





Article

Toward Comprehensive Chronic Kidney Disease Prediction Based on Ensemble Deep Learning Models

Deema Mohammed Alsekait ¹, Hager Saleh ^{2,*} , Lubna Abdelkareim Gabralla ¹, Khaled Alnowaiser ^{3,*}, Shaker El-Sappagh ^{4,5} , Radhya Sahal ⁶  and Nora El-Rashidy ⁷ 

¹ Department of Computer Science and Information Technology, Applied College, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia

² Faculty of Computers and Artificial Intelligence, South Valley University, Hurghada 84511, Egypt

³ College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al Kharj 11942, Saudi Arabia

⁴ Faculty of Computer Science and Engineering, Galala University, Suez 435611, Egypt; shaker.elsappagh@gu.edu.eg

⁵ Information Systems Department, Faculty of Computers and Artificial Intelligence, Benha University, Banha 13518, Egypt

⁶ School of Computer Science and Information Technology, University College Cork, T12 R229 Cork, Ireland

⁷ Machine Learning and Information Retrieval Department, Faculty of Artificial Intelligence, Kafrelsheikh University, Kafrelsheikh 13518, Egypt

* Correspondence: hager.saleh@fcih.svu.edu.eg (H.S.); k.alnowaiser@psau.edu.sa (K.A.)

Abstract: Chronic kidney disease (CKD) refers to the gradual decline of kidney function over months or years. Early detection of CKD is crucial and significantly affects a patient's decreasing health progression through several methods, including pharmacological intervention in mild cases or hemodialysis and kidney transportation in severe cases. In the recent past, machine learning (ML) and deep learning (DL) models have become important in the medical diagnosis domain due to their high prediction accuracy. The performance of the developed model mainly depends on choosing the appropriate features and suitable algorithms. Accordingly, the paper aims to introduce a novel ensemble DL approach to detect CKD; multiple methods of feature selection were used to select the optimal selected features. Moreover, we study the effect of the optimal features chosen on CKD from the medical side. The proposed ensemble model integrates pretrained DL models with the support vector machine (SVM) as the metalearner model. Extensive experiments were conducted by using 400 patients from the UCI machine learning repository. The results demonstrate the efficiency of the proposed model in CKD prediction compared to other models. The proposed model with selected features using `mutual_info_classi` obtained the highest performance.

Keywords: chronic kidney disease; machine learning; deep learning; ensemble learning



Citation: Alsekait, D.M.; Saleh, H.; Gabralla, L.A.; Alnowaiser, K.; El-Sappagh, S.; Sahal, R.; El-Rashidy, N. Toward Comprehensive Chronic Kidney Disease Prediction Based on Ensemble Deep Learning Models. *Appl. Sci.* **2023**, *13*, 3937. <https://doi.org/10.3390/app13063937>

Academic Editors: Dimitris Mourtzis and Yu-Dong Zhang

Received: 9 February 2023

Revised: 11 March 2023

Accepted: 14 March 2023

Published: 20 March 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

1.1. Overview

The kidney is the main organ that controls the human body's blood balance and blood pressure; it also plays a role in producing important hormones. Kidney Disease Improving Global Outcomes (KDIGO) defined chronic kidney disease (CKD) as a structural or functional abnormality of the kidney, which persists for more than three months [1].

Several conditions and disorders may lead to kidney diseases, including kidney stones, nephrolithiasis, kidney cyst formation, a rare blood disorder, muscle tissue breakdown, hemolytic uremic syndrome, blood clots, and glomerulonephritis, etc. [2,3].

Many cases of CKD do not present symptoms until the late stages of the disease; this makes estimating the actual prevalence of the disease complicated [4]. However, according to medical statistics, in 2005, CKD caused the death of approximately 38 million out of 57 million cases of CKD. During the COVID-19 period, the mortality rate among COVID-19

patients with CKD was 44.5%, whereas it was 4.5% among non-COVID-19 CKD patients. According to World Health Organization (WHO), by 2050, more than 150 million will have type 2 diabetes, which is considered the leading cause of several kidney disorders [5]. Common kidney abnormalities, such as hydronephrosis, cysts, and stones, are easily avoided and treated in early stages [6]. However, such disorders may lead to CKD and kidney tumors (i.e., cardiorenal syndrome, uremia). CKD is associated with primary adverse outcomes, but cardiovascular disease (CVD) is considered the leading cause of death in this population [7]. Consequently, screening and identification of chronic kidney disease patients during earlier stages can provide interventions that may modify the natural history and reduce the risk of progression to end-stage kidney disease and major cardiovascular events [8].

1.2. Problem Statement

An early diagnosis of such CKD is crucial and can save a patient's life. Medical experts utilized several fundamental approaches to gather precise insights about renal disease identification, such as medical examinations and lab tests (i.e., blood tests and urine tests). The blood test determines the glomerular filtration rate (GFR), which could be used to indicate kidney function. The urine test is used to determine albumin level, indicating whether the kidney is working correctly. With promising data sources that could help in medical diagnosis, developing robust and generalized diagnosis models that could assist medical experts and give accurate and timely decisions is crucial.

Recently, machine learning (ML) has contributed to developing effective models in the medical diagnosis domain, which could make accurate and timely decisions. Deep learning (DL) is a subsection of ML that seeks to find underlying links within a dataset through a set of operations that occurred while training. DL is a multilayer DL model that could theoretically handle nonlinear data; it significantly affects medical applications. For example, Fenglong et al. [9] utilized DL to develop a predictive model that predicts CKD based on several medical examination data. The same statement applies to [10]. Unless DL progresses in various applications, it has several limitations due to the heterogeneity of the medical data, which exacerbates the generalization and robustness of the developed model, resulting in misleading rules and reproducible diagnostic models. Therefore, the training process in DL does not always guarantee achievement of optimal weights and may end with a high-variance model. This challenge could be overcome by using diverse and various DL models. This process is called ensemble learning. Ensemble learning handles single model limitations by combining the advantage of traditional and ensemble learning to give more flexibility and a more generalization model [11]. Ensemble learning has two main types: the base learner and the diversity [12]. First, homogenous learning is achieved by using different data samples. Secondly, heterogeneous learning is achieved by using other models. The many combinations utilized to construct ensemble classifiers include bagging [13], boosting [14], and stacking [15]. In our study, the stacking ensemble model provides a generalized, robust, and flexible model. Several studies demonstrated that ensemble learning results in an accurate and effective model.

Choosing the optimal feature list is the main point in building an efficient model. Feature selection has been intensively investigated in the ML domain, achieving promising results in biomedical applications. Feature selections are classified into three main types: wrapper, filter, and embedded [16]. Our study utilized four feature selection methods to choose the optimal feature list. According to all the above, the main objective of the current paper is to develop an ensemble DL that could improve the prediction performance with the optimal feature subset. From a clinical perspective, our proposed feature list demonstrates the effectiveness in early prediction of CKD compared with the state of the art.

1.3. Paper Contribution

The main motivation for our work is summarized in the following points.

- We explore different feature selection techniques to select the optimal features from an AI perspective.
- We study the prediction of CKD from the medical side, working side by side with a medical expert to choose the most affected features from the medical side perspective.
- We choose the optimal feature subset that is selected from an AI perspective and confirmed medically.
- We develop novel stacking ensemble DL based on LSTM, CNN, and GRU in the base learning layer and SVM in the meta layer.
- We train and test the base-learning models based on the CKD dataset with different feature subsets.
- We compare the proposed ensemble model's performance with other DL models.
- We evaluate models' performance with standard metrics, such as precision, recall, F score, and accuracy.
- We ensure the superiority of our proposed model outcomes through calculating various evaluation metrics, as well as comparing with the state of the art.

1.4. Paper Organization

The rest of the paper is organized as follows. Section 2 describes the related work in the CKD domain. Dataset details and the proposed framework are discussed in Section 3. Section 4 shows the results and discussion. The paper concludes in Section 5.

2. Related Work

CKD is a critical and complex disease that affects a patient's life. Unfortunately, the early symptoms of CKD may be subtle, and several symptoms overlap with other diseases. For example, kidney diseases are usually associated with decreased albumin levels, increased blood pressure, and a high decrease in white blood cells, which may overlap with other diseases such as hypertension, liver diseases, anemia, and heart diseases. Therefore, the existence of an efficient and accurate system that could assist the medical expert in data analysis and prediction is highly required.

This section discusses the state of the art of CKD diagnosis and prediction. We mainly focus on ML and DL models and feature-selection methods. Several studies depend on historical and lab test data to provide promising predictive models. For example, Shahriar et al. [17] provided a predictive model for CKD based on Gaussian naive Bayes (GNB) and decision tree (DT); the authors utilized all features in the used dataset and provided the best performance based on GNB. Wassel et al. [18] used statistical, RF-, and DT-based models to predict the existence of CKD. They reported that RF gives the best accuracy, outperforming other algorithms. In [19], the authors utilized recursive feature elimination (RFE) to choose the feature list, then built an ensemble model of several algorithms, including NB, SVM, multilayer perceptron, and logistic regression (LR). Their proposed model achieved promising performance in terms of different classification metrics. In [20], the authors proposed a neural network (NN) SVM model to predict CKD based on patients' historical and clinical examination data. First, the data was preprocessed by replacing missing values with the column mean. The parameters were then calculated based on numerical analysis and then chose the optimal feature list based on information gain (IG). Their proposed model achieved the best accuracy. In [9], the authors utilized NN to develop a predictive model that predicts CKD based on medical examination, patient characteristics, and family history. Their proposed model achieved promising performance in terms of different evaluation metrics. The same is true in [21]; the only difference is the utilization of pretrained DNNs in building their predictive model. In [22], the authors explored the use of ML in analyzing datasets for CKD and employed LR and feedforward neural networks to achieve promising results. In [23], the authors selected features by using the XGBoost feature-selection algorithm and applied ML models: RF, SVM, RT, NN, and the bagging tree model (BTM) with selected features to predict CKD. In [24], the authors used recursive feature elimination (RFE) with ML models: SVM, KNN, DT, and RF for early

diagnosis to avoid developing kidney failure. The result showed that RF achieved the best performance. In [25], the authors proposed a DL model and compared its performance to ML models: SVM, KNN, LR, RF, and NB. RFE and chi-square were used to select important features from CKD. Their proposed approach achieved the best performance. In [26], the authors used J48 and random forest to predict the various stages of CKD. J48 provided better accuracy compared to RF. In [27], the authors used the chi-squared test to extract highly correlated features with the target. The SMOTE algorithm was used to handle unbalanced data and RF and SVM were applied with selected features. SVM achieved better results than RF with a 10-fold CV. In [28], the authors optimized different models: CC, ANN, and LSTM, which are called OCNN, OANN, and OLSTM, respectively. OCNN achieved the highest validation accuracy. In [29], the authors proposed multilayer perceptron to diagnose CKD and compared it with SVM and NB. Experiments showed that multilayer perceptron achieved the highest accuracy. In [30], the authors used LR, RF, SVM, KNN, and NB and a feedforward neural network. KNN was used to fill in the missing values. RF achieved the best accuracy. In [31], the authors compared ML models NB, MLP, quadratic discriminant analysis (QDA), RF, KNN, and LR by using different evaluating parameters. RF displayed the highest accuracy. In [32], the authors used ML models DT, LR, RF, and KNN to predict CKD disease. In [33], the authors applied LR, DT, and SVM and used the bagging ensemble method; the bagging ensemble obtained the best result.

Other studies depend on temporal electronic health records (EHR) to predict CKD; for example, Ren et al. [34] developed a predictive model for patients based on clinical and laboratory data. They first preprocess the data by using undersampling to maintain the balance of the data, using an autoencoder to encode text data, then building NN model-based bidirectional LSTM and autoencoder. The proposed model achieved 87.9 in terms of classification accuracy with 10-fold cross-validation. In [35], the authors developed a model that could predict the onset of CKD prediction within the next 3, 6, and 12 months based on data aggregated through 24 months. The authors used various ML algorithms, including (RF, LR, DT) and DL algorithms (CNN, Bi-LSTM). The best accuracy was obtained from the CNN model. Other studies use nontemporal features to develop a prediction model. For example, Song et al. [36] depend on nontemporal EHR data in building prediction models. They first utilized three feature-selection techniques to extract features from EHR, then used DNN to build a predictive model that achieved the best AUCROC. The same is true in [37], which builds predictive models for CKD among diabetes patients. The developed model is based on demographic data and lab test data.

3. Methodology

The paper aims to introduce a novel ensemble DL approach to detect CKD; feature-selection methods were used to select the optimal selected features. Medically, the chosen features affect CKD. As shown in Figure 1, the proposed framework has a set of phases: the selected dataset, preprocessing step, feature-selection methods, our proposed model, and optimization methods. Each stage will be described in detail as follows.

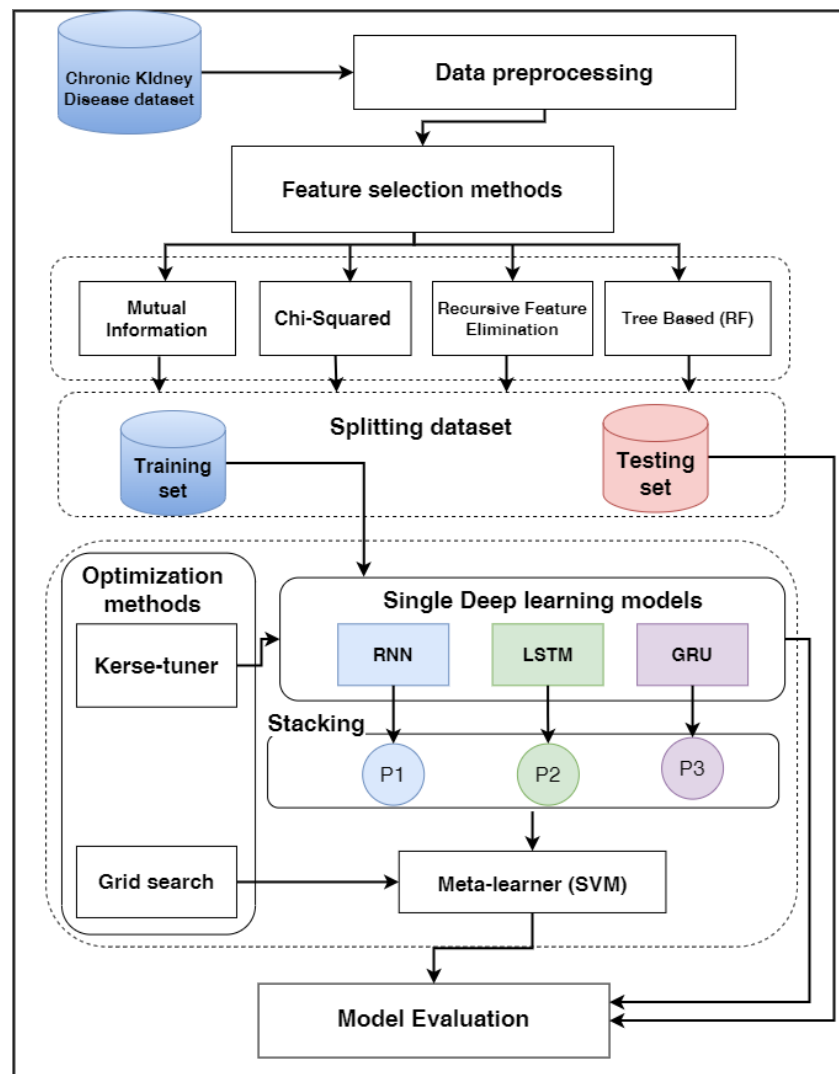


Figure 1. The phases of prediction CKD.

3.1. Dataset Description

The benchmark Chronic Kidney Disease dataset utilized in this study was aggregated from the UCI machine learning repository [38]. Many authors used this dataset to make experimental observations [23,25,39]. It includes data for 400 cases distributed between 150 negative CKD and 250 positive CKD. The utilized dataset consists of 24 features divided into 13 categorical features and 11 numeric features, and one class label has two values: 1 and 0 for positive CKD and negative CKD, respectively. Table 1 details the categorical and numeric features.

Table 1. Dataset description details.

#	Column Name	Abb	D.T	Range	Description
1	Age	Age	N	(2 to 90)	Patient's age in years
2	Blood Pressure	PB	N	(50 to 180)	Patient's blood pressure in mmHG
3	Specific gravity	SG	C	(1.025, 1.020, 1.015, 1.010, 1.005)	The ratio between urine density to water density
4	Albumin	AL	C	(0, 1, 2, 3, 4, 5)	Protein percentage in blood plasma
5	Sugar	SU	C	(0, 1, 2, 3, 4, 5)	The sugar level in blood plasma
6	Red blood cells	RBC	C	(Abnormal, Normal)	Percentage of red blood cells in blood plasma
7	pus cell	PC	C	(Abnormal, Normal)	White blood cells in urine
8	Pus cell clumps	PCC	C	(Abnormal, Normal)	Sign of bacterial infection
9	Bacteria	BA	C	(Present, Not Present)	Sign of bacterial existence in urine
10	Blood glucose random	BGR	N	(22 to 490)	A random test of glucose in the blood in mg/dL
11	Blood urea	BU	N	(1.5 to 391)	Percentage of urea nitrogen in blood plasma
12	Serum creatine	SC	N	(0.4, 76)	Creatine level in patient muscles in mg/dL
13	Sodium	SOD	N	(4.5 to 163)	Sodium mineral level in blood
14	Potassium	POT	N	(2.5 to 47)	Potassium mineral level in blood
15	Hemoglobin	HEMO	N	(3.1 to 17.8)	Red protein that responsible of transport oxygen in the blood
16	Packed cell volume	PCV	N	(9 to 43)	The volume of blood cells in a blood sample
17	White blood cell count	WC	N	(2200 to 4800)	Count of white blood cells in cells/cumm
18	Red blood cell count	RC	N	(2.1 to 8)	Count of red blood cells in millions/cumm
19	Hypertension	HTN	C	(Yes, No)	The condition where there is continuously high pressure in the blood vessels
20	Diabetes mellitus	DM	C	(Yes, No)	Impairment in the body's production or response to insulin, a condition of glucose metabolism that makes it difficult to maintain healthy levels of sugar
21	Coronary artery diseases	CAD	C	(Yes, No)	A common heart condition where the main blood channels feeding the heart, have trouble supplying enough nutrients, oxygen, and blood to the heart muscle
22	Appetite	APPET	C	(Good, Poor)	The desire to eat food
23	Pedal edema	PE	C	(Yes, No)	Swelling of the patient's body due to an injury or inflammation
24	Anemia	ANE	C	(Yes, No)	Insufficient healthy red blood cells to transport appropriate oxygen to the body's tissues
25	Class	Class	C	(CKD, Not CKD)	A positive or negative result in terms of having chronic kidney diseases

3.2. Data Preprocessing

Medical data usually suffer from different problems (i.e., the existence of null values or extreme values (outliers)) due to the sensor, network, and data entry failure. When it comes to building an ML model, dataset problems have a negative effect on ML and DL models. The main objective of this stage is to increase data quality by handling missing values and outliers.

- **Data encoding:** In the utilized dataset, a combination of categorical and numeric features exists. Unfortunately, two feature-selection techniques—ML and DL—perform better with numeric features than categorical ones. Therefore, we utilized the label encoder module in the Scikit-learn library to encode all categorical features.
- **Filling missing values:** Several statistical methods have been developed to deal with missing data, but it mainly depends on how much data is missing and how important the missing feature is [40]. Classic statistical techniques such as mean, maximum, and mode perform well with a low percentage of missing values (5% to 10%), with increasing percentage of missing values (20–50%); other complex techniques like expectation maximization [41] are appropriate. In our study, this is due to the low percentage of missing values. The feature means are used to impute missing values.
- **Removing outliers:** Outliers are values that lie far from the normal range of all feature values. It considers a critical problem in building a robust and generalized model [42]. In our current study, all data is analyzed from a statistical point of view to specify the outliers and then ensure the outlier's values from a medical point of view. All outliers in the utilized data were replaced by feature mean.

3.3. Feature-Selection Methods

Feature selection (FS) is the process of extracting a subset of features from the whole dataset, such that the new feature space is reduced within specific criteria [43]. FS considers a critical step in the feature engineering process as it allows the model to concentrate on the important features and remove the less important features. FS has several advantages in terms of model performance, including the following. (i) FS contributes to increasing the prediction accuracy, (ii) it reduces the model complexity, and (iii) it makes the developed model easier to understand and interpret. FS has three main types, including the filter approach, wrapper approach, and embedded approach [44]. Figure 2 show the main categories of FS. Our work explores three feature selection methods: the chi-square test (Chi2), recursive feature elimination (RFE), and mutual information.

- **The filter approach** tries to rank features based on descriptive and statistical measures. The optimum feature subset selection is based on the ranking of features according to the correlation of each feature with the desired output [45]. The types of filter methods are person correlation, information gain, and mutual information. The chi-squared test compute chi-squared stats between each nonnegative feature and class and calculates the score for each feature. This score is used to select the important features that have the highest score. We used Chi2 [46], a library built in Python. Mutual information (MI) [47] is a nonnegative value that expresses how dependent two random variables are on one another. Higher values indicate greater dependence, and it equals 0 only when two random variables are independent. We used `mutual_info_classif()` [47], a library built in Python.
- **The wrapper approach** mainly depends on the performance of the developed model. The core part of the wrapper approach is the utilized algorithm, which tries to find the optimum feature subset that gives the best performance. Initially, the process starts with a few features, and the performance is evaluated [48,49]. This process iterated with the different number of features until reaching the optimal feature subset. Types of wrapper methods include the forward feature selection, backward feature selection, and recursive feature elimination (RFE). By using RFE feature selection, each feature is ranked according to its score in order to help the model select the best features.

- The embedded approach considers part of the training process, and the feature-selection process is an integral part of the classification model [50]. This approach lies between the filter and wrapper approach, as the feature-selection process is made during the model tuning process. Lasso, ridge, and tree-based algorithms (i.e., decision tree, random forest, etc.) are the common ways of embedding feature selection [51]. We used RF, which uses a mean decrease impurity (Gini index) to estimate a feature’s importance.

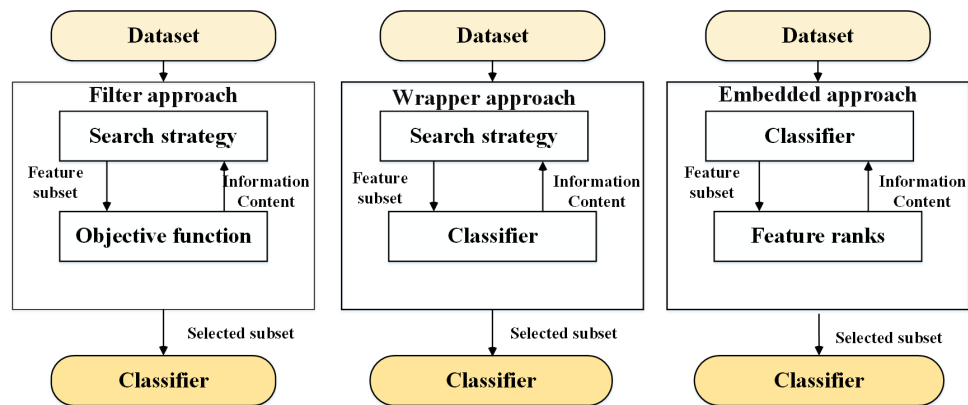


Figure 2. The different types of feature-selection methods.

3.4. The Proposed Model

In this study, we develop a novel prediction model for CKD. The model is divided into two main levels: Level 1 and Level 2, as shown in Figure 3.

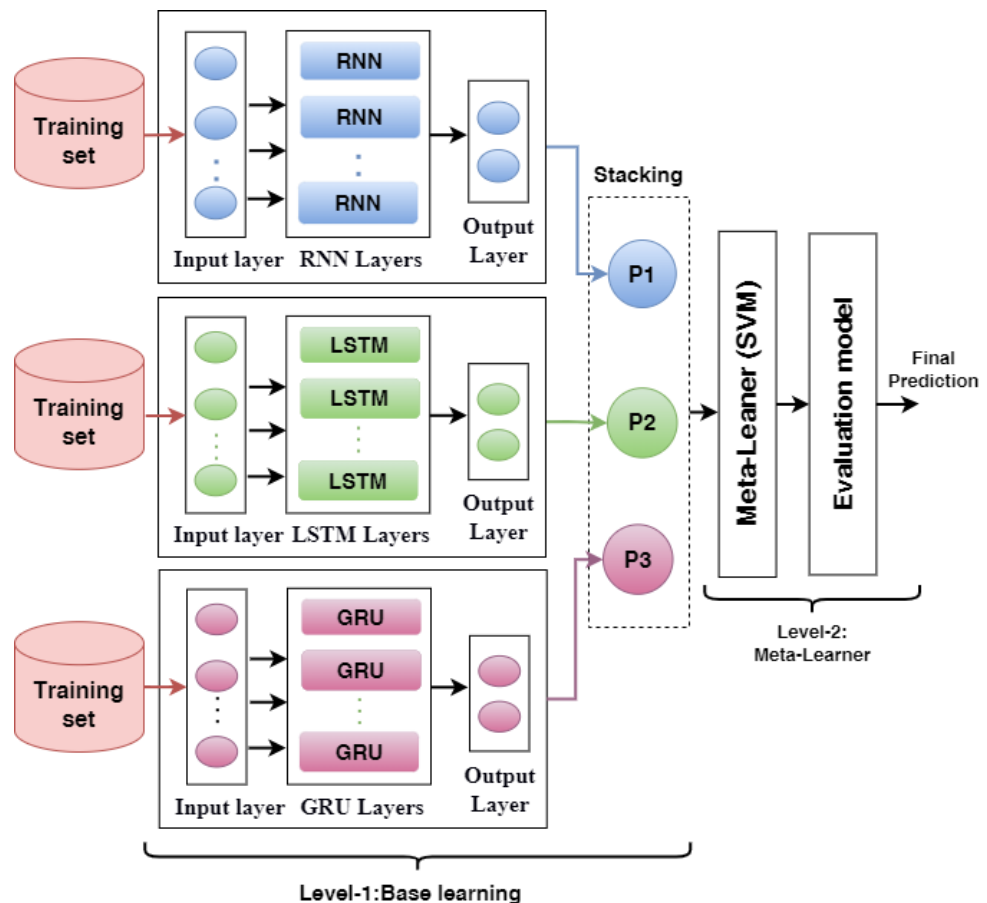


Figure 3. The proposed model of prediction CKD.

- The first level is the base learning, in which three optimized pretrained models are included— recurrent neural network (RNN), gated recurrent unit (GRU), and long short-term memory (LSTM)—with one and two hidden layers. Each model is loaded and frozen in all layers without the final layers. Each output probability prediction of the training set and testing set are combined in stacking training and testing stacking, respectively.
- The second level is called the metalearner level. We fuse the optimized base learner from the previous layer, train the SVM as a metalearner by using training stacking, and evaluate by using testing stacking by exploring the stacking ensemble's role in predicting the final output.

3.4.1. DL Model Architectures

We proposed and optimized three DL models, including the input, hidden, and output layers. The output layer has two neurons that are identical to the classes. The Adam optimizer [52] was employed, and the activation function is sigmoid [53]. Each DL model will be explained as follows.

- The RNN is a type of neural network that is best suited for sequence inputs when used with feedforward networks. The neural network will need to be modulated in order to recognize dependencies between data points in sequence data. RNNs can store previous input states or data to create the subsequent output of the [54].
- LSTM is an attention RNN architecture employed in the field of DL. LSTM has feedback connections. It can analyze complete data sequences in addition to single data points. A memory cell, called a "cell state", which preserves its state over time, plays a crucial role in an LSTM mode. The input gate, forget gate, and output gate are the three gates that regulate the addition and deletion of data from the cell state in the LSTM. New information from the current input that is added to the cell state is controlled by the input gate. The forget gate regulates what data is erased from memory. The output gate conditionally decides what to output from memory [55].
- The gated recurrent unit (GRU) is one of the most common RNN types. GRU uses an identical process as the RNN. GRU creates each recurrent unit to capture dependencies on various time scales. The GRU has gating units that control the information flow within the unit without using specific memory cells [56].

3.4.2. Optimization Methods

We used two optimization methods: Keras-Tuner and grid search with cross-validation to optimize DL and metalearner.

- In DL, Keras-Tuner is a scalable, easy-to-use framework for optimizing hyperparameters that deals with the problems associated with hyperparameter search [57]. For optimized DL models, we adapted the number of neurons in layer1 and layer2 for RNN, LSTM, and GRU: range (20,700).
- Grid search with cross-validation is employed to fine tune hyperparameters of SVM: C: [0.1, 1, 10, 100], gamma: [1, 0.1, 0.01, 0.001], kernel: ['rbf', 'poly', 'sigmoid'].

4. Experiments Results

This section presents the performance of the DL models with different layers and the proposed ensemble models on the CKD dataset by using feature-selection methods.

4.1. Experiment Setup

All models were implemented by using the TensorFlow library along with Keras. DL models were optimized by the Keras tuner. The metalearner classifier was optimized by grid search. All experiments were run using Google Colaboratory. In our work, we conducted two experimental results on the CKD dataset. First, it is split into two sets by using a stratified sampling method—80% training set and 20% testing sets—and secondly,

it is split into a 70% training set and a 30% testing set. We trained and optimized models by using the training set and evaluated models by using the testing set.

4.2. Performance Metrics

Several metrics are commonly utilized in evaluating classification performance, including precision, recall, F measure, and accuracy. All of them are calculated in terms of true positive (TP), false positive (FP), true negative (TN), and false negative (FN). TN showed that the result was negative, and it correctly returned as positive. For TP, the result returned as positive, and it was actually positive. In contrast, TP means that the result returned was positive, and it is actually positive, and the same for TN. We have

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{F1-score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \quad (4)$$

4.3. Training Parameters

For training RNN, LSTM, and GRU with layer1 and layer2, some of the hyperparameters were adapted: 50 epochs with a learning rate of 104, the loss function is binary_crossentropy, and the Adam optimizer was used with a batch size of 100. The values of the hyperparameters optimized by the Kerse tuner are listed below in Table 2.

Table 2. The best values of the hyperparameters of DL models.

Feature-Selection Methods	Models	Split 80:20	Split 70:30
		Number of Units	Number of Units
Chi-Squared	RNN Layer1	[490]	[330]
	RNN Layer2	[470, 90]	[430, 150]
	LSTM Layer1	[190]	[110]
	LSTM Layer2	[250, 230]	[450, 250]
	GRU Layer1	[470]	[310]
	GRU Layer2	[430, 310]	[330, 260]
REF	RNN Layer1	[170]	[430]
	RNN Layer2	[310, 150]	[410, 390]
	LSTM Layer1	[190]	[170]
	LSTM Layer2	[330, 250]	[330, 290]
	GRU Layer1	[290]	[150]
	GRU Layer2	[250, 120]	[370, 250]

Table 2. Cont.

Feature-Selection Methods	Models	Split 80:20	Split 70:30
		Number of Units	Number of Units
mutual_info_classi	RNN Layer1	[490]	[390]
	RNN Layer2	[370, 270]	[290, 230]
	LSTM Layer1	[490]	[450]
	LSTM Layer2	[230, 220]	[150, 150]
	GRU Layer1	[330]	[220]
	GRU Layer2	[250, 140]	[190]
Tree Based	RNN Layer1	[490]	[190]
	RNN Layer2	[370, 450]	[90, 90]
	LSTM Layer1	[290]	[330]
	LSTM Layer2	[310, 250]	[170, 90]
	GRU Layer1	[330]	[270]
	GRU Layer2	[210, 250]	[230, 220]

4.4. Feature-Selection Methods

In the experiments, we explore the essential features of applying feature-selection methods to the CKD dataset.

4.4.1. Features Scores by Chi-Squared

Figure 4 shows the score for each feature after applying chi-squared. We can see that wc has the highest score at 9701.05 and sg has the lowest score at 0.00503. In addition, bu and bgr have the second-best scores, approximately the same score at 2343.0971 and 2241.651, respectively. Moreover, sc, pcv have the third-best scores, approximately the same at 357.79 and 308.18, respectively. Pe, ane, rbc, sod, pcc, cad rc, ba, pot, sg registered the lowest score, below 0.005.

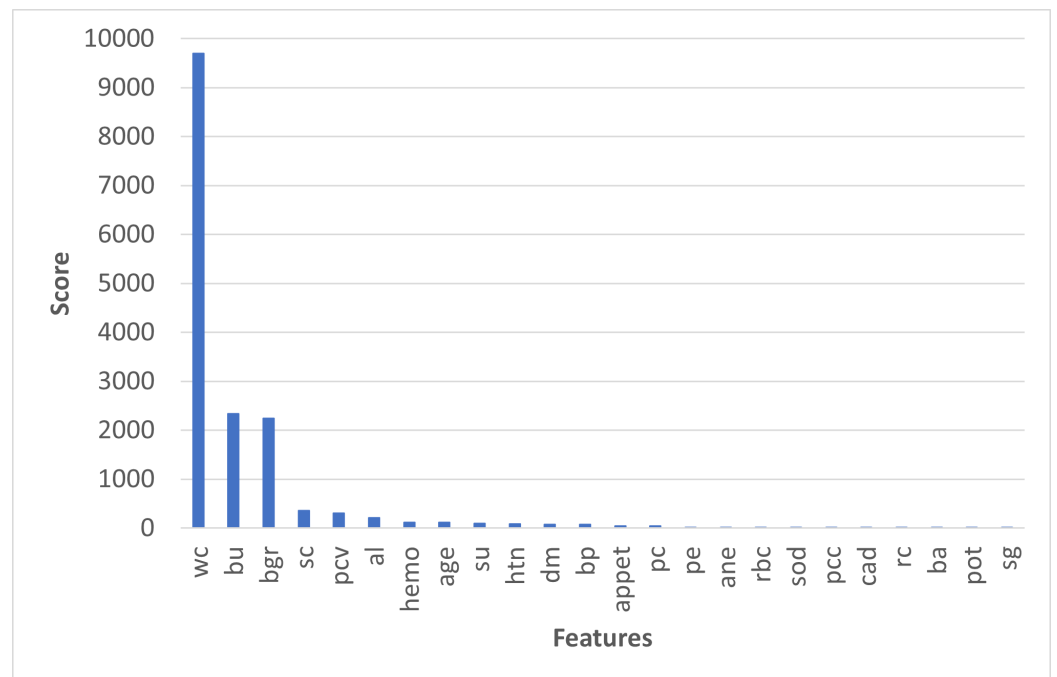


Figure 4. Features scores by chi-squared.

4.4.2. Features Scores by Mutual_Info

Figure 5 shows the score for each feature after applying mutual_info_classes. Hemo has the highest score at 0.460012675, and pcc has the lowest score at 0.035396758. In addition, pcv has the second-highest score at 0.41338565. Moreover, sc and rc have the third-best scores, approximately the same score at 0.379678331 and 0.373807073, respectively. Finally, sg and al have the fourth-best scores, approximately the same at 0.299693288 and 0.296662486, respectively.

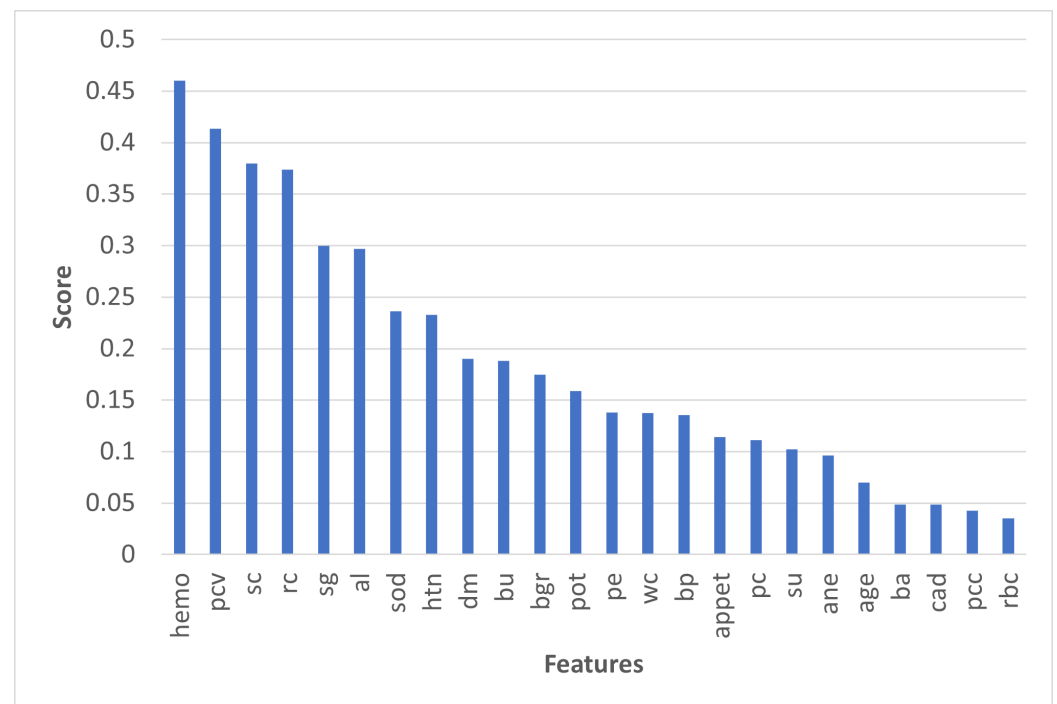


Figure 5. Feature scores by mutual_info_classif.

4.4.3. Features Ranking by RFE

Figure 6 shows the ranking of features selected by REF. REF sets ranking for each feature. Ranking 1 means the best features: age, bp, sg, al, bgr, bu, sc, sod, hemo, pcv, rc, htn, and dm. The worst feature with the highest ranking is ba at 12. The second-worst feature is cad at 11. The third-worst feature, pcc, has 10.

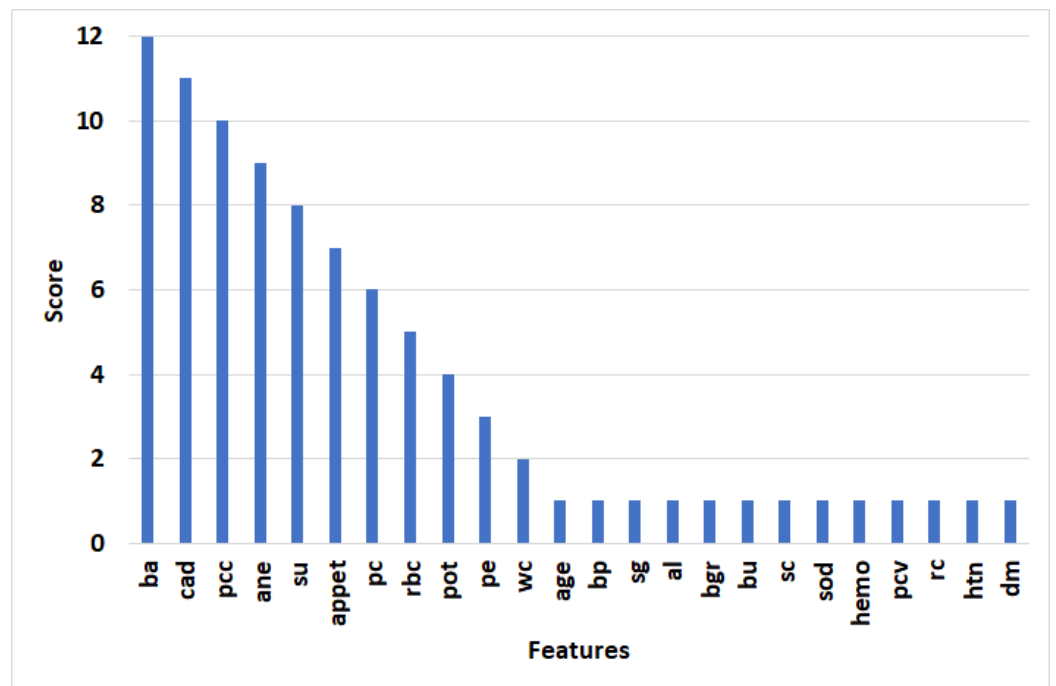


Figure 6. The ranking features selected by RFE.

4.4.4. Feature Importance by Tree_Based (RF)

Figure 7 shows the feature’s importance of Tree_based (RF). Sc has the highest importance at 0.17015. The pcv has the second-highest score at 0.161880. In addition, rc and al have approximately the same importance at 0.08108 and 0.0808, respectively. Finally, ane, pcc, and cad record the lowest importance at 0.000691, 0.000406, and 0.000138.

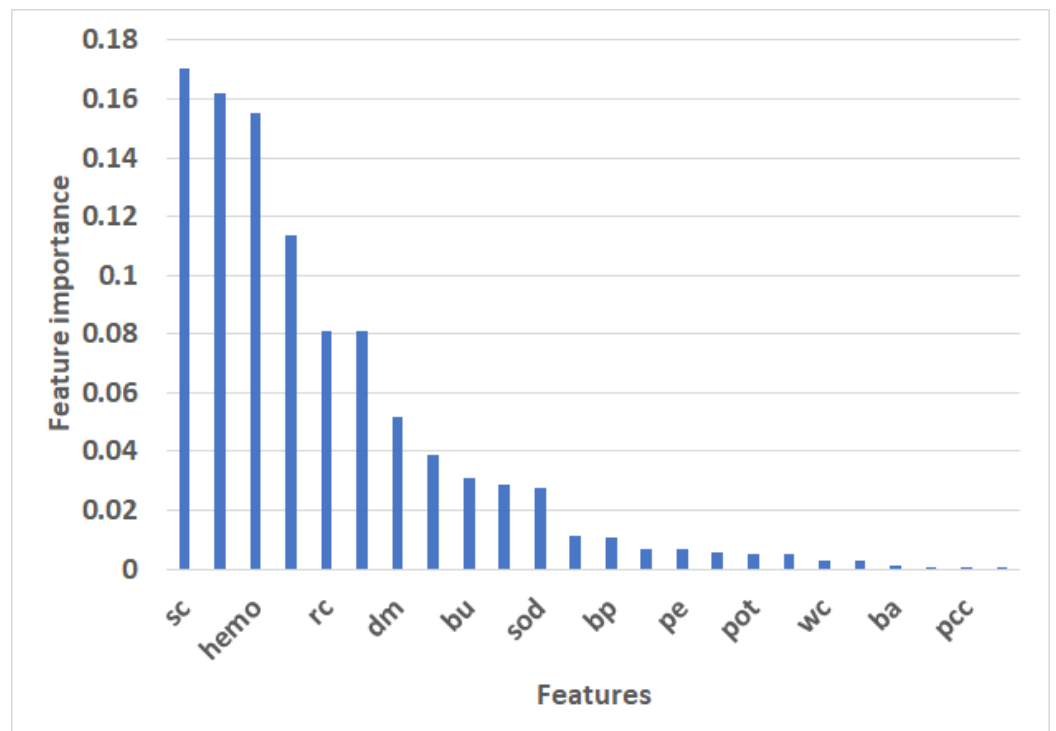


Figure 7. Feature Importance by Tree_Based (RF).

4.5. Performance of Applying DL Models and the Proposed Model with Feature-Selection Methods

In this subsection, we explore the performance of DL models (RNN, LSTM, and GRU) with layer1 and layer2 and the proposed model (Proposed-Layer1, Proposed-Layer2) with SVM as metalearners with different selected features by REF, Chi2, mutual_info and tree based. Proposed Layer1 refers to stacking of (RNN Layer1, LSTM Layer1, and GRU Layer1). Proposed-Layer1 refers to (stacking of RNN Layer1, LSTM Layer1, and GRU Layer1) and Proposed Layer12 refers to (stacking of RNN Layer2, LSTM Layer2, and GRU Layer2). We conducted two experiments with a different spiriting ratio for different experiments: Experimental 1 is explained the performance of splitting the CKD dataset into 80% training and 20% testing sets. Experimental 2 explains the performance of splitting the CKD dataset into 70% training and 30% testing sets.

4.6. Experimental 1

In this section, we explore the possibility of enhancing the performance of using a stacking ensemble deep learning model over DL models by using an 80% training and 20% testing set, as shown in Table 3. We make the following observations. The performance of Proposed Layer2 improved by about 1–3%. We can see that DL models with two layers achieve the best performance compared to DL models with one layer.

- For Chi2, Proposed Layer2 achieves the highest scores—95.0, 95.59, 95.0, and 95.05, in accuracy, precision, recall, and F1 score, respectively—compared to other models. Proposed Layer2 improved accuracy by 1, precision by 1.59, recall by 1, and F1 score by 1.05 compared to LSTM Layer2. LSTM Layer2 has the third-highest performance. GRU Layer1 registers the lowest performance accuracy = 86.25, precision = 87.62, recall = 86.25, and F1 score = 85.6.
- For mutual_info, we can see that in the table, Proposed Layer2 achieves the highest scores at 99.69, 99.71, 99.69, and 99.69 in accuracy, precision, recall, and F1 score, respectively, compared to other models. Proposed Layer2 improved accuracy by 2.19, precision by 2.05, recall by 2.18, and F1 score by 2.29 compared to LSTM Layer2. Proposed Layer1 registers the second-highest performance. LSTM Layer2 has the third-highest performance. RNN Layer1 registers the lowest performance accuracy = 93.75, precision = 94.32, recall = 93.75, and F1 score = 93.61.
- For RFE, Proposed Layer2 achieves the highest scores at 98.75, 98.79, 98.75, and 98.75 in accuracy, precision, recall, and F1 score, respectively, compared to other models. Proposed Layer2 improved accuracy by 2.22, precision by 2.39, recall by 2.22, and F1 score by 2.29 compared to RNN-Layer2. RNN Layer2 registers the third-highest accuracy = 96.53, precision = 96.4, recall = 96.53, and F1 score = 96.46. LSTM Layer1 registers the lowest performance accuracy = 91.25, precision = 91.68, recall = 91.25, and F1 score = 91.06.
- For Tree based, Proposed Layer2 achieves the highest accuracy, precision, recall, and F1 score at 99.38, 99.42, 99.38, and 99.38, respectively, compared to other models. Proposed Layer2 improved accuracy by 0.92, precision by 0.96, recall by 0.93, and F1 score by 0.92 compared to RNN Layer2. RNN Layer2 achieves the third-highest performance. RNN Layer1 registers the lowest performance accuracy = 96.25, precision = 96.59, recall = 96.25, and F1 score = 96.28.

Table 3. Performance of applying DL models and the proposed model with feature-selection methods by using 80–20 splitting.

Feature-Selection Methods	Models	Matrix Performance			
		Accuracy	Precision	Recall	F1-Score
Chi2	RNN Layer1	91.25	92.91	91.25	91.38
	RNN Layer2	92.5	92.77	92.5	92.38
	LSTM Layer1	88.75	89.6	88.75	88.38
	LSTM Layer2	94.0	94.0	94.0	94.0
	GRU Layer1	86.25	87.62	86.25	85.6
	GRU Layer2	93.75	93.89	93.75	93.68
	Proposed Layer1	93.75	93.74	93.75	93.73
	Proposed Layer2	95.0	95.59	95.0	95.05
mutual_info	RNN Layer1	93.75	94.32	93.75	93.61
	RNN Layer2	96.25	96.46	96.25	96.21
	LSTM Layer1	96.5	96.66	96.5	96.51
	LSTM Layer2	97.5	97.66	97.5	97.51
	GRU Layer1	95.0	95.18	95.0	95.03
	GRU Layer2	96.25	96.59	96.25	96.28
	Proposed Layer1	98.75	98.79	98.75	98.75
	Proposed Layer2	99.69	99.71	99.69	99.69
RFE	RNN Layer1	94.25	94.59	94.25	94.28
	RNN Layer2	96.53	96.4	96.53	96.46
	LSTM Layer1	91.25	91.68	91.25	91.06
	LSTM Layer2	95.0	95.05	95.0	94.96
	GRU Layer1	92.25	92.68	92.25	92.06
	GRU Layer2	95.0	95.37	95.0	94.92
	Proposed Layer1	97.5	97.66	97.5	97.51
	Proposed Layer2	98.75	98.79	98.75	98.75
Tree-based	RNN Layer1	96.25	96.59	96.25	96.28
	RNN Layer2	98.46	98.46	98.45	98.46
	LSTM Layer1	96.25	96.59	96.25	96.28
	LSTM Layer2	97.79	97.79	97.77	97.78
	GRU Layer1	97.5	97.66	97.5	97.51
	GRU Layer2	97.62	97.61	97.63	97.62
	Proposed Layer1	98.75	98.85	98.75	98.76
	Proposed Layer2	99.38	99.42	99.38	99.38

4.7. Experimental 2

In this section, we explore the possibility of enhancing the performance of using a stacking ensemble deep learning model over DL models by using a 70% training and a 30% testing set, as shown in Table 4. We make the following observations. The performance of Proposed Layer2 improved by about 1–3%. We can see that DL models with two layers achieve the best performance compared to DL models with one layer.

- For Chi2, Proposed Layer2 achieves the highest scores 94.9, 94.43, 94.01, and 94.62, in accuracy, precision, recall, and F1 score, respectively, compared to other models. Proposed Layer2 improved accuracy by 1.5, precision by 0.54, recall by 0.26, and F1 score by 0.94 compared to LSTM Layer2. LSTM Layer2 has the third-highest performance. GRU Layer1 registers the lowest performance accuracy = 89.67, precision = 89.5, recall = 89.93, and F1 score = 89.74.
- For mutual_info, Proposed Layer2 achieves the highest scores at 98.75, 98.88, 98.75, and 98.76 in accuracy, precision, recall, and F1 score, respectively, compared to other models. The proposed Layer2 enhanced precision by 1.02, recall by 0.8, and accuracy by 0.87 and F1 score by 0.85 compared to RNN Layer2. RNN Layer2 has the third-highest performance. LSTM Layer2 registers the lowest performance accuracy = 94.83, precision = 94.77, recall = 95.03, and F1 score = 94.9.
- For RFE, Proposed Layer2 achieves the highest scores at 96.31, 96.34, 96.23, and 96.28 in accuracy, precision, recall, and F1 score, respectively, compared to other models. Proposed Layer2 improved accuracy by 0.69, precision by 0.59, recall by 0.61, and F1 score by 0.7 compared to LSTM Layer1. LSTM Layer1 registers the third-highest performance. GRU Layer2 registers the lowest performance accuracy = 92.5, precision = 93.0, recall = 92.5, and F1 score = 92.34.
- For Tree based, Proposed Layer2 achieves the highest accuracy, precision, recall, and F1 score at 97.92, 98.19, 97.92, and 97.94, respectively, compared to other models. Proposed Layer2 improved accuracy by 2.89, precision by 3.17, recall by 2.8, and F1 score by 2.87 compared to RNN Layer2. RNN Layer2 achieves the third-highest performance. RNN Layer1 registers the lowest performance accuracy = 93.49, precision = 93.87, recall = 93.12, and F1 score = 93.46.

Table 4. Performance of applying DL models and the proposed model with features selection methods using 80–20 splitting.

Feature Selection Methods	Models	Matrix Performance			
		Accuracy	Precision	Recall	F1-Score
Chi2	RNN Layer1	90.67	90.58	90.93	90.74
	RNN Layer2	91.25	91.68	91.25	91.06
	LSTM Layer1	92.75	92.89	92.75	92.68
	LSTM Layer2	93.75	93.89	93.75	93.68
	GRU Layer1	89.67	89.58	89.93	89.74
	GRU Layer2	92.7	92.72	92.74	92.73
	Proposed Layer1	94.13	94.15	94.12	94.13
	Proposed Layer2	94.9	94.43	94.01	94.62
mutual_info	RNN Layer1	96.25	96.26	96.25	96.24
	RNN Layer2	97.88	97.86	97.95	97.91
	LSTM Layer1	96.08	96.08	96.06	96.07
	LSTM Layer2	94.83	94.77	95.03	94.9
	GRU Layer1	93.49	93.87	93.12	93.46
	GRU Layer2	96.31	96.34	96.23	96.28
	Proposed Layer1	98.33	98.48	98.33	98.34
	Proposed Layer2	98.75	98.88	98.75	98.76

Table 4. Cont.

Feature-Selection Methods	Models	Matrix Performance			
		Accuracy	Precision	Recall	F1 Score
RFE	RNN Layer1	94.11	94.66	92.68	93.44
	RNN Layer2	94.97	95.15	94.85	94.99
	LSTM Layer1	95.62	95.75	95.62	95.58
	LSTM Layer2	95.25	95.59	95.25	95.28
	GRU Layer1	93.75	94.07	93.75	93.65
	GRU Layer2	92.5	93.0	92.5	92.34
	Proposed Layer1	96.20	96.28	96.20	96.28
	Proposed Layer2	96.31	96.34	96.23	96.28
Tree-based	RNN Layer1	94.49	94.87	94.12	94.46
	RNN Layer2	95.03	95.02	95.12	95.07
	LSTM Layer1	94.83	94.77	95.03	94.9
	LSTM Layer2	94.38	94.62	94.38	94.3
	GRU Layer1	93.75	94.07	93.75	93.65
	GRU Layer2	93.49	93.87	93.12	93.46
	Proposed Layer1	97.5	97.66	97.5	97.51
	Proposed Layer2	97.92	98.19	97.92	97.94

4.8. Discussion

In this section, a discussion of the summarized experimental results is introduced. In addition, we discuss the best models for each feature-selection method. Moreover, we compare the proposed model with the literature studies and from the medical side.

4.8.1. The Best Models

Figure 8 shows the best models for each feature selection method of 20–80 splitting. The Proposed Layer2 with mutual_info achieves the highest accuracy, precision, recall, and F1 score at 99.69, 99.71, 99.69, and 99.69, respectively. The Proposed Layer2 with Chi2 registers the lowest and achieves the highest accuracy, precision, recall, and F1 score at 95.0, 95.59, 95.0, and 95.05, respectively.

Figure 9 shows the best models for each feature selection method of 30–70 splitting. The Proposed Layer2 with mutual_info achieves the highest accuracy, precision, recall, and F1 score at 94.9, 94.43, 94.01, and 94.62, respectively. The Proposed Layer2 with Chi2 registers the lowest and achieves the highest accuracy, precision, recall, and F1 score at 94.9, 94.43, 94.01, and 94.62, respectively.

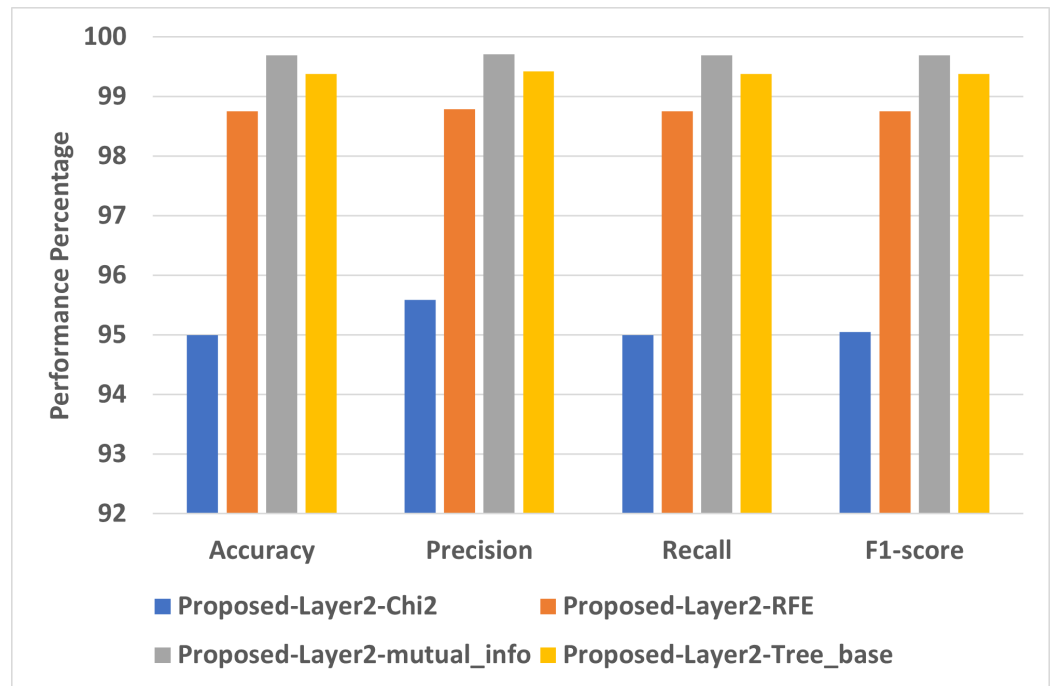


Figure 8. The best models for all feature selection methods of 20–80 splitting.

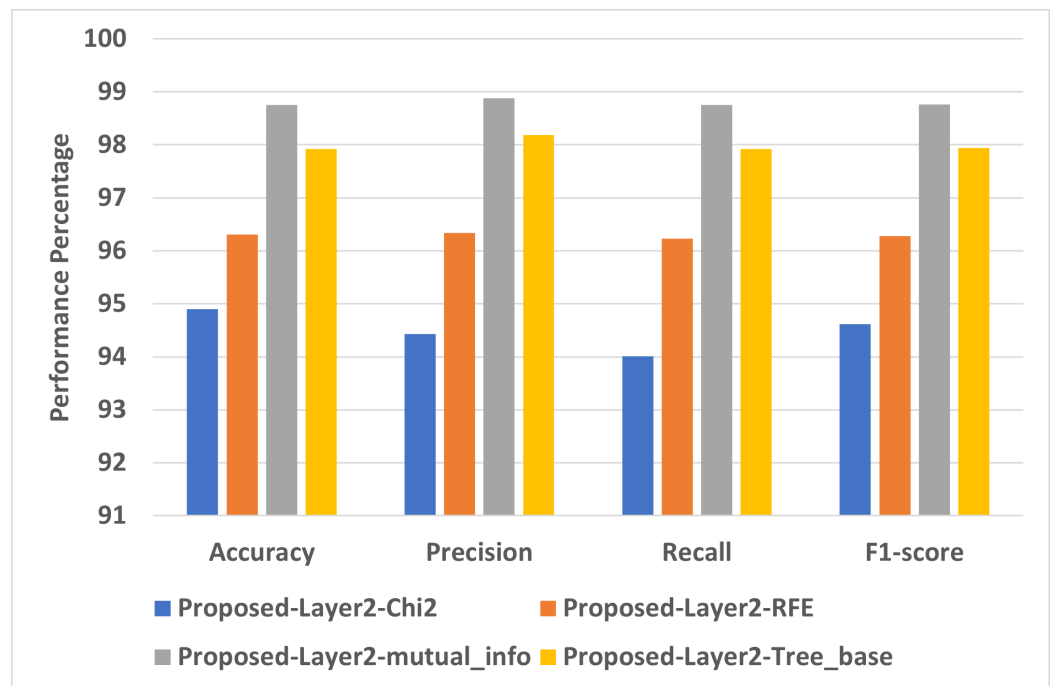


Figure 9. The best models for all feature selection methods of 30–70 splitting.

4.8.2. Comparison with Literature Studies

Table 5 compares previous studies and the proposed model. In [17,18,28], the data was partitioned into 80% for training and 20%. The data was partitioned into 70% for training and 30% for training. In [19,26,31], the authors made results by using 15-fold cross-validation.

We can see that the proposed model with mutual_info gives the highest accuracy at 99.69. In [17], GNB was registered as 94% of accuracy, and accuracy of 91% was recorded in [18]. In [20], NN SVM was recorded at 89% accuracy. In [9], autoencoder and NN had 94%. NN in [21] had 92% accuracy. In [58], CNN was recorded at 95.4. In [37], DNN had

74.7%. The authors used feature-selection methods. In [19], REF and the ensemble model recorded 94% accuracy. The authors [22] used a wrapper approach with LR, NN that was recorded with 96% accuracy.

Table 5. Comparison with literature studies.

Ref	FS	Models	Used Dataset	ACC
[17]	IG	GNB	Private Dataset	94
[18]	NO FS	NB, DT, RF	UCI ML repository	91
[19]	RFE	ensemble model (NB, SVM, MLP, DT)	UCI ML repository	94
[20]	NO FS	NN SVM	Data aggregated from 50 CKD and 50 control subjects	89
[9]	No FS	Auto encoder & NN	Two Private	93
[21]	No FS	NN	Dataset of 200 subjects aged more than 70 years	92
[22]	Wrapper approach	LR, NN	Private (100 subjects)	96
[37]	No FS	DNN	11,140 subjects' diabetes subjects with CKD	74.7
[26]	RFE	J48	UCI ML repository	85.5
[28]	No	OCNN	UCI ML repository	98.75
[31]	No	RF	UCI ML repository	99.75
The proposed model	mutual_info	Stacking ensemble	UCI ML repository	99.69

4.9. Model Explainability

Machine learning and deep learning have become integral to modern world functions. It is utilized to automate several tasks and discover patterns in data to make complex decisions. Like any decision-making tool, the degree of trust and confidence determines the decision [59]. Explainability is the process of providing information that clarifies why and how the model takes the decision and provides an explanation of the decision of the algorithm that could be understood by humans [60]. In order to make our proposed model more understandable, in this section, we explain our proposed model in terms of global and local explanations.

4.9.1. Global Feature Importance (Global Explainability)

To ensure the efficiency of our proposed model, we utilized the LIME library to show the behavior of the proposed model in terms of different features. As shown in Figure 10, each horizontal represents the impact of the feature in the overall decision. From Figure 10, we make the following conservation. (i) sg and hemo have the most significant impact on the overall decision. (ii) htn, sc, and PVC have a similar impact on prediction. (iii) All selected features greatly impact the overall decision. Figure 10 shows the density and the score of the high-impact features.

4.9.2. Local Explainability

In this subsection, we utilize LIME plots to explain each sample's output decision. As shown in Figure 11A,B, CKD in case 8 shows a case with a probability of 2 not to have a CKD with a probability of 1. It also shows the most impact features that contribute to moving the decision to the negative class (hemo = 8.00, Sc = 2.90, pvc = 24.00). In Figure 11B, the predicted class was 1, which was the actual class. Figure 11B shows the high-impact

features with their values that contribute to giving the positive class ($rc = 5.28$, $hemo = 17$, $sc = 0.70$, $PVC = 52$). These plots have high importance not only to show the predicted class versus the actual, but also to give a clear explanation of the reasons that move the final decision toward the class, which ensures the importance of our chosen features and their high impact on the final decision.

Global Term/Feature Importances

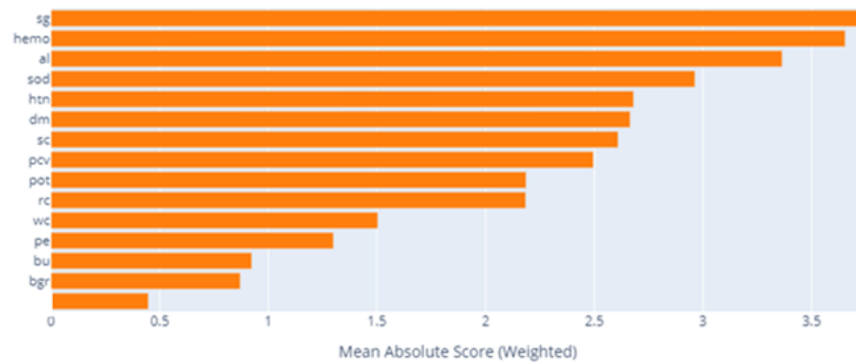
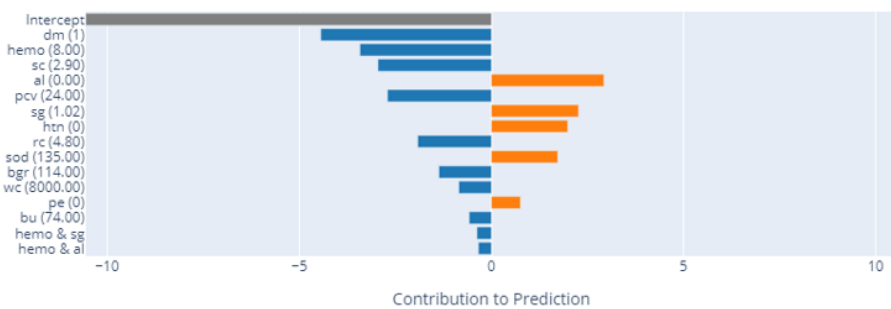


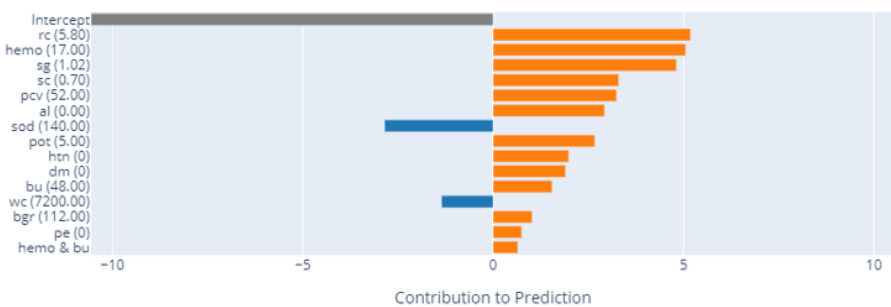
Figure 10. Global feature importance.

Local Explanation (Actual Class: 0 | Predicted Class: 0
Pr(y = 0): 1.000)



(A)

Local Explanation (Actual Class: 1 | Predicted Class: 1
Pr(y = 1): 1.000)



(B)

Figure 11. Local explanation of the proposed model.

4.10. Medical Side

The proposed model with `mutual_info` gives the best performance in several evaluation metrics (accuracy = 99.69, precision = 99.71, recall = 99.69, and F1 score = 99.69). The improved performance returned to the chosen and appropriate model and the selected feature selection, which extracts the essential features. This importance is also affirmed from the medical side. Traditionally, CKD is identified through medical examination with some lab tests, such as urine analysis and blood urea nitrogen (BUN). Notwithstanding the importance of these examinations, there are less common factors that should be considered in CKD prediction that have a significant effect in predicting CKD diseases. The following points summarize the importance of the chosen features in CKD prediction.

- Hemoglobin level was previously associated with heart failure, and many studies examined the relation between hemoglobin and kidney diseases and differences in the treatment process, and risks of death. For example, in [61], the authors aggregate the data of 722 adults from health plan records in California, analyze the correlation between hemoglobin level and other kidney functions, and then make the following observations. (i) The death level increased with lower hemoglobin with 95%, and the confidence interval ranged from 1.11 to 1.22 for hemoglobin from 12 to 12.8 g/dL, 94% with CI between 2.18 to 2.44 for hemoglobin (9.0 to 9.9). (ii) Relations are approximately the same for risk of hospitalization. (iii) The outcome of kidney functions significantly changed with hemoglobin level. (iv) This finding has no significant changes with systolic functions. The same is found true in [62].
- Packed cell volume (PVC) also has a significant effect on CKD [63]. In a study that was conducted to investigate the impact of the PVC, as well as reticulocyte count among 96 (62% male and 38% female) subjects aged 24–60 (mean age 35 ± 12.8) with CKD. The PVC calculated the Hawksley microhaematocrit centrifuge (Hawksley, UK). The mean PVC among CKD patients was 33 ± 7.98 among CKD compared to 37 ± 5.11 among control subjects. The difference was statistically significant ($p = 0.001$) [63].
- Serum creatine (SC) was utilized in a study conducted on 84 patients to determine whether serum creatine could be considered a marker in CKD. They concluded that there is a positive relationship between SC and EKD ($p < 0.001$) [64].
- In terms of red cells (RC), in the following study [65], the authors examined the relation between blood parameters such as WBC, RBC, and CKD. The study was conducted on patients aged 60–70 years with CKD. The results showed that CKD led to significant decreases in RBC count and lymphocyte count at 83.44%, and 76.1%, as well as a small decrease in platelets, counted at 6.28%, and WBC at 48.73% [66]. Urine tests could also be used to find red blood cells, which are used as an indicator of kidney disease (i.e., stone, cyst, failure, and bladder cancer infection) [67].
- Regarding sugar (SC) and diabetes mellitus (DM) and blood glucose random (BGR), several studies were conducted to study the correlation between diabetes mellitus and kidney disease occurrence [68], the authors concluded with several points as follows. (i) Nephrolithiasis, which is a symmetric disorder with CKD, increased with type 2 diabetes and metabolic syndrome. Glycemic control could contribute to delaying the progression of CKD. (ii) Dose adjustment among hypoglycemic patients is crucial. (iii) All drugs that are cleared by the kidney (i.e., glyburide) should be taken with caution, whereas other drugs that could be cleared by the liver and such sodium cotransporter (i.e., inhibitors) need to reduce in dose, practically when $GFR < 30$ mL/min.
- In terms of serum albumin (al), previous studies associate between kidney function decline and albumin. Recently, several studies explored serum albumin as a risk factor for CKD. For example, in [69], authors made a cohort study among CKD aged from 70 to 79, estimating the association through GFR values. The results showed that lower albumin levels were strongly associated with kidney function decline (-0.12 mL/min/ 1.77 m² per year, with a standard deviation of -0.01 , -0.020) when the results are divided into quarters. Serum albumin levels < 3.80 g/dL are

associated with kidney function decline (ratio 1.59; 1.22–2.27). This increased the risk of CKD incidence (ratio 1.29; 1.03–1.62). Urine albumin and creatinine levels are highly associated with kidney function decline (-0.08 mL/mil/1.72 per year for urine (ACR > 30 mg/g; -0.088 to -0.12) [70].

- In [71], authors explored the relationship between sodium, potassium, and kidney decline functions and concluded with a significant correlation between sodium and urea ($p = 0.005$, $r = 0.441$). In terms of the correlation between creatine and sodium, it did not show a significant relation ($p = 0.890$, $r = 0.023$).
- Blood pressure hypertension considers one of the main causes of kidney diseases as it leads to increased salt sensitivity, sympathetic tone, and upregulation of the aldosterone system [72]. Blood pressure also contributes to decreasing the slow progression of CKD as well as the risk of cardiovascular (CV) diseases [72]. Unless the certain relation between high blood pressure and CKD progression, a considerable debate still exists about optimal blood pressure [73].
- Blood urea considers the main source of nitrogen in a patient's body, which is filtered by the kidney to pass out through urine [74,75]. The main function of the kidney is to get rid of metabolic waste and maintain water PH. A high amount of urea in the blood significantly affects kidney function and may lead to kidney failure [76].
- In terms of blood edema (PE), it has a relation with kidney failure (AKF). Multiple organ dysfunction syndrome (MODS) is commonly associated with AKF, but edema occurs in septic patients with severe inflammatory response syndrome even without ARF (SIRS) [6].

5. Conclusions

The paper proposed an ensemble DL model for chronic kidney disease prediction by using different feature-selection methods. First, the data were preprocessed for encoding text features via label encoding, handling missing values, and detecting outliers. Secondly, different feature selection techniques were applied to choose the optimum feature list, including mutual information, chi-squared, RFE, and tree-based (RF). Thirdly, a stacking ensemble DL model was developed by combining the output of RNN, LSTM, and GRU in level-1 learning and using them to train and evaluate SVM in terms of metalearning in level 2. The performance of models was evaluated in terms of different metrics. The result showed that the proposed ensemble model with `mutual_info_classif` achieved the highest performance (accuracy = 99.69, precision = 99.71, recall = 99.69, and F1 score = 99.69).

Our proposed model is medically intuitive, as it is based on features that have proven to affect CKD diagnosis significantly. In the future, we intend to extend our work by collecting real datasets for university hospitals to ensure the generalization ability of our proposed model. Secondly, chronic kidney diseases overlap with other diseases; therefore, we intend to study and explore the correlation between all diseases and disorders and CKD health status. Thirdly, we intend to study the computational complexity of the realized algorithm.

Author Contributions: Methodology, H.S., L.A.G. and N.E.-R.; Software, H.S.; Validation, H.S.; Data curation, N.E.-R.; Writing—original draft, H.S., L.A.G., K.A. and N.E.-R.; Writing—review & editing, D.M.A., H.S., L.A.G., K.A., S.E.-S., R.S. and N.E.-R.; Visualization, R.S.; Funding acquisition, S.E.-S. All authors have read and agreed to the published version of the manuscript.

Funding: Nourah bint Abdulrahman University Researchers Supporting Project (PNURSP2023R435), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The direct link in the dataset citations will take the reader to all of the datasets that were utilized to support the study's assertions.

Acknowledgments: Nourah bint Abdulrahman University Researchers Supporting Project (PNURSP2023R435), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. We thank Kareem Nagaty Zayed (Kareem N. Zayed), the Lecturer of Nephrology at Mansoura University, Egypt, for his help and guidance throughout the study. The authors would like to acknowledge the support of the Deanship of Scientific Research at Prince Sattam bin Abdulaziz University.

Conflicts of Interest: All authors declare that they have no conflict of interest.

References

1. Eknayan, G.; Lameire, N.; Eckardt, K.; Kasiske, B.; Wheeler, D.; Levin, A.; Stevens, P.; Bilous, R.; Lamb, E.; Coresh, J.; et al. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Int.* **2013**, *3*, 5–14.
2. Kovesdy, C.P. Epidemiology of chronic kidney disease: An update 2022. *Kidney Int. Suppl.* **2022**, *12*, 7–11. [[CrossRef](#)]
3. Zhou, Y.; Yang, J. Chronic kidney disease: Overview. In *Chronic Kidney Disease*; Springer: Singapore, 2020; pp. 3–12.
4. Jha, V.; Garcia-Garcia, G.; Iseki, K.; Li, Z.; Naicker, S.; Plattner, B.; Saran, R.; Wang, A.Y.M.; Yang, C.W. Chronic kidney disease: Global dimension and perspectives. *Lancet* **2013**, *382*, 260–272. [[CrossRef](#)] [[PubMed](#)]
5. Wu, Y.; Yi, Z. Automated detection of kidney abnormalities using multi-feature fusion convolutional neural networks. *Knowl.-Based Syst.* **2020**, *200*, 105873. [[CrossRef](#)]
6. Swathi, K.; Vamsi Krishna, G. Prediction of Chronic Kidney Disease with Various Machine Learning Techniques: A Comparative Study. In *Smart Technologies in Data Science and Communication: Proceedings of SMART-DSC 2022*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 257–262.
7. Matsushita, K.; Ballew, S.H.; Wang, A.Y.M.; Kalyesubula, R.; Schaeffner, E.; Agarwal, R. Epidemiology and risk of cardiovascular disease in populations with chronic kidney disease. *Nat. Rev. Nephrol.* **2022**, *18*, 696–707. [[CrossRef](#)]
8. James, M.T.; Hemmelgarn, B.R.; Tonelli, M. Early recognition and prevention of chronic kidney disease. *Lancet* **2010**, *375*, 1296–1309. [[CrossRef](#)] [[PubMed](#)]
9. Ma, F.; Gao, J.; Suo, Q.; You, Q.; Zhou, J.; Zhang, A. Risk prediction on electronic health records with prior medical knowledge. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 1910–1919.
10. Navaneeth, B.; Suchetha, M. A dynamic pooling based convolutional neural network approach to detect chronic kidney disease. *Biomed. Signal Process. Control* **2020**, *62*, 102068. [[CrossRef](#)]
11. El-Rashidy, N.; Abuhmed, T.; Alarabi, L.; El-Bakry, H.M.; Abdelrazek, S.; Ali, F.; El-Sappagh, S. Sepsis prediction in intensive care unit based on genetic feature optimization and stacked deep ensemble learning. *Neural Comput. Appl.* **2022**, *34*, 3603–3632. [[CrossRef](#)]
12. Jayanthi, P. Machine learning and deep learning algorithms in disease prediction: Future trends for the healthcare system. In *Deep Learning for Medical Applications with Unique Data*; Elsevier: Amsterdam, The Netherlands, 2022; pp. 123–152.
13. Sun, Q.; Pfahringer, B. Bagging ensemble selection for regression. In Proceedings of the AI 2012: Advances in Artificial Intelligence: 25th Australasian Joint Conference, Sydney, Australia, 4–7 December 2012; Proceedings 25; Springer: Berlin/Heidelberg, Germany, 2012; pp. 695–706.
14. Odegua, R. An empirical study of ensemble techniques (bagging, boosting and stacking). In Proceedings of the Deep Learning IndabaX, Nairobi, Kenya, 25–31 August 2019. [[CrossRef](#)]
15. Sharafati, A.; Asadollah, S.B.H.S.; Al-Ansari, N. Application of bagging ensemble model for predicting compressive strength of hollow concrete masonry prism. *Ain Shams Eng. J.* **2021**, *12*, 3521–3530. [[CrossRef](#)]
16. Wah, Y.B.; Ibrahim, N.; Hamid, H.A.; Abdul-Rahman, S.; Fong, S. Feature Selection Methods: Case of Filter and Wrapper Approaches for Maximising Classification Accuracy. *Pertanika J. Sci. Technol.* **2018**, *26*, 329–340.
17. Rabby, A.S.A.; Mamata, R.; Laboni, M.A.; Ohidujjaman; Abujar, S. Machine learning applied to kidney disease prediction: Comparison study. In Proceedings of the 2019 IEEE 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 3–5 October 2019; pp. 1–7.
18. Walse, R.S.; Kurundkar, G.D.; Khamitkar, S.D.; Muley, A.A.; Bhalchandra, P.U.; Lokhande, S.N. Effective Use of Naïve Bayes, Decision Tree, and Random Forest Techniques for Analysis of Chronic Kidney Disease. In Proceedings of the International Conference on Information and Communication Technology for Intelligent Systems, Ahmedabad, India, 15–16 May 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 237–245.
19. Moreno-Sanchez, P.A. Chronic Kidney Disease Early Diagnosis Enhancing by Using Data Mining Classification and Features Selection. In Proceedings of the International Conference on IoT Technologies for HealthCare, Virtual Event, 24–26 November 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 61–76.
20. Akbari, A.; Swedko, P.J.; Clark, H.D.; Hogg, W.; Lemelin, J.; Magner, P.; Moore, L.; Ooi, D. Detection of chronic kidney disease with laboratory reporting of estimated glomerular filtration rate and an educational program. *Arch. Intern. Med.* **2004**, *164*, 1788–1792. [[CrossRef](#)]
21. Levey, A.S.; Inker, L.A.; Coresh, J. Chronic kidney disease in older people. *JAMA* **2015**, *314*, 557–558. [[CrossRef](#)]

22. Al Imran, A.; Amin, M.N.; Johora, F.T. Classification of chronic kidney disease using logistic regression, feedforward neural network and wide & deep learning. In Proceedings of the 2018 International Conference on Innovation in Engineering and Technology (ICIET), IEEE, Dhaka, Bangladesh, 27–28 December 2018; pp. 1–6.
23. Hassan, M.M.; Hassan, M.M.; Mollick, S.; Khan, M.A.R.; Yasmin, F.; Bairagi, A.K.; Raihan, M.; Arif, S.A.; Rahman, A. A Comparative Study, Prediction and Development of Chronic Kidney Disease Using Machine Learning on Patients Clinical Records. *Hum.-Cent. Intell. Syst.* **2023**, 1–13. [[CrossRef](#)]
24. Senan, E.M.; Al-Adhaileh, M.H.; Alsaade, F.W.; Aldhyani, T.H.; Alqarni, A.A.; Alsharif, N.; Uddin, M.I.; Alahmadi, A.H.; Jadhav, M.E.; Alzaharani, M.Y. Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. *J. Healthc. Eng.* **2021**, 2021, 1004767. [[CrossRef](#)] [[PubMed](#)]
25. Poonia, R.C.; Gupta, M.K.; Abunadi, I.; Albraikan, A.A.; Al-Wesabi, F.N.; Hamza, M.A. Intelligent diagnostic prediction and classification models for detection of kidney disease. *Healthcare* **2022**, *10*, 371. [[CrossRef](#)]
26. Ilyas, H.; Ali, S.; Ponum, M.; Hasan, O.; Mahmood, M.T.; Iftikhar, M.; Malik, M.H. Chronic kidney disease diagnosis using decision tree algorithms. *BMC Nephrol.* **2021**, *22*, 273. [[CrossRef](#)] [[PubMed](#)]
27. Swain, D.; Mehta, U.; Bhatt, A.; Patel, H.; Patel, K.; Mehta, D.; Acharya, B.; Gerogiannis, V.C.; Kanavos, A.; Manika, S. A Robust Chronic Kidney Disease Classifier Using Machine Learning. *Electronics* **2023**, *12*, 212. [[CrossRef](#)]
28. Mondol, C.; Shamrat, F.J.M.; Hasan, M.R.; Alam, S.; Ghosh, P.; Tasnim, Z.; Ahmed, K.; Bui, F.M.; Ibrahim, S.M. Early Prediction of Chronic Kidney Disease: A Comprehensive Performance Analysis of Deep Learning Models. *Algorithms* **2022**, *15*, 308. [[CrossRef](#)]
29. Sawhney, R.; Malik, A.; Sharma, S.; Narayan, V. A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease. *Decis. Anal. J.* **2023**, *6*, 100169. [[CrossRef](#)]
30. Qin, J.; Chen, L.; Liu, Y.; Liu, C.; Feng, C.; Chen, B. A machine learning methodology for diagnosing chronic kidney disease. *IEEE Access* **2019**, *8*, 20991–21002. [[CrossRef](#)]
31. Nishat, M.M.; Faisal, F.; Dip, R.R.; Nasrullah, S.M.; Ahsan, R.; Shikder, F.; Asif, M.A.A.R.; Hoque, M.A. A comprehensive analysis on detecting chronic kidney disease by employing machine learning algorithms. *EAI Endorsed Trans. Pervasive Health Technol.* **2021**, *7*, e1. [[CrossRef](#)]
32. Shamrat, F.J.M.; Ghosh, P.; Sadek, M.H.; Kazi, M.A.; Shultana, S. Implementation of machine learning algorithms to detect the prognosis rate of kidney disease. In Proceedings of the 2020 IEEE International Conference for Innovation in Technology (INOCON), Bangalore, India, 6–8 November 2020; pp. 1–7.
33. Pal, S. Chronic Kidney Disease Prediction Using Machine Learning Techniques. *Biomed. Mater. Devices* **2022**, *9*, 109. [[CrossRef](#)]
34. Ren, Y.; Fei, H.; Liang, X.; Ji, D.; Cheng, M. A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 131–138. [[CrossRef](#)]
35. Krishnamurthy, S.; Kapeleshh, K.; Dovgan, E.; Luštrek, M.; Gradišek Piletič, B.; Srinivasan, K.; Li, Y.C.; Gradišek, A.; Syed-Abdul, S. Machine learning prediction models for chronic kidney disease using national health insurance claim data in Taiwan. *medRxiv* **2020**. [[CrossRef](#)]
36. Song, X.; Waitman, L.R.; Hu, Y.; Yu, A.S.; Robins, D.; Liu, M. Robust clinical marker identification for diabetic kidney disease with ensemble feature selection. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 242–253. [[CrossRef](#)] [[PubMed](#)]
37. Jardine, M.J.; Hata, J.; Woodward, M.; Perkovic, V.; Ninomiya, T.; Arima, H.; Zoungas, S.; Cass, A.; Patel, A.; Marre, M.; et al. Prediction of kidney-related outcomes in patients with type 2 diabetes. *Am. J. Kidney Dis.* **2012**, *60*, 770–778. [[CrossRef](#)]
38. Chronic Kidney Disease Dataset. Available online: https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease (accessed on 8 February 2023).
39. Elhoseny, M.; Shankar, K.; Uthayakumar, J. Intelligent diagnostic prediction and classification system for chronic kidney disease. *Sci. Rep.* **2019**, *9*, 9583. [[CrossRef](#)] [[PubMed](#)]
40. Mahdhaoui, A.; Chetouani, M.; Cassel, R.S.; Saint-Georges, C.; Parlato, E.; Laznik, M.C.; Apicella, F.; Matorini, F.; Maestro, S.; Cohen, D. Computerized home video detection for motherese may help to study impaired interaction between infants who become autistic and their parents. *Int. J. Methods Psychiatr. Res.* **2011**, *20*, e6–e18. [[CrossRef](#)]
41. Moon, T.K. The expectation-maximization algorithm. *IEEE Signal Process. Mag.* **1996**, *13*, 47–60. [[CrossRef](#)]
42. Greco, L.; Luta, G.; Krzywinski, M.; Altman, N. Analyzing outliers: Robust methods to the rescue. *Nat. Methods* **2019**, *16*, 275–277. [[CrossRef](#)]
43. Yu, L.; Liu, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th International Conference on Machine Learning, Washington, DC, USA, 21–24 August 2003; Volume 3, pp. 856–863.
44. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
45. Venkatesh, B.; Anuradha, J. A review of feature selection and its methods. *Cybern. Inf. Technol.* **2019**, *19*, 3–26. [[CrossRef](#)]
46. Chi-Squared. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html (accessed on 8 February 2023).
47. Mutual Information. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html (accessed on 8 February 2023).
48. Lin, X.; Li, C.; Zhang, Y.; Su, B.; Fan, M.; Wei, H. Selecting feature subsets based on SVM-RFE and the overlapping ratio with applications in bioinformatics. *Molecules* **2017**, *23*, 52. [[CrossRef](#)] [[PubMed](#)]
49. Venkatesh, B.; Anuradha, J. A fuzzy gaussian rank aggregation ensemble feature selection method for microarray data. *Int. J. Knowl.-Based Intell. Eng. Syst.* **2020**, *24*, 289–301. [[CrossRef](#)]

50. Savić, M.; Kurbalija, V.; Ivanović, M.; Bosnić, Z. A feature selection method based on feature correlation networks. In Proceedings of the Model and Data Engineering: 7th International Conference (MEDI 2017), Barcelona, Spain, 4–6 October 2017; Proceedings 7; Springer: Berlin/Heidelberg, Germany, 2017; pp. 248–261.
51. Dy, J.G.; Brodley, C.E. Feature selection for unsupervised learning. *J. Mach. Learn. Res.* **2004**, *5*, 845–889.
52. Zhang, Z. Improved adam optimizer for deep neural networks. In Proceedings of the 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), IEEE, Banff, AB, Canada, 4–6 June 2018; pp. 1–2.
53. Wanto, A.; Windarto, A.P.; Hartama, D.; Parlina, I. Use of binary sigmoid function and linear identity in artificial neural networks for forecasting population density. *IJISTECH (Int. J. Inf. Syst. Technol.)* **2017**, *1*, 43–54. [\[CrossRef\]](#)
54. Basili, R.; Croce, D. *Recurrent Neural Networks*; IntechOpen: London, UK, 2008.
55. Skansi, S. *Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2018.
56. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Gated feedback recurrent neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 6–11 July 2015; pp. 2067–2075.
57. O'Malley, T.; Bursztein, E.; Long, J.; Chollet, F.; Jin, H.; Invernizzi, L. Keras tuner. Retrieved May 2019, 21, 2020.
58. Krishnamurthy, S.; Ks, K.; Dovgan, E.; Luštrek, M.; Gradišek Piletič, B.; Srinivasan, K.; Li, Y.C.; Gradišek, A.; Syed-Abdul, S. Machine learning prediction models for chronic kidney disease using national health insurance claim data in Taiwan. *Healthcare* **2021**, *9*, 546. [\[CrossRef\]](#)
59. Kakogeorgiou, I.; Karantzalos, K. Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *103*, 102520. [\[CrossRef\]](#)
60. Gulum, M.A.; Trombly, C.M.; Kantardzic, M. A review of explainable deep learning cancer detection models in medical imaging. *Appl. Sci.* **2021**, *11*, 4573. [\[CrossRef\]](#)
61. Go, A.S.; Yang, J.; Ackerson, L.M.; Lepper, K.; Robbins, S.; Massie, B.M.; Shlipak, M.G. Hemoglobin level, chronic kidney disease, and the risks of death and hospitalization in adults with chronic heart failure: The Anemia in Chronic Heart Failure: Outcomes and Resource Utilization (ANCHOR) Study. *Circulation* **2006**, *113*, 2713–2723. [\[CrossRef\]](#)
62. Navaneethan, S.D.; Bonifati, C.; Schena, F.P.; Strippoli, G.F. Evidence for optimal hemoglobin targets in chronic kidney disease. *J. Nephrol.* **2006**, *19*, 640–647.
63. Wallis, P.; Cunningham, J.; Few, J.; Newland, A.; Empey, D. Effects of packed cell volume reduction on renal haemodynamics and the renin-angiotensin-aldosterone system in patients with secondary polycythaemia and hypoxic cor pulmonale. *Clin. Sci.* **1986**, *70*, 81–90. [\[CrossRef\]](#) [\[PubMed\]](#)
64. Sit, D.; Basturk, T.; Yildirim, S.; Karagoz, F.; Bozkurt, N.; Gunes, A. Evaluation of the serum cystatin C values in prediction of indications for hemodialysis in patients with chronic renal failure. *Int. Urol. Nephrol.* **2014**, *46*, 57–62. [\[CrossRef\]](#)
65. Asaduzzaman, M.; Shobnam, A.; Farukuzzaman, M.; Gaffar, A.; Juliana, F.M.; Sharker, T.; Dutta, K.K.; Islam, M.J. Assessment of Red Blood Cell Indices, White Blood Cells, Platelet Indices and Procalcitonin of Chronic Kidney Disease Patients under Hemodialysis. *Int. J. Health Sci. Res.* **2018**, *8*, 98–109.
66. Roy, J.P.; Devarajan, P. Acute kidney injury: Diagnosis and management. *Indian J. Pediatr.* **2020**, *87*, 600–607. [\[CrossRef\]](#) [\[PubMed\]](#)
67. Kelly, C.J.; Brown, A.P.; Taylor, J.A. Artificial Intelligence in Pediatrics. In *Artificial Intelligence in Medicine*; Springer: Cham, Switzerland, 2020; pp. 1–18.
68. Doshi, S.M.; Friedman, A.N. Diagnosis and management of type 2 diabetic kidney disease. *Clin. J. Am. Soc. Nephrol.* **2017**, *12*, 1366–1373. [\[CrossRef\]](#) [\[PubMed\]](#)
69. Jarad, G.; Knutsen, R.H.; Mecham, R.P.; Miner, J.H. Albumin contributes to kidney disease progression in Alport syndrome. *Am. J. Physiol.-Ren. Physiol.* **2016**, *311*, F120–F130. [\[CrossRef\]](#)
70. Lang, J.; Katz, R.; Ix, J.H.; Gutierrez, O.M.; Peralta, C.A.; Parikh, C.R.; Satterfield, S.; Petrovic, S.; Devarajan, P.; Bennett, M.; et al. Association of serum albumin levels with kidney function decline and incident chronic kidney disease in elders. *Nephrol. Dial. Transplant.* **2018**, *33*, 986–992. [\[CrossRef\]](#)
71. Samsuria, I.K. The Relationship between sodium, potassium, and hypothyroidism in Chronic Kidney Disease (CKD) patients. *Bali Med. J.* **2019**, *8*, 264. [\[CrossRef\]](#)
72. Gentile, G.; McKinney, K.; Reboldi, G. Tight blood pressure control in chronic kidney disease. *J. Cardiovasc. Dev. Dis.* **2022**, *9*, 139. [\[CrossRef\]](#) [\[PubMed\]](#)
73. Mayr, F.B.; Yende, S.; Linde-Zwirble, W.T.; Peck-Palmer, O.M.; Barnato, A.E.; Weissfeld, L.A.; Angus, D.C. Infection rate and acute organ dysfunction risk as explanations for racial differences in severe sepsis. *JAMA* **2010**, *303*, 2495–2503. [\[CrossRef\]](#) [\[PubMed\]](#)
74. Salazar, J.H. Overview of urea and creatinine. *Lab. Med.* **2014**, *45*, e19–e20. [\[CrossRef\]](#)
75. Griffin, B.R.; Faubel, S.; Edelstein, C.L. Biomarkers of drug-induced kidney toxicity. *Ther. Drug Monit.* **2019**, *41*, 213. [\[CrossRef\]](#)
76. Takemoto, Y.; Naganuma, T. Kidney function tests. *Jpn. J. Clin. Urol.* **2012**, *66*, 274–278.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.