

# Context dependent bidirectional deep learning and Bayesian gaussian auto-encoder for prediction of kidney disease

Jayashree M<sup>1,2</sup>, Anitha N<sup>3</sup>

<sup>1</sup>East Point College of Engineering and Technology, Visvesvaraya Technological University, Belagavi, India

<sup>2</sup>Department of Information Science and Engineering, CMR Institute of Technology, Bangalore, India

<sup>3</sup>Department of Computer Science, BNM Institute of Technology, Bangalore, India

## Article Info

### Article history:

Received Jul 31, 2024

Revised Feb 26, 2025

Accepted Mar 26, 2025

### Keywords:

Chronic kidney disease

Context dependent

Long short-term memory

Prediction

SoftMax activation

## ABSTRACT

Chronic kidney disease (CKD) has emerged as a significant global health issue, leading to millions of premature deaths annually. Early prediction of CKD is crucial for timely diagnosis and preventive measures. While various deep learning (DL) methods have been introduced for CKD prediction, achieving robust quantification results remains challenging. To address this, we propose the context-dependent bi-directional DL and Bayesian gaussian autoencoder (CDBDP-BGA) method for CKD prediction. This approach utilizes clinical parameters and symptoms from a structured dataset. By incorporating context dependence into the bi-directional long short-term memory (Bi-LSTM) model, CDBDP-BGA efficiently redistributes the representation of information, enhancing its modeling capabilities. Feature selection is optimized using a BGA-based algorithm, which employs the Bayesian gaussian function. The SoftMax activation function classifies CKD into five distinct stages based on estimated-glomerular filtration-rate (eGFR), considering both symptoms (texture and numerical features) and clinical parameters (age, sex, and creatinine). Simulation results using two datasets demonstrate that CDBDP-BGA outperforms conventional methods, achieving 97.4% accuracy without eGFR and 98.7% with eGFR.

This is an open access article under the [CC BY-SA](#) license.



## Corresponding Author:

Jayshree M

Department of Information Science Engineering, East Point College of Engineering and Technology

Visvesvaraya Technological University

Belagavi, Karnataka, India

Email: jaya.rajendra2024@gmail.com

## 1. INTRODUCTION

Chronic kidney disease (CKD) refers to kidney damage resulting from the inability to filter blood properly. The kidney's primary function is to remove excess water and waste from the blood, which are then excreted via urine. Due to a lack of early disease diagnosis, the mortality rate associated with CKD has recently increased. Various methods have been developed to assist doctors in minimizing mortality by employing sophisticated computer-based detection techniques. Early detection of CKD is of utmost importance in the field of research, as the disease frequently presents itself only after substantial kidney damage has taken place. This early detection has the potential to save numerous lives and greatly decrease mortality rates associated with CKD. Saif *et al.* [1] explained, a deep ensemble method was proposed, employing a majority voting function for prediction outcomes, resulting in substantial improvements in prediction accuracy. Despite these improvements in precision, recall, and accuracy, the training time for early detection was not addressed. To tackle this, a pipeline processing electronic health records (EHRs) using recurrent neural network (RNN) was designed to predict CKD progression through distinct stages.

This method, called long-short term-memory (LSTM) RNN or kidney disease progression [2], achieved high prediction accuracy. However, while improvements were observed in the precision-recall curve, the overall performance of binary classification was not analyzed. Due to dataset length and redundancy, many existing approaches produce incorrect predictions, diagnosing individuals with mild CKD symptoms as severe cases and administering inappropriate therapies. To improve prediction accuracy, data mining with self-tuning spectral clustering using K-mode was proposed in [3]. However, real-time deployment challenges arise as patient data continuously updates, with no efficient method to incorporate new data [4]. To address this, a novel self-correcting mechanism for RNN was introduced, resulting in improved receiver operating characteristic (ROC) curve performance. Another study employed an LSTM-RNN focusing on the error factor [5].

Moreover, a review of ensemble techniques for CKD prediction was conducted in [6], and another predictive model for kidney transplant endpoints was presented in [7]. Diabetic kidney disease progression using biomarkers and deep learning (DL) was proposed in [8], and a survey of CKD prediction outcomes along with patient requirements and preferences was conducted in [9]. Zhu *et al.* [10] states, a regression model analyzed temporal trends to reduce CKD incidence. Another study focused on bone disorders and CKD [11], while a prognostic model for CKD and type 2 diabetes was presented in [12]. A review of existing methods and future directions was conducted in [13], and an early CKD prediction model was proposed in [14], showing improved sensitivity and specificity through regression analysis. Given the increasing global significance of CKD as a mortality source, designing a computer-aided diagnostic (CAD) method for automatic CKD diagnosis is essential. Raihan *et al.* [15], the extreme gradient boosting (XGBoost) classifier algorithm accurately and precisely predicted CKD presence. New markers like eGFR were used in [16] for early CKD prediction, analyzing the relationship between patient data vectors and outcomes. Despite improvements in accuracy and precision using deep ensemble and RNN, the training time involving clinical parameters and symptoms in kidney disease prediction remains a challenge. Hence, the contribution of the work are as follows:

- To address the issues which the existing approaches failed, this work presents a context dependent bi-directional DL and Bayesian gaussian autoencoder (CDBDP-BGA) for robust quantification of prediction of kidney disease.
- Context-dependent Bi-LSTM (CD-Bi-LSTM) network is introduced with context dependency factor for textual feature extraction, demonstrating consistent performance improvements against multiple existing methods and lowering training time. CD-Bi-LSTM network is capable of trading off between detection accuracy and training time than deep ensemble method and LSTM-RNN.
- To select optimal numerical features among the essential features of CKD, BGA-based numerical feature selection algorithm is presented.
- The performance of prediction was assessed using various metrics i.e., precision, recall, F-measure, accuracy and training time. Results are discussed by evaluating CDBDP-BGA and comparing state-of-the-art work and using the same datasets.

## 2. LITERATURE SURVEY

CKD is a crucial and comprehensive public health concern. Over the past few years, its high occurrence rate, rate of hospitalization, cost associated with medication and its poor prognosis, has had an extensive influence on patient's quality of life. A deep ensemble method was proposed, employing a majority voting function for prediction outcomes, resulting in substantial improvements in prediction accuracy [1]. Despite these improvements in precision, recall, and accuracy, the training time for early detection was not addressed. To tackle this, a pipeline processing EHRs using RNN was designed to predict CKD progression through distinct stages. This method, called LSTM-RNN or kidney disease progression [2], achieved high prediction accuracy. However, while improvements were observed in the precision-recall curve, the overall performance of binary classification was not analyzed. Kidney prediction was evaluated by utilizing Berden classification to predict the risk of end stage [17]. Employing this classification model resulted in the improvement of ROC. Hybrid technique called Pearson correlation for feature selection and hybrid classifiers was employed in [18] by the improvement of accuracy score in an extensive manner. A review on presence and onset of CKD was presented in [19] with DL. The study [20] and [21], a systematic review for detection and prediction methods in CKD progression was analyzed in detail. A prediction method with quantitative risk representatives for detecting CKD at the earliest stage was presented in [22].

Ensemble learning using boosting techniques was proposed in [23] taking into considerations the clinical parameters for CKD prediction. ensemble learning resulted in the improvement of accuracy and minimized the run time in a significant manner. An in-depth analysis of clinical outcomes in patients with CKD and eGFR was presented in [24]. Yet another time-varying cox model was applied in [25] for analyzing

the occurrence of CKD. Early CKD detection was presented in [26] by combination of parallel categorization algorithm. An in-depth validation of eGFR for CKD is employed using the progression mechanism in [27]. An elaboration review on unsupervised learning of CKD was investigated in [28]. A systematic review on mortality prediction among kidney patient was presented in [29]. Yet another intelligence diagnosis mechanism employing DL classifier was proposed in [30] to focus on the precision and recall aspects. In [31], presented an approach for filling null values in missing data using different machine learning (ML) approaches, where XGBoost provided better results achieving F-score of 97%. Jayashree and Anitha [32], presented an approach for kidney disease detection where various ML approaches were applied. The findings show that the XGBoost provided better results achieving 98.33% accuracy. Motivated by above mentioned works in literature, though the review on CKD detection accuracy aspects were considered, however focus on the training time aspects were limited. Certain reviews despite making a thorough study on considering the training time for disease detection, the precision and accuracy aspects were not measured. To address on these aspects CDBDP-BGA was applied considering clinical parameters. Elaborate description of CDBDP-BGA method is provided in following sections.

### 3. METHOD

The given methodology, presented in Figure 1 depicts the design which was utilized to carry out the experiments. It incorporated data collection made from structured CKD training and unstructured Twitter testing dataset, textual feature extraction, numerical feature selection and finally, classification using clinical parameters relative to the symptoms for prediction of kidney disease and performance evaluation.

As illustrated in Figure 1, the structured CKD training and unstructured Twitter testing datasets were considered as input. the CDBDP-BGA method underwent three stages: first, textual feature extraction was performed employing CD-Bi-LSTM. Second, numerical feature selection was done using BGA. Finally, for prediction of kidney disease, both the textual feature and numerical features were combined and the SoftMax function was applied for the obtained clinical parameters along with the symptoms to measure eGFR for classifying different stages.

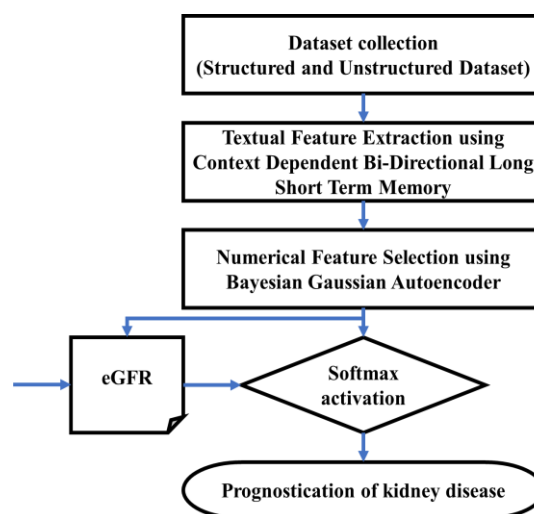


Figure 1. CDBDP-BGA for prediction of kidney disease

#### 3.1. Dataset

The structured training dataset used in our work for prediction of kidney disease at an early stage considering both clinical parameters and set of critical symptoms was taken from <https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease>. CKD dataset was obtained over a year of two-month consisting of 400 different sample instances. From the overall 400 different sample instances, 250 different sample instances were identified to be CKD patients and on the other hand, 150 different sample instances were identified to be healthy participants. Also, each sample instance consisted of 25 attributes based on the measured data via blood test. Here, the first 24 attributes were independent whereas the last one attribute was a dependent attribute and among the overall 24 attributes, 11 attributes are numeric whereas other remaining 14 attributes are categorical. The unstructured data was acquired from health-news Twitter

dataset from <https://archive.ics.uci.edu/dataset/438/health+news+in+twitter>. This health news in Twitter dataset consisted of health news acquired from 15 different major health news agencies. The dataset consisted of 58,000 instances. The structured CKD dataset was used for training and unstructured health-news dataset was used for testing. By employing structured CKD training dataset and an unstructured health news testing dataset, the proposed CDBDP-BGA for the prediction of kidney disease method is designed in the following sections.

### 3.2. Context dependent bi-directional long short-term memory-based keyword extraction

In society, people suffer from a variety of diseases, including diabetes and kidney disease. Among these, kidney disease is considered a global health issue. Risk analysis for kidney disease has been discussed using several methods. Moreover, unstructured health news testing datasets, often extracted from Twitter, typically contain two main types of information: textual explanations and various physical rules. Keyword or feature extraction from these unstructured datasets is crucial for improving the quality of kidney datasets, ensuring that DL models can efficiently learn patterns and make accurate predictions. RNN [1], with their feedback loops, are particularly suitable for processing sequential data, such as news from 15 major health agencies, and can be trained using back-propagation. However, they face issues like gradient problem when modeling long information sequences. To address these challenges, a model called CD-Bi-LSTM network has been designed to process unstructured health news from Twitter datasets, as presented in Figure 2. This model aims to enhance the accuracy and efficiency of kidney disease risk analysis by effectively handling and extracting relevant information from these unstructured data sources.

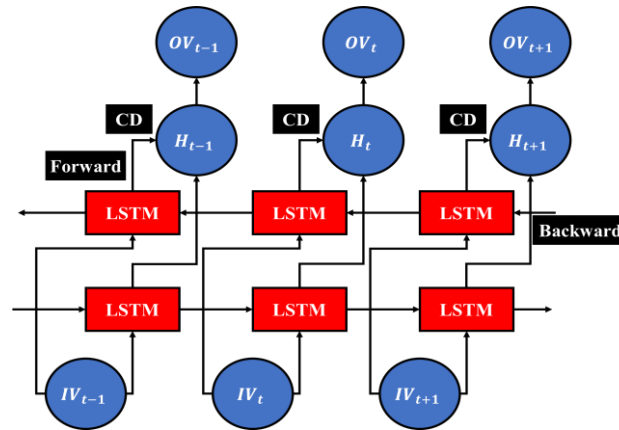


Figure 2. Block diagram of CD-Bi-LSTM-based feature extraction model

In the proposed work of CD-Bi-LSTM network (hidden state), the Twitter information in this hidden state can be updated by gate structure in a constant manner via context dependency. The proposed CD-Bi-LSTM network is used for processing a sequence of tweet data obtained from 15 major health new agencies. It contains two LSTM layers, one for processing input (i.e., input vector) in the forward direction and the other for processing tweet information (i.e., context information) in the backward direction. The intuition behind this model is that by processing data in both forward and backward directions via context dependency, the model is proficient in comprehending the correlation between sequences (i.e., knowing the previous and succeeding tweets in a Twitter account). With the unstructured health news testing dataset extracted using Twitter dataset, the sample instances comprises of health news obtained from more than 15 major health news agencies, to name a few being BBC and CNN. The sample instances are formulated within input vector matrix as (1).

$$IV = \begin{bmatrix} S_1 T_1 & S_1 T_2 & \dots & S_1 T_M \\ S_2 T_1 & S_2 T_2 & \dots & S_2 T_M \\ \dots & \dots & \dots & \dots \\ S_N T_1 & S_N T_2 & \dots & S_N T_M \end{bmatrix} \quad (1)$$

From (1) the input vector matrix 'IV' is formulated by taking into considerations the sample instances 'S<sub>N</sub>' for the corresponding tweets 'T<sub>M</sub>' obtained from 15 major health news agencies of differing

size and varied news. Under the definite rules, three gate structures decides what the corresponding tweet information is stored, updated or forgotten in the corresponding internal state. The mathematical formula for updating these three gate structures is given in (2) to (6).

$$For_t = \sigma(W_{For}[IV_t, H_{t-1}]B_{For}) \quad (2)$$

$$Inp_t = \sigma(W_{Inp}[IV_t, H_{t-1}] + B_{Inp}) \quad (3)$$

$$Out_t = \sigma(W_{Out}[IV_t, H_{t-1}] + B_{Out}) \quad (4)$$

$$C_t = For_t \cdot C_{t-1} + Inp_t \cdot \tanh(W_C[IV_t, H_{t-1}] + B_C) \quad (5)$$

$$H_t = Out_t \cdot \tanh(C_t) \quad (6)$$

From the (2) to (6) ' $IV_t$ ', ' $H_t$ ', and ' $C_t$ ' represents the input state, hidden state and cell state at time instance ' $t$ ' with trainable weight matrices denoted by ' $W_{For}$ ', ' $W_{Inp}$ ', ' $W_{Out}$ ', and ' $W_C$ ' forget gate, input gate, output gate and cell state in addition to biases for corresponding gates denoted as ' $B_{For}$ ', ' $B_{Inp}$ ', ' $B_{Out}$ ', and ' $B_C$ ' activated by sigmoid function ' $\sigma$ ' respectively. Moreover, the structure of CD-Bi-LSTM network is designed to model the context dependency from the preceding text and the succeeding text. Context-dependent memory results in the improved recall when the context during storage or encoding is similar as the context during retrieval or decoding. To model this, the CD-Bi-LSTM network employing two parallel layers both in forward and backward layers, the hidden unit is formulated as (7) and (8). In (7) and (8), ' $\vec{H}_t$ ' and ' $\overleftarrow{H}_t$ ' represents the output of LSTM in the forward layer and backward layer respectively. Finally, these two outputs of LSTM in the forward layer and backward layer are combined to formulate the overall output (9).

$$\vec{H}_t = \overrightarrow{LSTM}(IV_t, H_{t-1}) \quad (7)$$

$$\overleftarrow{H}_t = \overleftarrow{LSTM}(IV_t, H_{t+1}) \quad (8)$$

$$TOV_t = \vec{H}_t + \overleftarrow{H}_t \quad (9)$$

In (9), the textural feature representations are extracted according to context dependency both from the preceding text and the succeeding text. In this manner, by employing context dependency in Bi-LSTM assists in describing the basis disease symptoms which in turn aids in obtaining useful information behind the texts in an accurate manner. The Algorithm 1 is used for feature extraction. As given in Algorithm 1, using the unstructured tweet dataset, the input vector is subjected to textual feature extraction by contextual dependency in the Bi-LSTM network model.

#### Algorithm 1. CD-Bi-LSTM-based textual feature extraction

Input Unstructured dataset ' $DS$ ', Samples instances ' $DSS=\{S_1, S_2, \dots, S_N\}$ ', Tweets ' $T = \{T_1, T_2, \dots, T_M\}$ '

Output Convergent-Efficient Context-Dependent Feature Extraction ' $TOV_t$ '

Step 1 **Initialize** ' $N$ ', ' $M$ '

Step 2 **Begin**

Step 3 **For** each Unstructured dataset ' $DS$ ' with Samples instances ' $DSS$ '

Step 4 Formulate input vector matrix as given in (1)

Step 5 Formulate forget gate, input gate and output gate as given in (2), (3) and (4).

Step 6 Mathematically formulate cell state and hidden state as given in (5) and (6)

Step 7 Mathematically formulate two parallel layers both in forward and backward layers as given in (7) and (8)

Step 8 Combine the two outputs of LSTM in the forward layer and backward layer to generate context dependent textual feature extraction (i.e., representation) as given in (9)

Step 9 Return textual feature extraction (i.e., representation) ' $TOV_t$ '

Step 10 **End for**

Step 11 **End**

### 3.3. Bayesian gaussian autoencoder-based numerical feature selection

Clinical data often contain numerical features where some values are highly correlated while others are not. Using these values directly can negatively impact task performance. Previous work has demonstrated that using RNN for kidney disease progression [2] can achieve high prediction accuracy. However, this study

did not address dimensionality reduction or the handling of highly correlated numerical features. By learning low-dimensional representations of high-dimensional data, feature selection can retain useful numerical features for predicting kidney disease. Yet, selecting useful numerical features from high-dimensional data remains a challenging task. To address this issue, we employ BGA in this work. The Bayesian gaussian function is introduced in a specialized hidden layer, enhancing precision in selecting non-redundant features. Therefore, we use this BGA-based numerical feature selection model to identify numerical features with highly correlated values. Figure 3 illustrates the structure of the BGA-based numerical feature selection model.

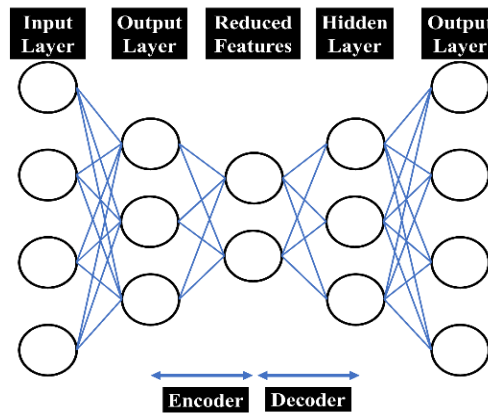


Figure 3. Structure of BGA-based numerical feature selection model

As illustrated in the Figure 3, an autoencoder comprises of two parts, an encoder function ' $f_{\varphi}(IV)$ ' and decoder function ' $f_{\psi}(RF)$ ' respectively, where ' $IV$ ' is the input vector that represents the set of features and ' $RF$ ' denotes the set of reduced features. In addition, an input layer where the input vector ' $IV$ ' forms as the input and in the hidden layer (i.e., two hidden layers employed in our work) the process of encoding and decoding is performed to generate reduced features set (i.e., reduced features selected). Finally, in the decoder side reconstruction is performed with minimal reconstruction loss. To start with the autoencoder is evaluated by how well the decoder reconstructs the data from encoder by means of a loss function using (10).

$$Rec_{loss} = \arg \min_{\varphi, \psi} \left| IV - \left( f_{\psi} \left( f_{\varphi}(IV) \right) \right) \right|^2 \quad (10)$$

From (10), initially the reconstruction loss function ' $Rec_{loss}$ ', is modeled based on ' $\varphi$ ' biases of encoder as well as decoder and the weights ' $\psi$ ' respectively for each tweet in the corresponding Twitter account. Autoencoder in our work maps the numerical values vector ' $Vec$ ' into a hidden representation by means of an encoder function using (11). Followed by which the reconstruction performed by the decoder is mathematically formulated using (12). In (11) and (12), ' $h$ ' denotes the rectified linear unit (ReLU) activation function which selects the most non-redundant numerical features. The challenge now remains in ascertaining the optimal ' $\theta = (\psi, \varphi)$ ', hence, Bayesian gaussian function is used to minimize reconstruction loss ' $Rec_{loss}$ ' using (13).

$$h = h^{(enc)} f_{\varphi}(IV) h(\varphi^{(enc).Vec+1}) \quad (11)$$

$$NOV_t = h^{(dec)} f_{\psi}(RF) h(\psi^{(dec).h^{(enc)} f_{\varphi}(IV)}) \quad (12)$$

$$Prob(IV|\Theta) = \frac{1}{2\pi} \exp \left[ -\frac{1}{2\pi} \sum_{i=1}^N (NOV_i - Exp(IV_t - IV'_t)) \right] \quad (13)$$

In (13), the probability of minimizing reconstruction loss ' $Rec_{loss}$ ' is evaluated by means of output of tweets from specified Twitter account ' $Exp(IV_t - IV'_t)$ ' for ' $IV_t$ ' input vector features represented by ' $t$ ' sample instances. Finally, we obtain a fine-tuned representation ' $h^{(dec)} f_{\psi}(RF)$ ' of discrete numerical values

or reduced numerical features selection with minimal reconstruction loss. The Algorithm 2 is used for improving the precision and accuracy rate of prediction of kidney disease, also, a feature selection algorithm employing BGA is used.

#### Algorithm 2. BGA-based numerical feature selection

Input Unstructured dataset 'DS', Samples instances 'DSS = {S<sub>1</sub>, S<sub>2</sub>, ..., S<sub>N</sub>}', Tweets 'T = {T<sub>1</sub>, T<sub>2</sub>, ..., T<sub>M</sub>}'

Output Reconstruction Loss Minimized Reduced Features Selected

Step 1 **Initialize** 'N', 'M', textual feature extraction (i.e., representation) results 'TOV<sub>t</sub>'

Step 2 **Begin**

Step 3 **For** each Unstructured dataset 'DS' with Samples instances 'DSS' and textual feature extraction (i.e., representation) results 'TOV<sub>t</sub>'

Step 4 **//Input layer**  
Define number of input nodes i.e., from the input vector matrix

Step 5 **//Hidden layer 1 - encoder**  
Formulate reconstruction loss function as given in (10)

Step 6 Formulate encoder function as given in (11)

Step 7 **//Hidden layer 2 - decoder**  
Formulate decoder function as given in (12)

Step 8 Determine optimal 'θ' as given in (13)

Step 9 **//Output layer**  
Return features selected 'NOV<sub>t</sub>' (i.e., reduced features)

Step 10 **End for**

Step 11 **End**

### 3.4. SoftMax activated prediction of kidney disease

Finally, in this section prediction of kidney disease at an early stage by means of clinical parameters with symptoms based on eGFR using SoftMax activation function is designed. To start with the textual feature extraction (i.e., representation) results 'TOV<sub>t</sub>' and numerical features selected 'NOV<sub>t</sub>' (i.e., reduced features) is combined and mathematically represented using (14).

$$h = \begin{bmatrix} TOV_t \\ NOV_t \end{bmatrix} \quad (14)$$

From (14), using ReLU, the combined results is obtained for further prediction of kidney disease. Finally, employing the SoftMax activation function along with the clinical parameters and with the symptoms arrived based on the three distinct features, i.e., age, sex and creatinine, the equations for obtaining five stages based on the eGFR is mathematically formulated (15). From (15), by using numerical features and textual feature symptoms, results along with the clinical parameter values obtained, eGFR prediction of kidney disease at an early stage are said to be made both precisely and accurately. The Algorithm 3 is used for prediction of kidney disease at an early stage, where the textual features and numerical features are combined for classification.

$$\sigma(eGFR)_i = \frac{e^{eGFR_i}}{\sum_{j=1}^K e^{eGFR_j}}, \text{ for } (i = 1, 2, \dots, K) \quad (15)$$

#### Algorithm 3. SoftMax activated prediction for kidney disease

Input Unstructured dataset 'DS', Samples instances 'DSS = {S<sub>1</sub>, S<sub>2</sub>, ..., S<sub>N</sub>}', Tweets 'T = {T<sub>1</sub>, T<sub>2</sub>, ..., T<sub>M</sub>}'

Output Robust Quantification

Step 1 **Initialize** 'N', 'M', textual feature extraction (i.e., representation) results 'TOV<sub>t</sub>', numerical features selected 'NOV<sub>t</sub>' (i.e., reduced features)

Step 2 **Begin**

Step 3 **For** each Unstructured dataset 'DS' with Samples instances 'DSS', textual feature extraction (i.e., representation) results 'TOV<sub>t</sub>' and numerical features selected 'NOV<sub>t</sub>'

Step 4 Formulate rectifier activation function by combining the textual feature extraction and numerical features selected results as given in (13)

Step 5 Formulate SoftMax activation function along with the clinical parameters and with the symptoms as given in (14)

Step 6 For female with 'creatinine < 62 μmol/L', eGFR (mL/min/1.73m<sup>2</sup>) = '144\*(Cr/61.6)<sup>(-0.329)</sup> \* (0.993)<sup>Age</sup>'

```

Step 7      For female with 'creatinine >62 µmol/L', eGFR (mL/min/1.73m2) = '144 * ( $\frac{Cr}{61.6}$ )-1.209 *
(0.993)Age, Formulate decoder function as given in (12)
Step 8      For female with 'creatinine <80 µmol/L', eGFR (mL/min/1.73m2) = '144 * ( $\frac{Cr}{79.2}$ )-0.411 *
(0.993)Age,
Step 9      For female with 'creatinine >80 µmol/L', eGFR (mL/min/1.73m2) = '144 * ( $\frac{Cr}{79.2}$ )-1.209 *
(0.993)Age,
Step 10     If 'eGFR ≥ 90'
Step 11     Then patient is in Stage 1 (i.e., kidney damaged with normal)
Step 12     End if
Step 13     If 'eGFR is between 60 and 89'
Step 14     Then patient is in Stage 2 (i.e., kidney damaged with mildly decreased)
Step 15     End if
Step 16     If 'eGFR is between 30 and 59'
Step 17     Then patient is in Stage 3 (i.e., moderately decreased)
Step 18     End if
Step 19     If 'eGFR is between 15 and 29'
Step 20     Then patient is in Stage 4 (i.e., severely decreased)
Step 21     End if
Step 22     If 'eGFR < 15'
Step 23     Then patient is in Stage 5 (i.e., kidney failure)
Step 24     End if
Step 25     End for
Step 26     End

```

#### 4. RESULTS AND DISCUSSION

For the experimentation of CDBDP-BGA, experimentation was conducted in an Intel Core i5-6200U CPU @ 2.30GHz 4 cores with 4 Gigabytes of DDR4 RAM. Also, the existing deep-ensemble approach [1] and RNN [2] were also experimented on the same platform. All the codes were written in Python. The structured CKD dataset and unstructured health-news Twitter dataset were used for evaluation. Experimental evaluations were conducted considering five performance metrics, precision, recall, F-measure, accuracy and training time. To ensure fair comparisons same structured and unstructured dataset was applied to the three methods, CDBDP-BGA (with and without eGFR), [1], [2] and evaluated for an average of 10 simulation runs.

##### 4.1. Performance analysis of training time

Training time or time consumed in training the samples for prediction of kidney disease with both clinical parameters and symptoms were evaluated using (16). In (16) the training time ' $TT$ ' is measured based on the samples ' $Samples_i$ ' and the time consumed in performing overall prediction of kidney disease is ' $Time (CKD\ diagnosis)$ '. It is measured in terms of milliseconds (ms).

$$TT = \sum_{i=1}^N Samples_i * Time (CKD\ diagnosis) \quad (16)$$

Table 1 lists the tabulation results of training time by substituting the values in (16) for two existing methods, deep ensemble [1] RNN [2] and proposed CDBDP-BGA.  $TT$  was reduced using the proposed CDBDP-BGA method by 29% compared to [1] and 38% compared to [2].

Table 1. Tabulation of training time using proposed CDBDP-BGA method, deep ensemble [1] and RNN [2]

Samples	Training time (ms)		
	CDBDP-BGA	Deep ensemble	RNN
500	125	165	240
1,000	145	200	255
1,500	155	215	270
2,000	168	245	285
2,500	185	280	315
3,000	205	315	330
3,500	225	338	345
4,000	240	355	375
4,500	285	380	390
5,000	315	405	415
500	125	165	240



#### 4.2. Performance analysis of precision, recall, accuracy and F-measure

The performance metrics such as precision and recall were applied to the unstructured sample instances from a sample space. Precision and recall are formulated using (17) and (18) respectively. From the (17) and (18) precision '*Pre*' and recall '*Rec*' are evaluated based on the true positive rate (i.e., diseased patients identified as diseased) '*TP*', false positive rate (i.e., diseased patients identified as normal samples) '*FP*' and the false negative rate (i.e., normal samples identified as diseased patients) '*FN*' respectively. The efficiency of classifier was measured employing the F-measure. The F-measure was mathematically formulated and is presented using (19). In (19) F-measure '*F - measure*', is evaluated by considering the precision '*Pre*' and recall '*Rec*' rate. Finally, accuracy or prediction kidney disease accuracy is evaluated using (20). In (20), accuracy '*Acc*' is measured using the true positive rate (i.e., diseased patients identified as diseased) '*TP*', '*FP*' indicates false positive (i.e., diseased patients identified as normal samples) and the false negative rate (i.e., normal samples identified as diseased patients) '*FN*' and true negative rate (i.e., diseased patients identified as normal samples) '*TN*' respectively.

$$Pre = \frac{TP}{TP+FP} * 100 \quad (17)$$

$$Rec = \frac{TP}{TP+FN} * 100 \quad (18)$$

$$F - measure = 2 * \frac{Pre*Rec}{Pre+Rec} \quad (19)$$

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (20)$$

Figure 4 given above shows the graphical representations of precision, recall, accuracy and F-measure with eGFR by substituting the values in (17) to (20). From the Figure 4 it is inferred that the four-performance metrics, precision, recall, accuracy and F-measure with eGFR using the proposed CDBDP-BGA method is found to be comparatively better than [1] and [2]. Also, with 500 samples provided as input, the true positive rate using the three methods was observed to be 485, 470, and 455. In a similar manner, the false positive rate using the three methods with eGFR was found to be 15, 30, and 45. As a result the overall precision with eGFR was found to be 97%, 94%, and 91%. In a similar manner, the false negative rate using the three methods was found to be 50, 65, and 100, therefore hypothesizing the recall rate to be 97%, 87.5%, and 81.98% respectively. Finally, the accuracy and F-measure with eGFR was found to be 97.4%, 96.8%, 96.4% and 97%, 90.82%, 86.25% respectively. Figure 5 shows pictorial representations of precision, recall, accuracy and F-measure without eGFR by substituting the values in (17) to (20). In figure, four parameters of CDBDP-BGA method are better without eGFR better than [1] and [2]. In a similar manner without eGFR, the precision using the three methods were observed to be 94%, 92%, 90%, the recall rate using the proposed CDBDP-BGA method and existing methods [1] and [2] were found to be 82.45%, 79.31%, and 77.58%. Finally, the prediction kidney disease accuracy without eGFR for the three methods were found to be 98.7%, 98%, 97.1% with an F-measure of 87.84%, 85.18%, and 83.32%.

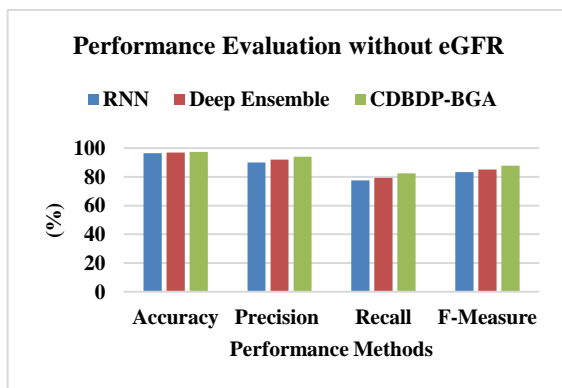


Figure 4. Graphical representations of precision, recall, accuracy and F-measure with eGFR

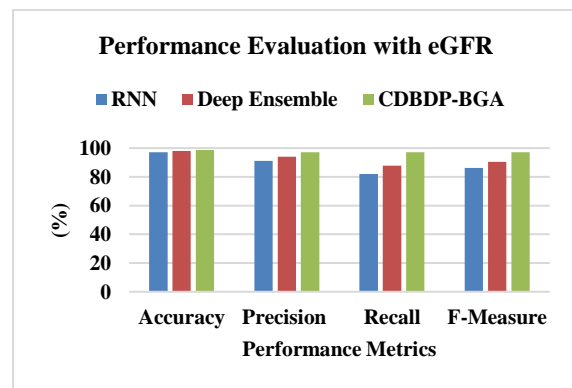


Figure 5. Graphical representations of precision, recall, accuracy and F-measure without eGFR

From the above Figure 4 and Figure 5 two inferences are made. First, four performance evaluation metrics, precision, recall, F-measure and accuracy with eGFR are found to be better than without application of eGFR. Second, four performance evaluation metrics, precision, recall, f-measure and accuracy for prediction of kidney disease are found to be comparatively better using proposed CDBDP-BGA method than [1] and [2]. The reason behind the improvement was due to the application of identifying the textual feature representation and numerical feature selection separately using contextual dependent Bi-LSTM and BGA. Also, both the textual feature representation and numerical feature selected results were applied finally for prediction of kidney disease. In the classification stage, SoftMax activation function along with eGFR and the clinical parameters (i.e., the numerical features selected) were employed for prediction of kidney disease. This in turn finally resulted in the improvement of precision, recall, F-measure and accuracy in a significant manner.

## 5. CONCLUSION

Prediction of kidney disease with both clinical parameters and symptoms pave way for efficiency diagnosis. Hence, desirable work is considered that may assist in analyzing the prediction of kidney illness, therefore reducing the mortality to a greater extent. Past research works underscore prediction of kidney disease employing different conventional and non-conventional methods, to name a few being, ML, DL, and so on. In this work, a CDBDP-BGA for prediction of kidney disease is proposed. The experimentation results validated that the CDBDP-BGA method imparts better results in performance metrics like, precision, recall, f-measure, accuracy and training time compared to the conventional methods. In future, the different preprocessing is utilized to estimate missing data for prediction of kidney disease in early stage with minimum time.

## FUNDING INFORMATION

This work has not been funded by any source.

## AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Jayashree M	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Anitha N												✓		

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

## CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

## DATA AVAILABILITY

- The Chronic Kidney Disease dataset used in this study can be accessed from the UCI Machine Learning Repository from <https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease>.
- The Twitter Health News dataset is available at the following link: <https://archive.ics.uci.edu/dataset/438/health+news+in+twitter>.
- The comparison results presented in this study are based on data from Reference [1] and Reference [2].




## REFERENCES

- [1] D. Saif, A. M. Sarhan, and N. M. Elshennawy, "Deep-kidney: an effective deep learning framework for chronic kidney disease prediction," *Health Information Science and Systems*, vol. 12, no. 1, Dec. 2024, doi: 10.1007/s13755-023-00261-8.
- [2] Y. Zhu, D. Bi, M. Saunders, and Y. Ji, "Prediction of chronic kidney disease progression using recurrent neural network and electronic health records," *Scientific Reports*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-49271-2.




- [3] P. Pradeepa and M. K. Jeyakumar, "Modelling of IDBN with LSNN based optimal feature selection for the prediction of CKD using real time data," *Multimedia Tools and Applications*, vol. 82, no. 4, pp. 6309–6344, Feb. 2023, doi: 10.1007/s11042-022-13561-0.
- [4] H. Du, Z. Pan, K. Y. Ngiam, F. Wang, P. Shum, and M. Feng, "Self-correcting recurrent neural network for acute kidney injury prediction in critical care," *Health Data Science*, vol. 2021, Jan. 2021, doi: 10.34133/2021/9808426.
- [5] L. Men, N. Ilk, X. Tang, and Y. Liu, "Multi-disease prediction using LSTM recurrent neural networks," *Expert Systems with Applications*, vol. 177, p. 114905, Sep. 2021, doi: 10.1016/j.eswa.2021.114905.
- [6] T. M. Alenezi, T. H. Sulaiman, M. Abdelrazek, and A. M. AbdelAziz, "Predictive modeling of kuwaiti chronic kidney diseases (KCKD): leveraging electronic health records for clinical decision-making," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 2, pp. 86–94, 2024, doi: 10.14569/IJACSA.2024.0150211.
- [7] T. Wang, R. G. Qiu, and M. Yu, "Predictive modeling of the progression of alzheimer's disease with recurrent neural networks," *Scientific Reports*, vol. 8, no. 1, Jun. 2018, doi: 10.1038/s41598-018-27337-w.
- [8] S. M. Swaminathan *et al.*, "Novel biomarkers for prognosticating diabetic kidney disease progression," *International Urology and Nephrology*, vol. 55, no. 4, pp. 913–928, Oct. 2023, doi: 10.1007/s11255-022-03354-7.
- [9] D. E. M. van der Horst *et al.*, "Predicting outcomes in chronic kidney disease: needs and preferences of patients and nephrologists," *BMC Nephrology*, vol. 24, no. 1, Mar. 2023, doi: 10.1186/s12882-023-03115-3.
- [10] W. Zhu, M. Han, Y. Wang, and G. Wang, "Trend analysis and prediction of the incidence and mortality of CKD in China and the US," *BMC Nephrology*, vol. 25, no. 1, Mar. 2024, doi: 10.1186/s12882-024-03518-w.
- [11] Y. Li *et al.*, "Machine learning based biomarker discovery for chronic kidney disease–mineral and bone disorder (CKD-MBD)," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, Feb. 2024, doi: 10.1186/s12911-024-02421-6.
- [12] S. A. Saputro *et al.*, "External validation of prognostic models for chronic kidney disease among type 2 diabetes," *Journal of Nephrology*, vol. 35, no. 6, pp. 1637–1653, Jan. 2022, doi: 10.1007/s40620-021-01220-w.
- [13] S. F. Adenwalla, P. O'Halloran, C. Faull, F. E. M. Murtagh, and M. P. M. Graham-Brown, "Advance care planning for patients with end-stage kidney disease on dialysis: narrative review of the current evidence, and future considerations," *Journal of Nephrology*, vol. 37, no. 3, pp. 547–560, Jan. 2024, doi: 10.1007/s40620-023-01841-3.
- [14] J. Zhao *et al.*, "An early prediction model for chronic kidney disease," *Scientific Reports*, vol. 12, no. 1, Feb. 2022, doi: 10.1038/s41598-022-06665-y.
- [15] M. J. Raihan, M. A. M. Khan, S. H. Kee, and A. Al Nahid, "Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP," *Scientific Reports*, vol. 13, no. 1, Apr. 2023, doi: 10.1038/s41598-023-33525-0.
- [16] E. Kanda, B. I. Epureanu, T. Adachi, T. Sasaki, and N. Kashihara, "New marker for chronic kidney disease progression and mortality in medical-word virtual space," *Scientific Reports*, vol. 14, no. 1, Jan. 2024, doi: 10.1038/s41598-024-52235-9.
- [17] R. Takeda *et al.*, "Development of a kidney prognostic score in a Japanese cohort of patients with antineutrophil cytoplasmic autoantibody vasculitis," *Kidney International Reports*, vol. 9, no. 3, pp. 611–623, Mar. 2024, doi: 10.1016/j.ekir.2024.01.007.
- [18] H. Khalid, A. Khan, M. Zahid Khan, G. Mehmood, and M. Shuaib Qureshi, "Machine learning hybrid model for the prediction of chronic kidney disease," *Computational Intelligence and Neuroscience*, vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/9266889.
- [19] Y. Hashimoto, H. Omura, and T. Tanaka, "Presence and onset of chronic kidney disease as a factor involved in the poor prognosis of patients with essential thrombocythemia," *Advances in Hematology*, vol. 2024, no. 1, Jan. 2024, doi: 10.1155/2024/9591497.
- [20] D. K. E. Lim *et al.*, "Prediction models used in the progression of chronic kidney disease: a scoping review," *PLoS ONE*, vol. 17, no. 7 July, p. e0271619, Jul. 2022, doi: 10.1371/journal.pone.0271619.
- [21] F. Chen, P. Kantagowit, T. Nopsopon, A. Chuklin, and K. Pongpirul, "Prediction and diagnosis of chronic kidney disease development and progression using machine-learning: protocol for a systematic review and meta-analysis of reporting standards and model performance," *PLoS ONE*, vol. 18, no. 2 February, p. e0278729, Feb. 2023, doi: 10.1371/journal.pone.0278729.
- [22] S. M. Lee, S. H. Kim, and H. J. Yoon, "Prediction of incident chronic kidney disease in a population with normal renal function and normo-proteinuria," *PLoS ONE*, vol. 18, no. 5 May, p. e0285102, May 2023, doi: 10.1371/journal.pone.0285102.
- [23] S. M. Ganie, P. K. D. Pramanik, S. Mallik, and Z. Zhao, "Chronic kidney disease prediction using boosting techniques based on clinical parameters," *PLoS ONE*, vol. 18, no. 12 December, p. e0295234, Dec. 2023, doi: 10.1371/journal.pone.0295234.
- [24] M. E. Grams *et al.*, "Predicting timing of clinical outcomes in patients with chronic kidney disease and severely decreased glomerular filtration rate," *Kidney International*, vol. 93, no. 6, pp. 1442–1451, Jun. 2018, doi: 10.1016/j.kint.2018.01.009.
- [25] Y. Lin, H. Shao, V. Fonseca, A. H. Anderson, V. Batuman, and L. Shi, "A prediction model on incident chronic kidney disease among individuals with type 2 diabetes in the United States," *Diabetes, Obesity and Metabolism*, vol. 25, no. 10, pp. 2862–2868, Jun. 2023, doi: 10.1111/dom.15177.
- [26] V. Singh and D. Jain, "A hybrid parallel classification model for the diagnosis of chronic kidney disease," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 2, pp. 14–28, 2023, doi: 10.9781/ijimai.2021.10.008.
- [27] M. Gregorich *et al.*, "Development and validation of a prediction model for future estimated glomerular filtration rate in people with type 2 diabetes and chronic kidney disease," *JAMA Network Open*, vol. 6, no. 4, p. E231870, Apr. 2023, doi: 10.1001/jamanetworkopen.2023.1870.
- [28] L. Antony *et al.*, "A comprehensive unsupervised framework for chronic kidney disease prediction," *IEEE Access*, vol. 9, pp. 126481–126501, 2021, doi: 10.1109/ACCESS.2021.3109168.
- [29] P. Hansrivijit *et al.*, "Prediction of mortality among patients with chronic kidney disease: a systematic review," *World Journal of Nephrology*, vol. 10, no. 4, pp. 59–75, Jul. 2021, doi: 10.5527/wjn.v10.i4.59.
- [30] M. Elhoseny, K. Shankar, and J. Uthayakumar, "Intelligent diagnostic prediction and classification system for chronic kidney disease," *Scientific Reports*, vol. 9, no. 1, Jul. 2019, doi: 10.1038/s41598-019-46074-2.
- [31] M. Jayashree and N. Anitha, "Hybrid machine learning based approach to reduce the features for prediction of long-term renal ailment," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 10, no. 4, pp. 48–56, Feb. 2022, doi: 10.1201/9781003589273-79.
- [32] M. Jayashree and N. Anitha, "Feature reduction set for the prediction of renal disease using ensemble methods and optimal hyperplane algorithms," in *Lecture Notes in Networks and Systems*, vol. 922 LNNS, Springer Nature Singapore, 2024, pp. 279–289.

---

**BIOGRAPHIES OF AUTHORS**

**Ms. Jayashree M**    received her bachelor of engineering degree in computer science and engineering from GVIT, BU, Karnataka, India in 1999. She completed her M.Tech from MSRIT, Bengaluru, VTU, Karnataka in 2007. Currently pursuing her Ph.D., from VTU Belagavi, and her research area includes machine learning, artificial intelligence, and natural language processing. She has worked in various reputed colleges in the city and has 20 years of teaching experience. She is a life member of technical bodies like ISTE. She can be contacted at email: jayashree.raju2006@gmail.com.



**Dr. Anitha N**    received her bachelor in engineering in computer science and engineering from the Bangalore University in 2001, her master of engineering in information technology from U.V.C.E (BU) in 2006 and her Ph.D. from the Visvesvaraya Technological University in 2016. Her research interests are machine learning, artificial intelligence, data mining, big data analytics. She can be contacted at email: anitha.mhp@gmail.com.