been presented [35].

Banarjee, Noor, and colleagues (2019) [10] attempted to predict chronic kidney disease (CKD) using Random Forest, Naive Bayes, Support Vector Machine (SVM), and potassium level data, among other methods. The study used a dataset that had 16 parameters, including the name of the item, the total amount of fat, the salt level, the number of calories, and the potassium level. Accuracy and potassium levels were used to evaluate the performance of the machine. It was possible to achieve an overall accuracy of 99.75 percent by using a dietary system that provided a meal chart that assisted in maintaining the health of CKD patients even further by maintaining the salt level.

The use of a data mining tool has been chosen as the quantitative technique for the automated categorization of chronic kidney disease from non-chronic kidney disease in this study (NCKD). In the pursuit of better prediction and classification of kidney disorders, several approaches have been tried.

This paper describes the classification of CKD and NCKD using multilayer precision neural network (MPLNN). Two databases, one from open source UCI repository CKD database and another from local hospital were considered for the study.

3. MATERIALS AND METHODS

3.1 Study Dataset

Two databases are used in the proposed research. The first database (DB1) contains kidney dialysis data obtained after ethical clearance from a local general hospital. [2] [17] [19] [20]. The suggested study omitted attributes that were judged to be negligible. The exact details are shown in Table 1.

<table>
<thead>
<tr>
<th>No</th>
<th>Attribute</th>
<th>Type</th>
<th>Units or Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ID NO.</td>
<td>Numerical</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>DOB</td>
<td>Numerical</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Gender</td>
<td>Nominal</td>
<td>Male/Female</td>
</tr>
<tr>
<td>4</td>
<td>Race</td>
<td>Nominal</td>
<td>Chinese/Indian/Malay/Iban</td>
</tr>
<tr>
<td>5</td>
<td>Blood Type</td>
<td>Nominal</td>
<td>A-B-/-0-</td>
</tr>
<tr>
<td>6</td>
<td>Weight Pre</td>
<td>Numerical</td>
<td></td>
</tr>
</tbody>
</table>

Only n = 180 of the 230 patient pathology data (both nominal and numerical values) recorded over a two-year period were judged to be significant. The baseline characteristics of the research population (n = 180) are shown in Table 2. For the sake of brevity, the mean ± SD is presented for each variable.

Because of insufficient data, only four qualities were taken into consideration: creatinine, urea, sodium, and potassium. Patients/subjects were classified as having chronic kidney disease (CKD) or not having chronic kidney disease (NCKD) based on the priority region (low, medium, or high) in which each feature was found, as well as the length of time they had spent in the dialysis centre. Table 3
displays the data that has been labelled. Training with MLPNN was carried out using male data, female data, and male-female data in accordance with the four attributes identified. The three groups of patients (male, female, and male-female) that were considered for the local database were as follows: The University of California, Irvine Machine Learning Repository (DB2) [14] [30] is used for the second study, which makes use of open source data from that repository. There were a total of n=400 subjects whose data was saved. There were a total of 25 quality options available to choose from. Table 4 lists the characteristics of DB2 at its most basic level. Table 5 displays the data that has been labelled.

Table 2. Baseline characteristics of DB1 (n=180)

<table>
<thead>
<tr>
<th>Variable</th>
<th>CKD</th>
<th>NCKD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex (Male=1, Female=0)</td>
<td>0.544828±0.499713</td>
<td>0.428571±0.502096445</td>
</tr>
<tr>
<td>Age (Years)</td>
<td>42.48966±8.589494</td>
<td>43.25714±9.586580256</td>
</tr>
<tr>
<td>Creatinine (mgs/dl)</td>
<td>4.13931±0.947606</td>
<td>6.354826±0.245360308</td>
</tr>
<tr>
<td>Urea (mgs/dl)</td>
<td>25.15862±10.18078</td>
<td>25.25714±9.974505316</td>
</tr>
<tr>
<td>Sodium (mEq/L)</td>
<td>4.66±0.562534</td>
<td>4.162857±0.842295782</td>
</tr>
<tr>
<td>Months (Numeric)</td>
<td>18.13793±9.862683</td>
<td>0 ± 0</td>
</tr>
</tbody>
</table>

Table 3. Labelled data (DB1)

<table>
<thead>
<tr>
<th>n</th>
<th>CKD</th>
<th>NCKD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKD</td>
<td>145</td>
<td></td>
</tr>
<tr>
<td>NCKD</td>
<td>35</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Baseline characteristics of DB2 (n=400)

<table>
<thead>
<tr>
<th>Variable</th>
<th>CKD</th>
<th>NCKD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>54.54132±17.70215</td>
<td>46.51678±16.0349</td>
</tr>
<tr>
<td>Blood Pressure</td>
<td>79.625±16.04422</td>
<td>75.7222±20.54956</td>
</tr>
<tr>
<td>Specific Gravity</td>
<td>1.013918±0.070286</td>
<td>1.022414±0.084652</td>
</tr>
<tr>
<td>Albumin</td>
<td>1.722488±1.374507</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>Sugar</td>
<td>0.76699±1.346338</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>Blood Glucose</td>
<td>175.4198±92.64777</td>
<td>107.7222±20.54956</td>
</tr>
<tr>
<td>Blood Urea</td>
<td>27.3800±58.6515</td>
<td>32.79861±11.73122</td>
</tr>
<tr>
<td>Serum Creatinine</td>
<td>4.414916±6.941535</td>
<td>0.868986±0.264187</td>
</tr>
<tr>
<td>Sodium</td>
<td>133.9018±16.0937</td>
<td>141.731±12.67433</td>
</tr>
</tbody>
</table>

In order to make efficient training of the network model, four cases were considered for the DB2. Attributes which were closely correlated were grouped together and thereby multi parameters were employed for training the MLPNN model. Based on this, four cases were considered (Table 6)

Table 5. Labelled Data of DB2

<table>
<thead>
<tr>
<th>n</th>
<th>CKD</th>
<th>NCKD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKD</td>
<td>250</td>
<td></td>
</tr>
<tr>
<td>NCKD</td>
<td>150</td>
<td></td>
</tr>
</tbody>
</table>

3.2 Data Normalization

In order to ensure that the applied data suits well for training using the MLPNN, a sample normalization function was applied [11].

Data Normalization

\[
\text{Data Normalization} = \frac{(\text{Actual attribute value} - \text{Lower limit})}{(\text{Upper limit of the attribute} - \text{Lower limit of the attribute} + 1)}
\]

3.3. MLPNN

The multilayer perception neural network (MLPNN) has found widespread use in pattern recognition [13] [32] [33] [35]. This model is used in the majority of clinical decision-making because of its self-tolerance, robustness, and generalization capabilities. Figure 1 depicts the MLPNN model that
will be used in the proposed research. Three-seven inputs were selected for open-source databases, and four nodes were evaluated for local databases.

The MLPNN was trained using a variety of nominal and numerical attribute combinations. Using a trial and error process, the network model was optimized in terms of hidden neurons, activation function, learning algorithm, learning rate, and learning momentum. The models are assessed using performance criteria such as sensitivity, specificity, and classification accuracy. In terms of network architecture, kernel selection, and decision rules formulated to distinguish CKD from NCKD, the efficiency of the data mining framework is based on supervised models.

The performance of the MLPNN classifier is evaluated by estimating the parameters sensitivity, specificity and overall classification accuracy. Figure 2 shows the proposed framework.

The network model was optimally configured to obtain better classification [24] [30] [31]. The performance of the proposed framework is evaluated using the following parameters:

\[
\text{Sensitivity} \% = \frac{TP}{TP + TN} \quad (1)
\]

\[
\text{Specificity} \% = \frac{TN}{FP + TN} \quad (2)
\]

\[
\text{Positive Precision} \% = \frac{FP}{TP + FP} \quad (3)
\]

\[
\text{Negative Precision} \% = \frac{FN}{TN + FN} \quad (4)
\]

\[
\text{Accuracy} \% = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (5)
\]

\[
\text{Error Rate} \% = \frac{(FP + FN)}{(TP + FP + TN + FN)} \quad (6)
\]

\[
F - \text{Score} = \frac{(2 \ast TP)}{(2 \ast TP + FP + TN)} \quad (7)
\]

\[
\text{MCC} = \frac{\text{SQR} \left ( (TP + T) + (TN + F) \right )}{\text{SQR} \left ( (TP + F) + (F + TN) \right )} \quad (8)
\]

\[
\text{Kappa Test} = \frac{\text{Expected Agreement} - \text{Observed Agreement}}{100} \quad (9)
\]

Where

\[
\text{Observed Agreement} = \% \text{(Overall Accuracy)}
\]

\[
\text{Expected Agreement} = \frac{(TP + FN) \ast (TP + FN) \ast (TN + FN) \ast (FP + TN)}{(TP + FP) \ast (F + FN)}
\]

\[
\text{TP} = \text{Correctly recognized as CKD attributes}
\]

\[
\text{FN} = \text{Incorrectly recognized as NCKD attributes}
\]

\[
\text{TN} = \text{Correctly recognized as NCKD attributes}
\]

\[
\text{FP} = \text{Incorrectly recognized as CKD attributes}
\]

4. RESULT AND DISCUSSION

The kidney dialysis attributes that were recorded were normalized using [27]. Then, the MLPNN model was trained with the results of this process. For this proposed study, 60% and 40% of the training and testing models were chosen, respectively. To make sure the MLPNN model was set up correctly, different work parameters were looked at, and the one shown in Table 7 was used. The MLPNN was trained in a way that was supervised.

During the simulation, it was discovered that a learning rate of 0.8 in conjunction with a momentum of 0.6 produces consistent outcomes. The maximum number of epochs was set at 1000, and the Mean Square Error (MSE) criterion was employed to determine whether the network had reached its convergence. The performance of the network throughout the training phase is depicted in Figures 3a and 3b. As shown in Table 7, the MSE plots for both datasets were created using the Levenberg-Marquardt (LM) training algorithm.

<table>
<thead>
<tr>
<th>Neural Network Parameter</th>
<th>Values Selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hidden Neuron</td>
<td>HN = 1,5,10</td>
</tr>
<tr>
<td>Learning / Training algorithm</td>
<td>LM</td>
</tr>
<tr>
<td>Activation Function: Input layer hidden -output</td>
<td>Log sigmoid</td>
</tr>
<tr>
<td>Stopping criteria: MSE</td>
<td>0.01</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>0.5-0.8</td>
</tr>
<tr>
<td>Learning momentum</td>
<td>0.6</td>
</tr>
</tbody>
</table>

It can be observed from the Figures 3a and 3b that the convergence was found to be good with single
hidden neurons. As the iteration increases, the network reach to yield better convergence.

Figure 3a: Performance of the MLPNN (DB1) during training phase

Figure 3b: Performance of the MLPNN (DB2) during training phase

Figures 4 and 5 show the MLPNN classification performance using the test samples (training mode, HN=1 was found to be obtained for all training algorithms). For better brevity, results were shown by considering all learning/training algorithm during the training phase.

According to the results of the proposed study, the MLPNN produces good classification accuracy when the LM training strategy is applied. This can be seen in Figures 4 and 5. When using the open source dataset, the accuracy is 92.78 percent in all situations, but the local data set has an accuracy of 93.22 percent in all cases. Tables 8, 9, 10, and 11 demonstrate the overall performance measures for the classifiers based on the LM algorithm.

Table 8. Performance measures for DB1

<table>
<thead>
<tr>
<th>Case</th>
<th>SE (%)</th>
<th>SP (%)</th>
<th>AC (%)</th>
<th>PP (%)</th>
<th>NP (%)</th>
<th>ER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>88.75</td>
<td>90.00</td>
<td>89.00</td>
<td>02.73</td>
<td>33.33</td>
<td>11.00</td>
</tr>
<tr>
<td>2</td>
<td>95.38</td>
<td>100.00</td>
<td>96.25</td>
<td>∞</td>
<td>16.66</td>
<td>03.75</td>
</tr>
<tr>
<td>3</td>
<td>95.17</td>
<td>91.42</td>
<td>94.40</td>
<td>02.12</td>
<td>17.94</td>
<td>05.55</td>
</tr>
</tbody>
</table>

Table 9. Performance measures for DB2

<table>
<thead>
<tr>
<th>Case</th>
<th>SE (%)</th>
<th>SP (%)</th>
<th>AC (%)</th>
<th>PP (%)</th>
<th>NP (%)</th>
<th>ER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>95.00</td>
<td>92.00</td>
<td>93.84</td>
<td>05.00</td>
<td>08.00</td>
<td>06.15</td>
</tr>
<tr>
<td>2</td>
<td>85.60</td>
<td>93.33</td>
<td>88.50</td>
<td>04.46</td>
<td>20.45</td>
<td>11.50</td>
</tr>
<tr>
<td>3</td>
<td>92.40</td>
<td>93.33</td>
<td>92.75</td>
<td>04.14</td>
<td>11.94</td>
<td>07.25</td>
</tr>
<tr>
<td>4</td>
<td>96.40</td>
<td>95.33</td>
<td>96.00</td>
<td>02.82</td>
<td>05.92</td>
<td>04.00</td>
</tr>
<tr>
<td>Average</td>
<td>92.35</td>
<td>93.50</td>
<td>92.78</td>
<td>04.10</td>
<td>11.58</td>
<td>07.22</td>
</tr>
</tbody>
</table>

Table 10. Fidelity measures for DB1

<table>
<thead>
<tr>
<th>Case</th>
<th>F Score</th>
<th>MCC</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.93</td>
<td>0.28</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>0.98</td>
<td>0.56</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>0.96</td>
<td>0.347</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Table 11. Fidelity measure for DB2

<table>
<thead>
<tr>
<th>Case</th>
<th>F Score</th>
<th>MCC</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.967611</td>
<td>0.286408</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>0.98008</td>
<td>0.479818</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>0.90101</td>
<td>0.156854</td>
<td>1.00</td>
</tr>
<tr>
<td>4</td>
<td>0.976</td>
<td>0.39</td>
<td>1.00</td>
</tr>
</tbody>
</table>

For the DB1 it can be inferred that the proposed scheme yields high classification accuracy compared to the earlier methods reported in the literature. Fig 6 shows the ROC analysis of the MLPNN classifier. The area under the curve confirms the suitability of the proposed classifier for CKD-NCKD.

5. CONCLUSIONS

It is proposed that a neural network-based data mining tool be used for the classification of chronic and non-chronic kidney disease in the present study. To conduct the research, two separate databases were evaluated, and the renal dialysis properties were treated as a multi-feature for the purpose of training the neural network model. For the classification, a feed-forward multilayer sensory neural network was taken into consideration. In terms of hidden neuron activation function and learning algorithm, the network parameters were configured in the most optimal way. The classification accuracy was used to evaluate the performance of the suggested scheme in terms of learning rate. The results of the simulation demonstrate that the classification accuracy for datasets 1 and 2 was 93.22 percent and 92.78 percent, respectively. As evidenced by the area under the curve calculated using ROC analysis, the suggested neural network model for the CKD-NCKD classification problem performs admirably.

REFERENCES


