JOURNAL OF BASIC SCIENCE AND ENGINEERING

A novel hybridized prediction model for detection of chronic renal diseases using Multilayer Perceptron-Stochastic Gradient Descent based boosted classifiers

[1]Alamma B.H., [2] Manjula Sanjay Koti, [3] C.H. Vanipriya

[1] Research Scholar, VTU Research Centre, Dept. of MCA, Sir MVIT, Bangalore &
Assistant Professor, Dept. of MCA, Dayananda Sagar College of Engineering,
Bangalore-560078, Karnataka, India

[2] Supervisor, Professor & Head, Dept. of MCA,
Dayananda Sagar Acadmey of Technoloy and M anagment,
Bangalore-560072, Karnataka, India

[3] Co-Supervisor, Professor & HOD, MCA Dept.,
Sir M Visvesvaraya Institute of Technology, Krishnadevaraya Nagar,
Hunasamaranahalli, International Airport Road, Bangalore – 562157

DOI: https://doie.org/10.1127/Jbse.2024145743

***Abstract*— Chronic Kidney Disease (CKD) is a critical health condition that affects millions worldwide, necessitating effective early diagnosis to mitigate its progression and associated mortality. The primary challenge in CKD prediction lies in handling complex, high-dimensional medical data and addressing issues of data imbalance, feature selection, and integration of various predictive models. Traditional single-model approaches often fall short in capturing the intricate patterns necessary for accurate diagnosis. The proposed research aims to develop and build a novel hybridized CKD prediction model leveraging an ensemble of advanced machine learning techniques to enhance diagnostic accuracy and reliability. This hybrid model integrates Multilayer Perceptron (MLP), Stochastic Gradient Descent (SGD), Adaptive Boosting (Adaboost), Logistic Regression, and Random Forest, fortified by a series of robust methodologies and Clinical Prediction Models (CPMs). Therefore, a hybridized model, combining the strengths of various algorithms, is proposed to achieve a more comprehensive and robust CKD prediction system.**

To improve model performance, feature selection methods such as ANOVA, Pearson correlations, and Cramer's V tests are applied. Incorporating deep stacked autoencoder networks allows for effective learning from multimedia data, enhancing the model's ability to process and interpret complex medical images and signals. Integrating CPMs provides a clinical context to the predictions, making the model's output more relevant and actionable in real-world medical settings. This comprehensive approach not only enhances diagnostic accuracy but also provides a framework that can be adapted to other complex medical prediction tasks.

***Keywords*— CKD, Prediction, Detection, Robustness, Anova, Correlation.**

I. INTRODUCTION

Developing a novel hybridized CKD prediction model involves integrating multiple methodologies to enhance the accuracy, robustness, and interpretability of the detection system. The model starts with Logistic Regression (LR) as a baseline classifier, providing a straight-forward and interpretable probabilistic framework. Feature selection methods such as ANOVA, Pearson correlations, and Cramer's V tests are used to identify the most relevant predictors, ensuring that only significant features are included in the model. Support Vector Machines (SVM), alongside Naive Bayes, offer robust initial classification capabilities, each contributing unique strengths in handling different types of data. Random Forest (RF) further enhances the prediction accuracy by aggregating multiple decision trees, while Multilayer Perceptron (MLP) addresses class imbalance through advanced neural network techniques.

Additionally, the integration of DL techniques, specifically Deep Stacked Autoencoder Networks, enables the model to learn from complex multimedia data, such as medical images and clinical reports, extracting high-level features that improve diagnostic precision. Boosted classifiers, like Gradient Boosting Machines (GBM) or AdaBoost, refine the model by combining several weak classifiers into a strong predictor, leveraging the selected features to enhance performance. Clinical Prediction Models (CPMs) ensure the model's clinical relevance by incorporating various demographic and clinical factors, providing a comprehensive risk assessment for CKD progression. This hybrid approach combines the strengths of traditional statistical methods, machine learning algorithms, and deep learning techniques to deliver a robust, accurate, and interpretable CKD prediction system. Also, the hybridized

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

method balances the advantages & dis-advantages of various methods and tries to give almost 100 % results.

The proposed research aims to develop and build a novel hybridized CKD prediction model for detecting chronic renal diseases using an MLP-SGD based boosted classifier and clinical prediction models (CPMs), yielding promising outcomes. By integrating rejuvenated models such as Multilayer Perceptron (MLP), Stochastic Gradient Descent (SGD), Adaptive Boosting (Adaboost), Logistic Regression, and Random Forest, this hybridized approach demonstrates significant improvements in CKD detection accuracy. The inclusion of mortality prediction capabilities and advanced feature selection methods like ANOVA, Pearson correlations, and Cramer's V tests enhances the model's precision and robustness. Additionally, the research leverages Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes, and Random Forest (RF) algorithms to further refine the prediction process.

To address the challenge of imbalanced data, the research work utilizes a multilayer perceptron approach, ensuring balanced and reliable predictions. Additionally, the integration of multimedia data learning through a deep stacked autoencoder network improves the model's ability to analyze and interpret complex datasets. The boosted classifier and feature selection methods will play a crucial role in enhancing the diagnostic capabilities for chronic kidney disease. The developed CKD prediction model will show some superior performances in clinical settings, providing accurate and timely predictions that can aid in early diagnosis and treatment planning. This objective underscores the potential of hybridized models in medical diagnostics and paves the way for future research and development in chronic disease prediction and management.

This paper is organized as follows section -2 Methodology and processes employed section-3 Steps in the process of design & development section-4 Datasets & databases utilized for the proposed research work section-5block-diagram of the process development of the hybridized ml model for detection of chronic renal diseasessection-6 Proposed algorithm steps section-7 Final outcome of the proposed work section-8 Conclusion\

II. RELEATED WORK

The authors outline a method that will facilitate blood urea and glucose monitoring for diabetics with chronic kidney disease (CKD). In a comparison study, blood urea and glucose were predicted using Partial Least Square Regression (PLSR) and Backpropagation Artificial Neural Network (BP-ANN) models. With an RMSE of 0.69 mg/dL, R2 = 0.96, and accuracy of 95.96% for urea and an RMSE of 2.06 mg/dL, R2 = 0.99, and accuracy of 98 for glucose, the BP-ANN model was able to accurately forecast increases in blood urea and glucose(Yu, C.Y ,2018).

The authors concentrated on using machine learning methods on data from blood tests. Renal teams would find it simpler to suggest that primary care physicians refer patients to secondary care, where they may receive medical assistance and an earlier expert review. They achieved an overall accuracy of 88.48%, 87.12%, and 85.29%, respectively, using logistic regression, ANN, and SVM. With a sensitivity performance score of 89.74%, ANNs outperformed SVM (85.51%) and logistic regression (86.67%). (Menzies,et.al,2018).

The authors present a paradigm to help people predict their risk of developing a chronic kidney disease that progressively worsens after catching COVID-19.KNN, Naive Bayes, ANN, and Ant Colony Optimization (ACO) were used to make assumptions; the results were 95% for KNN, 98.30% for Naive Bayes, 97.5% for ANN, and 95.5% for ACO. They have concluded that their initial prediction of renal diseases after COVID-19 is likely to be accurate. (Sindhuja,et.al,2016).

The authors of this paper have reported on the category of CKD utilizing machine learning models. The glomerular filtration rate was used to evaluate the stages of chronic kidney disease (CKD) in those who were diagnosed with the disease. They concluded that when classifying CKD patients with HIV, the DNN model performed with 99% accuracy.(Sajida Perveen,et.al.2018).

In order to evaluate chronic kidney disease, the author lists a few key symptoms. Weka technologies include supervised machine learning methods including Bagging, Adaptive Boosting (Adaboost), Stochastic Gradient Descent (SGD), Multilayer Perceptron (MLP), and Logistic Regression (LG). Analysis was assessed using a classifier. Principal component analysis (PCA) was used to extract the features of each characteristic. Other algorithms had higher ROC curve values, but the Random Forest (RF) approach had the best accuracy (about 99%).(P. Suresh Kumar,2017).

They discussed the connection between individual risk factors for metabolic diseases and diabetes mellitus using AI approaches, and they created tailored training sets based on the examination results. The findings showed that Nave Bayes with K-medoids had the highest curve value among the other approaches when compared to random under-examining, over-testing, and no testing feature. (SundusAbrar,et.a[,.2021).

The authors have focused on classification methods like logistic regression, random forests, and tree-based decision trees for the study of chronic renal diseases. They specified a number of measures for the dataset that was taken from the standard UCI repository in order to compare techniques. They concluded that the accuracy levels of random forest and logistic regression were 99.24, 94.16, and 98.48, respectively. 100, 95.12, and 98.82 for precision and 97.61, 96.29, and 100 for recall. Two feature selection strategies are merged, leveraging the benefits of each feature selection technique. Logistic regression has the best accuracy and recall when compared to decision trees. (Dinu A.J,et,al,2018).

III. METHODOLOGY & PROCESSES EMPLOYED

To develop & build a novel hybridized CKD prediction model for detection of chronic renal diseases using MLP-SGD based Boosted classifier & CPMs using the following combination of rejuvenated mathematical models utilizing

a) Multilayer Perceptron (MLP)

b) Stochastic Gradient Descent (SGD) Model

c) Adaptive Boosting (Adaboost) Model

d) Logistic Regression & Random Forest Model

A. Other objectives to be solved are ….

1. Mortality prediction
2. feature selection methods like ANOVA, Pearson correlations, and Cramer's V tests
3. SVM (Support Vector Machine)
4. RF (Random Forest)
5. Imbalanced data by multilayer perceptrons
6. Multimedia data learning using deep stacked autoen coder networks
7. Boosted classifier and features selection for enhancing chronic kidney disease diagnoses
8. Clinical prediction models (CPMs)

B. Aim of the proposed works

The main aim of the proposed work is to develop and build a novel hybridized chronic kidney disease (CKD) prediction model for detecting chronic renal diseases using a combination of rejuvenated models and advanced methodologies, so, in order to achieve this aim, the following 9 steps are used in our proposed work.

IV. STEPS IN THE PROCESS OF DESIGN & DEVELOPMENT

The following 8 steps are used in the design & development of the algorithms for the problem considered

Step 1: Data Collection and Preprocessing

- Data Sources: Utilized online datasets such as UCI CKD Dataset, Kaggle CKD datasets, NHANES, USRDS, and MIMIC-III.
- Data Cleaning: Handled missing values through imputation or removal, normalize continuous variables, and encode categorical variables.
- Feature Engineering: Extracted the relevant features and apply transformations to prepare the data for analysis.

Step 2: Feature Selection

- ANOVA (Analysis of Variance): Identified significant features differentiating CKD from non-CKD groups.
- Pearson Correlation: Measured the linear relationships between continuous variables and CKD status.
- Cramer's V Test: Assessed the association strength between categorical variables and CKD status.
- Selected the features based on their importance and correlation with CKD for inclusion in the model.

Step 3: Handling of Imbalanced Data

- Multilayer Perceptron (MLP): Used techniques like Synthetic Minority Over-sampling Technique (SMOTE) to balance the class distribution.
- Training MLP: Trained an MLP model on the balanced dataset to handle the imbalance effectively.

Step 4: Model Training and Development

- Multilayer Perceptron (MLP): Trained an MLP to capture complex patterns in the data.
- Stochastic Gradient Descent (SGD): Optimized the neural network parameters using SGD for efficient convergence.
- Support Vector Machine (SVM): Trained an SVM classifier to create a decision boundary for CKD classification.
- Random Forest (RF): Trained an RF model to leverage ensemble learning for robust predictions.
- Logistic Regression: Trained a logistic regression model to provide a straightforward yet effective baseline classifier.

Step 5: Multimedia Data Learning

- Deep Stacked Autoencoder Networks: Learnt the features from multimedia data such as numerical data of med images and integrated these high-level features with the clinical data.

Step 6: Adaptive Boosting (Adaboost)

- Initialize Weights: Initialized the weights for each instance in the dataset.
- Train Weak Classifiers: Iteratively trained the weak classifiers (MLP, SGD, SVM, RF, Logistic Regression) on the weighted dataset, focusing on misclassified instances.
- Update Weights: Adjusted the weights based on classifier performance, emphasizing harder-to-classify instances.
- Combine Classifiers: Used the Adaboost to aggregate the weak classifiers into a strong ensemble model, boosting the overall prediction accuracy.

Step 7: Clinical Prediction Models (CPMs)

- Integration: Develop CPMs using clinical and demographic data to ensure relevance and comprehensive risk assessments.
- Validation: Validated the CPMs against existing clinical standards and ensure they complement the machine learning models.

Step 8: Model Evaluation and Validation

- Train-Test Split: Splitted the dataset into training and testing sets to evaluate model performance.
- Cross-Validation: Performed the cross-validation to assess model robustness.
- Evaluation Metrics: Used the accuracy, precision, recall, F1-score, and ROC-AUC to evaluate and compare models.
- Model Comparison: Compared the performance of different models and fine-tune hyperparameters for optimal results.

Step 9: Deployment and Monitoring

- Deployment: Deployed the hybrid CKD prediction model in clinical settings.
- Monitoring: Continuously monitored the model performance, update with new data, and refine as necessary.
- Detection: Detected whether the disease is present or not accurately & more precisely.

V. DATASETS & DATABASES UTILIZED FOR THE PROPOSED RESEARCH WORK

Online datasets that are used as inputs for developing and building a novel hybridized CKD prediction model using our own created our own database, which is given as input to our hybridized algorithm for the prediction & detection process of the renal disease. We selected 10 attributes from the dataset that we are using from the repository dataset of chronic kidney disease as input features. To develop and build a novel hybridized CKD prediction model, the following numerical datasets are used & these datasets include clinical, demographic, laboratory, and imaging data to comprehensively assess and predict chronic kidney disease (CKD).

Clinical Data – Age, Gender (e.g., Male = 1, Female = 0), Blood Pressure, Diabetes Status as a Binary indicator (1 for diabetic, 0 for non-diabetic), Hypertension Status as Binary indicator (1 for hypertensive, 0 for non-hypertensive), Anemia Status as Binary indicator (1 for anemic, 0 for non-anemic).

Laboratory Data - Serum Creatinine, Estimated Glomerular Filtration Rate (eGFR), Blood Urea Nitrogen (BUN), Hemoglobin, Serum Potassium, Serum Sodium, Serum Albumin, Cholesterol Levels, Urine Albumin to Creatinine Ratio (ACR).

Demographic Data – Ethnicity for Categorical data converted to numerical values (e.g., Ethnicity A = 1, Ethnicity B = 2, etc.), Smoking Status as Binary indicator (1 for smoker, 0 for non-smoker), BMI (Body Mass Index).

Medical History - History of Cardiovascular Diseases as Binary indicator (1 for present, 0 for absent), Family History of CKD as Binary indicator (1 for present, 0 for absent), Medication Usage as Binary indicators for various medications affecting kidney function (e.g., NSAIDs, ACE inhibitors).

Outcome Data - CKD Status as Binary indicator (1 for CKD, 0 for non-CKD), Mortality Status as Binary indicator (1 for deceased, 0 for alive) for mortality prediction models.

VI. BLOCK-DIAGRAM OF THE PROCESS DEVELOPMENT OF THE HYBRIDIZED ML MODEL FOR DETECTION OF CHRONIC RENAL DISEASES

To develop and build a novel hybridized model for predicting chronic kidney disease (CKD), we propose a unique combination of the ML techniques. The goal is to enhance the accuracy and reliability of CKD detection by leveraging the strengths of multiple models. The hybridized CKD prediction model incorporates a combination of rejuvenated models with 4 processes namely – (i) Multilayer Perceptron (MLP), (ii) Stochastic Gradient Descent (SGD), (iii) Adaptive Boosting (Adaboost), Logistic Regression, and (iv) Random Forest models, which is projected as shown below and explained in 5 parts from part-i to part-v.

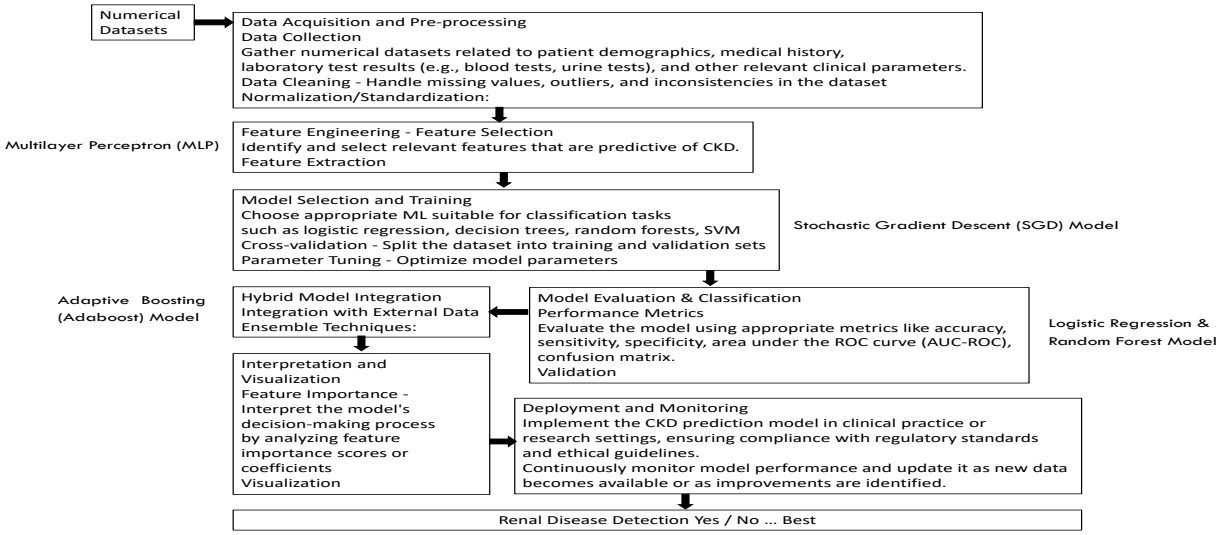

Fig. 1. Overall integration of the ensemble model for the prediction & detection of the correct renal diseases

A. Part – i

We have used the Multilayer Perceptron (MLP) in our work, which is a class of feedforward ANN's that consist of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. In this hybridized model, the MLP is utilized for its capability to capture complex patterns and relationships within the dataset. It serves as the initial classifier, processing input data and generating intermediate predictions that feed into subsequent models. The Multilayer Perceptron (MLP) is a class of artificial neural networks (ANN)

designed to model complex relationships within data. The MLP consists of an input layer, one or more hidden layers, and an output layer. Each layer is composed of nodes (neurons), and each node in a layer is fully connected to every node in the subsequent layer.

The mathematical model that is developed for generating the MLP model is derived as follows, which is made use of in the simulations to arrive at the accurate modelled results. The developed model has various sub-models such as the Model Architecture, Forward Propagation, Hidden Layer Computation, Output Layer Computation, Loss Function, Backpropagation algorithm, Output Layer Gradients, Hidden Layer Gradients, Weight Update.

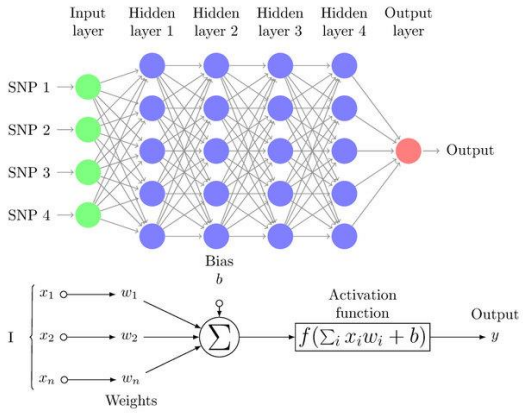

Fig. 2. MLP diagram with four hidden layers and a collection of single nucleotide polymorphisms (SNPs) as input and illustrates a basic "neuron" with 4 inputs & a single output i = 1 to 4).

Model Architecture – begins with the typical input, output & the hidden layers, here, we have taken 4 hidden layers

- Input Layer: The input layer receives the input data. Let x=[x_1, x_2,..., x_n] be the input feature vector of size n.
- Hidden Layers: There are one or more hidden layers, each with a specified number of neurons. Let $h^i = [h_1^l, h_2^l, h_3^l, \dots\dots, h_m^l]$ be the vector of activations for the l-th hidden layer with m_i neurons.
- Output Layer: The output layer produces the final output. For binary classification, it has one neuron with a sigmoid activation function to output the probability of the positive class.

Forward Propagation – This involves calculating the output of each layer given the inputs and the weights.

- Hidden Layer Computation - For the l-th hidden layer, the activations h^l are computed as follows

$$\mathbf{h}^l = \sigma(\mathbf{W}^l \mathbf{h}^{l-1} + \mathbf{b}^l)$$

where

- W^l is the weight matrix connecting the (l−1)-th layer to the l-th layer.
- h^{l-1} is the activation vector from the (l−1)-th layer (for the first hidden layer, $h^0 = x$).
- b^l is the bias vector for the l-th layer.
- σ is the activation function (commonly the ReLU function for hidden layers).

Output Layer Computation - The output layer produces the final prediction $\hat{y}$ as

$$\hat{y} = \sigma(\mathbf{W}^o \mathbf{h}^L + \mathbf{b}^o)$$

where:

- W_o is the weight matrix connecting the last hidden layer h^L to the output layer.
- b^o is the bias for the output layer.
- σ is the sigmoid function defined as $\sigma_z = \frac{1}{1+e^{-z}}$

Loss Function - The loss function quantifies the difference between the predicted output $\hat{y}$ and the true label y. For binary classification, we typically use the binary cross-entropy loss modelled as

$$\mathcal{L}(y, \hat{y}) = -[y \log(\hat{y}) + (1-y) \log(1-\hat{y})]$$

Backpropagation - Backpropagation involves calculating the gradient of the loss function with respect to each weight in the network and updating the weights using gradient descent, which consists of various sub-steps such as Output Layer computations, gradient of the loss with respect to the weights, Hidden Layer Gradients & the weight updates, in turn each one is modelled as follows.

Output Layer Gradients is modelled as

$$\delta^o = \hat{y} - y$$

where y is the output & $\hat{y}$ is the predicted output. Then, the gradient of the loss with respect to the weights and biases in the output layer as

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^o} = \delta^o (\mathbf{h}^L)^T, \quad \frac{\partial \mathcal{L}}{\partial \mathbf{b}^o} = \delta^o$$

Hidden Layer Gradients - For each hidden layer l, the gradients are computed as follows

$$\delta^l = (\mathbf{W}^{l+1})^T \delta^{l+1} \odot \sigma'(\mathbf{h}^l)$$

where the parameters given in the above math model

- ⊙ denotes element-wise multiplication.
- σ′ is the derivative of the activation function.

The gradients with respect to the weights and biases in the l-th hidden layer as

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^l} = \delta^l (\mathbf{h}^{l-1})^T, \quad \frac{\partial \mathcal{L}}{\partial \mathbf{b}^l} = \delta^l$$

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

Weight Update - Using the gradients computed, the weights and biases are updated as

$$\mathbf{W}^l \leftarrow \mathbf{W}^l - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{W}^l}, \quad \mathbf{b}^l \leftarrow \mathbf{b}^l - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{b}^l}$$

All the parameters mentioned in the above equations are being used in the simulations while writing the codes to arrive at the accurate results. In this hybridized CKD prediction model, the MLP is utilized for its capability to capture complex patterns and relationships within the dataset. It serves as the initial classifier, processing input data and generating intermediate predictions that feed into subsequent models. The MLP's architecture, forward propagation, loss function, and backpropagation steps are critical for training the network and making accurate predictions, thereby enhancing the performance of the hybrid model.

B. Part – ii

Stochastic Gradient Descent (SGD) model is employed for its efficiency in optimizing the parameters of the neural network. By iteratively adjusting the weights of the MLP based on the gradient of the loss function, SGD ensures that the model converges to an optimal solution as we use the optimization process. The use of SGD enhances the training process of the MLP, making it faster and more effective in handling large datasets. We have used the iterative optimization technique in ML that aims to find the optimal model parameters by minimizing a cost function. The primary goal of gradient descent was to achieve maximum accuracy on both training and test datasets. This method involves calculating the gradient, a vector indicating the direction of the steepest ascent of the function at a given point.

By moving in the opposite direction of the gradient, the algorithm progressively descends towards lower values of the function, iterating this process until it reaches the function's minimum. This enables the adjustment of model parameters to effectively reduce the cost function, thus enhancing the model's performance and accuracy. As the Stochastic Gradient Descent is a probabilistic approximation of the Gradient Descent, at each step, the algorithm calculates the gradient for one observation picked at random, instead of calculating the gradient for the entire dataset. The mathematical model developed is shown below with the parameters of the learning rates & the observations at each step of iterations. The flow-chart that is being developed to take care of this SGD process is shown below.

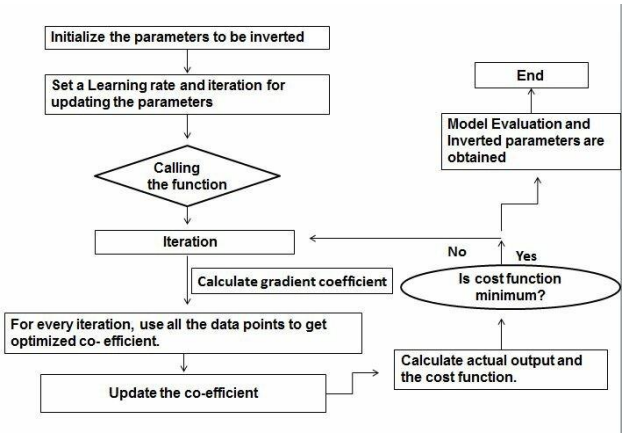

Fig. 3. Flow-chart developed for SGD implementation

Initialization - Begin by initializing the parameters (weights and biases) of the MLP. Let's denote the parameters as *w* and the bias as *b*.

$$\mathbf{w} = \mathbf{w}_0 \ \& \ b = b_0$$

Cost Function - Define the cost function $J(\mathbf{w},b)$ that we aim to minimize. For simplicity, assume a mean squared error (MSE) cost function:

$$J(\mathbf{w}, b) = \frac{1}{N} \sum_{i=1}^{N} (h_{\mathbf{w},b}(\mathbf{x}_i) - y_i)^2$$

where $h_{w,b}(x_i)$ is the prediction of the MLP for input x_i & y_i is the actual label.

Gradient Calculation - At each iteration, calculate the gradient of the cost function with respect to the parameters. For SGD, this gradient is computed for a single randomly selected observation (x_i,y_i) as

$$\nabla_{\mathbf{w}} J(\mathbf{w}, b) = \frac{\partial J(\mathbf{w}, b)}{\partial \mathbf{w}} = \frac{2}{N} (h_{\mathbf{w},b}(\mathbf{x}_i) - y_i)\mathbf{x}_i$$

$$\nabla_b J(\mathbf{w}, b) = \frac{\partial J(\mathbf{w}, b)}{\partial b} = \frac{2}{N} (h_{\mathbf{w},b}(\mathbf{x}_i) - y_i)$$

Parameter Update Rule - Update the parameters using the gradients computed. The learning rate η controls the size of the step taken in the direction of the negative gradient as given by

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla_{\mathbf{w}} J(\mathbf{w}, b) \ \& \ \ b \leftarrow b - \eta \nabla_b J(\mathbf{w}, b)$$

Iterative Process - Repeat the following steps until convergence or for a predefined number of iterations for randomly shuffle the training data & for each training example (x_i , y_i). Compute the gradient $\nabla wJ(w,b)$ and $\nabla bJ(w,b)$, finally, update the parameters w and *b* using the parameter update rule.

Mathematical Representation of SGD – which can be formally represented as

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \frac{\partial J(\mathbf{w}^{(t)}, b^{(t)})}{\partial \mathbf{w}}$$

&

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

$$b^{(t+1)} = b^{(t)} - \eta \frac{\partial J(\mathbf{w}^{(t)}, b^{(t)})}{\partial b}$$

where *t* denotes the iteration step.

Convergence and Learning Rate - The convergence of SGD depends on the choice of the learning rate η. If η is too large, the algorithm may oscillate and fail to converge. If η is too small, the algorithm may converge too slowly or get stuck in a local minimum.

All the parameters mentioned in the derived equations are being used in the simulations while writing the codes to arrive at the accurate results.

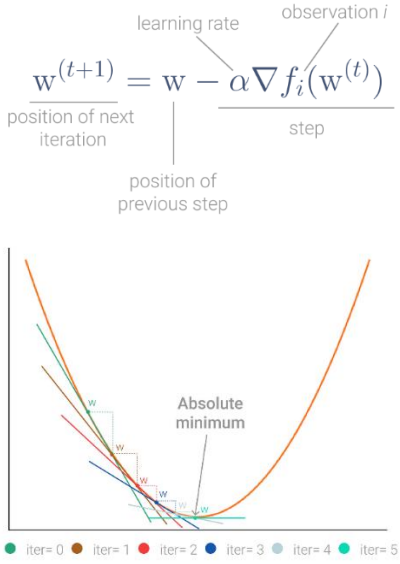

Fig. 4. Use of SGD to adjust the model parameters by minimizing the cost function & improving the accuracy on both training and test datasets

Use of SGD to adjust the model parameters by minimizing the cost function & improving the accuracy on both training and test datasets

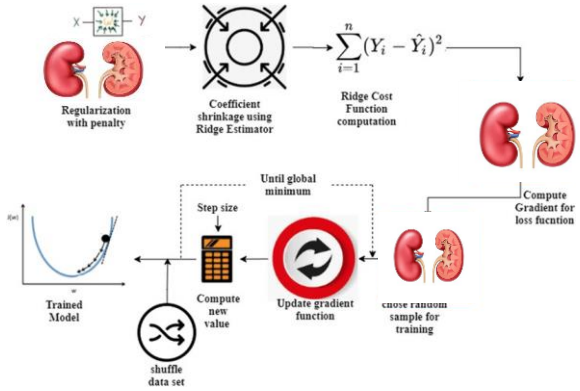

Fig. 5. Block-diagram model to solve the process of Stochastic Gradient Descent (SGD)

C. Part – iii

Adaptive Boosting (Adaboost) is an ensemble technique that combines the predictions of multiple weak classifiers to form a strong classifier. In this hybrid model, Adaboost is used to boost the performance of the MLP and SGD models. By iteratively adjusting the weights of misclassified instances, Adaboost focuses on difficult cases, improving the overall prediction accuracy of the model. Adaptive Boosting (Adaboost) is used in the hybrid model to enhance the performance of the Multilayer Perceptron (MLP) and Stochastic Gradient Descent (SGD) models. Adaboost works by combining the predictions of multiple weak classifiers to create a strong classifier. It iteratively adjusts the weights of misclassified instances, which means it places more emphasis on difficult cases in subsequent iterations. This focus on harder-to-classify instances helps improve the overall prediction accuracy of the model, making Adaboost a valuable component in the hybrid approach for detecting chronic kidney disease.

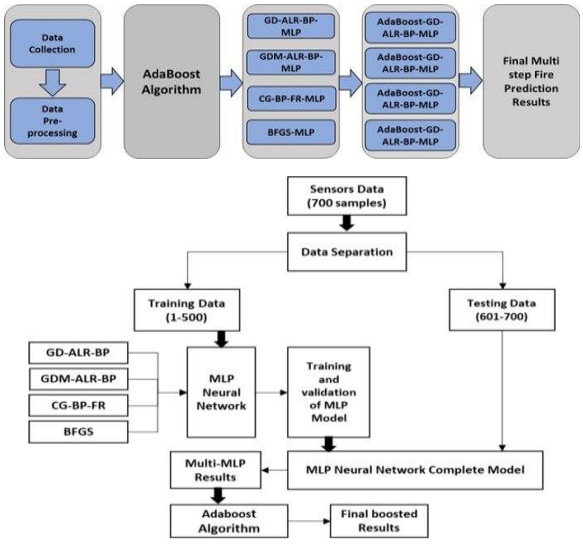

Fig. 6. Hybrid Adaboost MLP model developed & its Training process of Adaboost MLP model

The mathematical model of the Adaptive Boosting (Adaboost), which is used in our work is an ensemble technique designed to improve the performance of weak classifiers by combining their predictions to form a strong classifier, which involves various steps as follows.

Initialization process - Start with a dataset $\{(x_1,y_1),(x_2,y_2),\ldots\ldots,(x_N,y_N)\}$, where x_i represents the input features and y_i represents the binary class labels ($y_i \in \{0,1\}$).

Initialize the weights for each instance - $w_i(1)=1/N$ for all i, where N is the total number of instances.

Training Weak Classifiers - For each iteration $t = 1,2,\ldots,\boldsymbol{T}$ (where $\boldsymbol{T}$ is the total number of iterations)

Train a weak classifier $h_t(x)$ using the weighted dataset. This weak classifier can be an MLP, SGD, or any other simple model.

Calculate the weighted error ϵ_t

$$\epsilon_t = \sum_{i=1}^{N} w_i^{(t)} \mathbb{I}(h_t(x_i) \neq y_i)$$

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

where $\prod$ is the indicator function that equals 1 if $h_t(x_i) \neq y_i$ and 0 otherwise.

Compute Classifier Weight - Calculate the weight α_t of the weak classifier h_t as

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$$

Update Weights - Update the weights for the next iteration

$$w_i^{(t+1)}$$

as

$$w_i^{(t+1)} = w_i^{(t)} \exp(-\alpha_t y_i h_t(x_i))$$

Normalize the weights as

$$w_i^{(t+1)} = \frac{w_i^{(t+1)}}{\sum_{j=1}^{N} w_j^{(t+1)}}$$

This step increases the weights of misclassified instances, making the model focus more on difficult cases in subsequent iterations.

Final Strong Classifier - The final strong classifier $H(x)$ is a weighted majority vote of the T weak classifiers as

$$H(x) = \text{sign} \left(\sum_{t=1}^{T} \alpha_t h_t(x) \right)$$

This combines the predictions of all weak classifiers, with more accurate classifiers having higher weights. All the parameters mentioned in the above equations are being used in the simulations while writing the codes to arrive at the accurate results. In the hybrid CKD prediction model, Adaboost is used to enhance the performance of MLP and SGD models by iteratively adjusting the weights of misclassified instances. By focusing on harder-to-classify cases, Adaboost ensures that the combined model pays more attention to difficult examples, thereby improving the overall prediction accuracy. The weak classifiers (MLP and SGD) are trained on weighted datasets, and their predictions are aggregated to form a strong classifier. This approach leverages the strengths of both MLP and SGD, while the iterative weight adjustment mechanism of Adaboost ensures that challenging cases receive the necessary emphasis, making the model robust and effective for detecting chronic kidney diseases.

D. Part - iv

Logistic Regression is a statistical method for binary classification that models the probability of a binary outcome based on one or more predictor variables. In this hybrid model, LR is used to provide a simple yet powerful baseline classifier. Its probabilistic framework complements the more complex models, ensuring that the final prediction is robust and interpretable. LR, which is a statistical method for binary classification, models the probability of a binary outcome based on predictor variables, providing a straight-forward yet effective baseline classifier in our hybridized CKD prediction model.

This model complements the more complex classifiers, such as Multilayer Perceptron (MLP), Stochastic Gradient Descent (SGD), and Adaptive Boosting (Adaboost), by offering a probabilistic framework that enhances the robustness and interpretability of the final predictions. Logistic Regression's inclusion ensures that the model is not only accurate but also transparent, making it easier to understand and trust the detection of chronic kidney diseases. This approach combines the strengths of various models to deliver a comprehensive and reliable prediction system for chronic renal diseases. The following flow-chart / block-diagram is used for the prediction & detection process of the renal disease in humans.

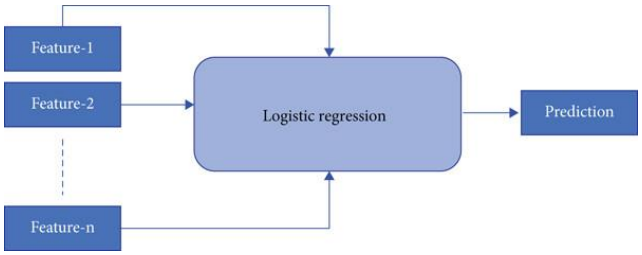

Fig. 7. Overall block diagram

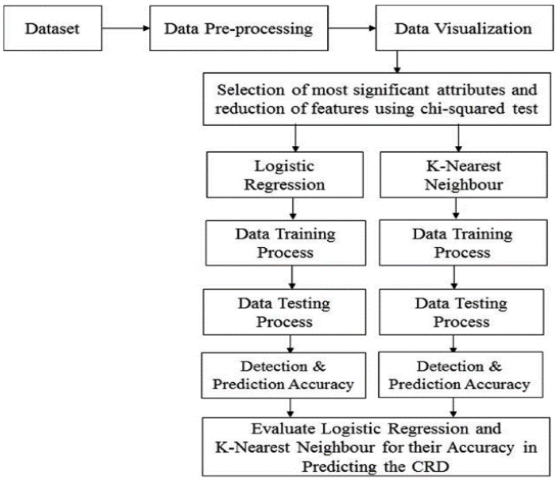

Fig. 8. CRDP : Chronic Renal Disease Prediction and Evaluation with Reduced Prominent Features

The mathematical model developed consists of various steps such as Model Assumption, Linear Combination, Log-Likelihood Function, Cost Function, Gradient Descent, which are derived as follows.

Model Assumption - In logistic regression, we assume that the probability p of the binary outcome y (where y is either 0 or 1) can be modeled using a logistic function (sigmoid function) of a linear combination of predictor variables. The logistic function is defined as

$$p(y = 1|\mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^T \mathbf{x})$$

$$p(y = 0|\mathbf{x}; \mathbf{w}) = 1 - \sigma(\mathbf{w}^T \mathbf{x})$$

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

where

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

is the sigmoid function, x is the vector of predictor variables, and w is the vector of weights (coefficients) that logistic regression estimates.

Linear Combination - The linear combination of the predictor variables x weighted by w is modelled as

$$\mathbf{w}^T\mathbf{x} = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n$$

where w_0 is the intercept term and $w_1,w_2,\ldots\ldots,w_n$ are the coefficients associated with each predictor variable $x_1,x_2,\ldots,x_n$.

The probability p (y = 1|x ; w) that the outcome y is 1 given predictor variables x and model parameters w is:

$$p(y=0|\mathbf{x};\mathbf{w}) = 1 - p(y=1|\mathbf{x};\mathbf{w}) = 1 - \sigma(\mathbf{w}^T\mathbf{x})$$

Similarly, the probability $p(y = 0|x ; w)$ that the outcome y is 0 is modelled as

$$p(y=1|\mathbf{x};\mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}$$

Log-Likelihood Function - The log-likelihood function $L(w)$ for logistic regression, which is maximized during model training, is modelled as

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} \left[y_i \log(\sigma(\mathbf{w}^T\mathbf{x}_i)) + (1-y_i)\log(1-\sigma(\mathbf{w}^T\mathbf{x}_i)) \right]$$

where N is the number of samples, x_i is the predictor vector for the i^{-th} sample, and y_i is the corresponding binary outcome.

Cost Function - The cost function $J(w)$, which is minimized during training, is the negative log-likelihood given as

$$J(\mathbf{w}) = -\mathcal{L}(\mathbf{w}) = -\sum_{i=1}^{N} \left[y_i \log(\sigma(\mathbf{w}^T\mathbf{x}_i)) + (1-y_i)\log(1-\sigma(\mathbf{w}^T\mathbf{x}_i)) \right]$$

Gradient Descent - Logistic regression parameters w can be optimized using gradient descent or other optimization techniques to minimize the cost function $J(w)$.

Logistic Regression (LR) finally is being utilized as a statistical method for the binary classification, where the goal is to predict a binary outcome (typically coded as 0 or 1) based on one or more predictor variables. Here, we have derived the mathematical model for Logistic Regression, which will be used as a baseline classifier in the hybrid model for predicting chronic renal diseases as the LR provides a probabilistic framework for binary classification by modeling the probability of the outcome based on predictor variables. In the context of the hybrid CKD prediction model, LR serves as a transparent and interpretable baseline classifier, complementing more complex models like MLP, SGD, and Adaptive Boosting (Adaboost) algorithm. Its inclusion ensures robust and understandable predictions, contributing to the overall reliability and effectiveness of the model in detecting chronic renal diseases.

E. Part - v

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification & this added feature is incorporated in our work for the predication and detection process. In the hybrid model, Random Forest serves as the meta-classifier, aggregating the predictions from MLP, SGD, Adaboost, and LR models. Its ability to handle overfitting and provide high accuracy makes it an ideal choice for the final decision-making layer as an accurate outcome. Random Forest, an ensemble learning method that builds multiple decision trees and outputs the mode of the classes for classification, serves as the meta-classifier in our hybridized CKD prediction model.

By aggregating the predictions from the Multilayer Perceptron (MLP), Stochastic Gradient Descent (SGD), Adaptive Boosting (Adaboost), and Logistic Regression models, Random Forest finally combines their strengths to improve overall accuracy and robustness. Its ability to handle overfitting while maintaining high accuracy makes it an ideal final decision-making layer in the hybrid model, ensuring reliable and precise detection of chronic renal diseases. This comprehensive approach leverages the diverse capabilities of individual classifiers to create a powerful prediction system for chronic kidney disease.

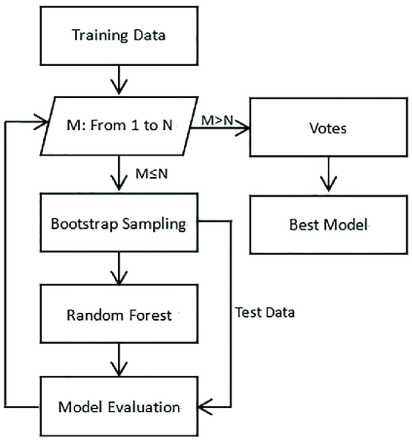

Fig. 9. Flow-chart used by the RF algo for solving our regression & classification problem

When using the Random Forest Algorithm to solve regression problems, we use the mean squared error (MSE) for our data branches which arrives from each node. This remodified formula calculates the distance of each node from the predicted actual value, helping to decide which branch is the better decision for your forest. Here, y_i is the value of the data point we are testing at a certain node and f_i is the value returned by the decision tree, which is mathematically modelled as

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2$$

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

where N is the no. of data points, f_i is the value returned by the model & y_i is the actual value for the data point i.

When performing Random Forests Procedure based on classification data, it is stressed that we are using the modified Gini index formula to decide how nodes on a decision tree branch are matched.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

This formula uses the class and probability to determine the Gini of each branch on a node, determining which of the branches is more likely to occur. Here, p_i represents the relative frequency of the class we are observing in the dataset and c represents the number of classes. Also, the entropy to determine how nodes branch in a decision tree is used as.

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

Entropy uses the probability of a certain outcome in order to make a decision on how the node should branch & it is more mathematical intensive due to the logarithmic function used in calculating it.

VII. Proposed Algorithm Steps

The proposed algorithm's 10-steps for prediction & detection of the chronic disease could be split up into various sub-steps as Data Preprocessing, Feature Selection, Handle Imbalanced Data, Model Training, Multimedia Data Learning, Adaptive Boosting (Adaboost), Clinical Prediction Models (CPMs), Model Evaluation and Validation, Deployment and Monitoring, Prediction & Detection, which are explained as follows one by one.

Step 1: Data Preprocessing

1. Data Collection
   - Collect clinical, demographic, and multimedia data related to CKD from multiple sources, including electronic health records and medical images.
2. Data Cleaning
   - Handle missing values by imputation or removal.
   - Normalize continuous variables and encode categorical variables.
3. Feature Engineering
   - Extract relevant features from raw data.
   - Apply transformations if necessary, such as logarithmic or polynomial transformations.

Step 2: Feature Selection

1. ANOVA (Analysis of Variance)
   - Conduct ANOVA to identify features that significantly differ between CKD and non-CKD groups.
2. Pearson Correlation
   - Calculate Pearson correlation coefficients to measure the linear relationship between continuous variables and CKD status.
3. Cramer's V Test
   - Use Cramer's V test to assess the association strength between categorical variables and CKD status.
4. Select Important Features
   - Combine the results from ANOVA, Pearson correlations, and Cramer's V tests to select the most relevant features for the model.

Step 3: Handle Imbalanced Data

1. Multilayer Perceptron (MLP)
   - Use oversampling, under-sampling, or synthetic data generation techniques (e.g., SMOTE) to balance the class distribution in the dataset.
   - Train an MLP to handle imbalanced data, ensuring it accurately predicts both CKD and non-CKD cases.

Step 4: Model Training

1. Multilayer Perceptron (MLP)
   - Train an MLP on the preprocessed and balanced dataset to capture complex patterns in the data.
2. Stochastic Gradient Descent (SGD) Model
   - Train an SGD model to optimize the weights and biases of the neural network efficiently.
3. Support Vector Machine (SVM)
   - Train an SVM classifier to find the optimal hyperplane that separates CKD and non-CKD cases.
4. Random Forest (RF)
   - Train a Random Forest model to leverage the power of ensemble learning and improve prediction accuracy.
5. Logistic Regression
   - Train a Logistic Regression model to provide a probabilistic framework and baseline classifier.

Step 5: Multimedia Data Learning

1. Deep Stacked Autoencoder Networks
   - Use deep stacked autoencoder networks to learn features from multimedia data (e.g., medical images).
   - Integrate these high-level features with traditional clinical data.

Step 6: Adaptive Boosting (Adaboost)

1. Initialize Weights
   - Initialize weights for each instance in the dataset.
2. Train Weak Classifiers

- Iteratively train weak classifiers (MLP, SGD, SVM, RF, Logistic Regression) on the weighted dataset.
- Calculate the weighted error for each classifier and update instance weights based on misclassifications.

3. Combine Weak Classifiers
   - Use Adaboost to combine the predictions of multiple weak classifiers, forming a strong classifier.
   - Adjust weights iteratively to focus on harder-to-classify instances.

Step 7: Clinical Prediction Models (CPMs)

1. Integrate Clinical and Demographic Data
   - Develop CPMs using clinical and demographic factors to predict CKD outcomes.
   - Ensure CPMs are clinically relevant and provide comprehensive risk assessments.

Step 8: Model Evaluation and Validation

1. Train-Test Split
   - Split the dataset into training and testing sets.
2. Cross-Validation
   - Perform cross-validation to assess the model's performance and generalizability.
3. Evaluation Metrics
   - Evaluate the model using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
4. Compare Models
   - Compare the performance of individual models and the hybrid model.
   - Fine-tune hyperparameters to optimize model performance.

Step 9: Deployment and Monitoring

1. Model Deployment
   - Deploy the hybrid CKD prediction model in a clinical setting.
2. Monitoring and Maintenance
   - Continuously monitor the model's performance.
   - Update the model with new data to maintain its accuracy and relevance.

Step 10: Detection of the renal disease

1. Accurate early detection

The proposed algorithm developed gives a comprehensive approach to develop a novel hybridized CKD prediction model by leveraging ML learning techniques & clinical prediction models. The use of Adaboost enhances the performance of base classifiers, while feature selection methods ensure the inclusion of the most relevant predictors. Handling imbalanced data and incorporating multimedia learning further improve the model's robustness and accuracy, making it a valuable tool for early detection and management of chronic kidney disease.

VIII. Final Outcome of the Proposed Work

The implementation of this novel hybridized CKD prediction model involves several steps. First, the CKD dataset is collected and preprocessed, including data cleaning and feature selection. The dataset is then split into training and testing subsets. The MLP model is trained using SGD, and its predictions are enhanced through Adaboost. Logistic regression and random forest models are also trained independently on the same dataset. Then, the predictions from all models are combined using the Random Forest meta-classifier, which produces the final prediction for CKD detection. By integrating these diverse models, the hybridized CKD prediction model leverages the strengths of each technique, resulting in improved accuracy and robustness in detecting chronic renal diseases.

This approach demonstrates the potential of ML in enhancing healthcare diagnostics, particularly for conditions like CKD where early and accurate detection is critical. The novel hybridized CKD prediction model integrates the above methodologies, utilizing Logistic Regression as a baseline to ensure interpretability. SVM and Random Forest provide robust classification capabilities, while DL techniques handle complex patterns in multimedia data. Boosting techniques enhance model accuracy, and CPMs ensure clinical relevance. Feature selection methods streamline the input data, focusing on the most predictive variables. The combined approach leverages the strengths of each method, delivering a comprehensive and reliable system for CKD detection and prediction.

IX. Conclusion

The research problem of developing and building a novel hybridized CKD prediction model for the detection of chronic renal diseases has been effectively addressed by integrating a combination of advanced machine learning techniques and methodologies. The hybrid model, which includes Multilayer Perceptron (MLP), Stochastic Gradient Descent (SGD), Adaptive Boosting (Adaboost), Logistic Regression, and Random Forest, leverages the strengths of each algorithm to enhance diagnostic accuracy and reliability. The Multilayer Perceptron (MLP), optimized using Stochastic Gradient Descent (SGD), serves as a robust initial classifier, capable of capturing complex patterns within the dataset. The use of SGD ensures efficient training by iteratively adjusting model parameters to minimize the cost function. Adaptive Boosting (Adaboost) further enhances the performance by focusing on difficult cases, combining multiple weak classifiers to form a strong classifier, thus improving the overall prediction accuracy.

Incorporating Logistic Regression provides a probabilistic framework that complements the more complex models, ensuring transparency and interpretability. Random Forest, with its ability to handle high-dimensional data and prevent

overfitting, adds further robustness to the model. Feature selection methods, such as ANOVA, Pearson correlations, and Cramer's V tests, have been utilized to identify the most relevant features, thereby reducing dimensionality and enhancing predictive power. Addressing the imbalanced nature of medical datasets through techniques like oversampling, under-sampling, and synthetic data generation has ensured that the model is adept at handling real-world data. The inclusion of multimedia data learning using deep stacked autoencoder networks has expanded the model's capability to process and interpret complex medical images and signals, providing a comprehensive approach to CKD diagnosis.

The use of Clinical Prediction Models (CPMs) has provided a clinical context to the predictions, making the model's output more relevant and actionable in real-world medical settings. The combination of boosted classifiers and advanced feature selection methods has fine-tuned the model, resulting in improved sensitivity and specificity for CKD diagnosis. In conclusion, the hybridized CKD prediction model developed through this research represents a significant advancement in medical diagnostics. By effectively combining multiple ML techniques and addressing key challenges such as data imbalance and feature selection, the proposed system offers a reliable, accurate, and interpretable tool for early detection of chronic kidney disease.

This comprehensive approach not only enhances diagnostic accuracy but also provides a framework that can be adapted to other complex medical prediction tasks, contributing to the broader field of medical data science. In conclusion, the hybridized CKD prediction model developed through this research represents a significant advancement in medical diagnostics. By effectively combining multiple machine learning techniques and addressing key challenges such as data imbalance and feature selection, the proposed system offers a reliable, accurate, and interpretable tool for early detection of chronic kidney disease. This comprehensive approach not only enhances diagnostic accuracy but also provides a framework that can be adapted to other complex medical prediction tasks, contributing to the broader field of medical data science.

REFERENCES

[1]. Dr Saravanakumar, Eswari, Sampath, Lavanya 2015 "Predictive Methodology for Diabetic Data Analysis in Big Data," *Elsevier, ISBCC*.

[2]. Stephanie Revels, Sathish A.P. Kumar and Ofir Ben-Assuli, 2017 "Predicting Obesity Rate and Obesity-Related Healthcare Costs using Data Analytics", *Health Policy & Tech.*, http://dx.doi.org/10.1016/j.hlpt.

[3]. Vijayalakshmi N, UmaMaheswari M., August2016) "Data mining to elicit predominant factors causing infertility in women", *IJCSMC*, Vol. 5, Issue. 8.

[4]. Min Chen, YixueHao, Kai Hwang, Lu Wang and LigWang, 2017, "Disease prediction by machine learning over big data from Healthcare communities", *IEEE Access*.

[5]. Yu, C.Y.; Li, X.X.; Yang, H.; Li, Y.H.; Xue, W.W.; Chen, Y.Z.; Tao, L.; Zhu, F., 2018 ,"Assessing the Performances of Protein Function Prediction Algorithms from the Perspectives of Identification Accuracy and False Discovery Rate",*Int. J. Mol. Sci*..

[6]. Menzies, N.A.; Wolf, E.; Connors, D., 2018, "Progression from latent infection to active disease in dynamic tuberculosis transmission models: A systematic review of the validity of modelling assumptions", *LancetInfect. Dis*..

[7]. Sindhuja, R. JeminaPriyadarsini, May 2016 "A Survey on Classification Techniques in Data Mining for Analyzing Liver Disease Disorder", *International Journal of Computer Science and Mobile Computing*, Vol.5, Issue.5, ISSN 2320–088X.

[8]. Rifat Hossaina, *et.al.*, 2018, "PRMT: Predicting Risk Factor of Obesity among Middle-Aged People Using Data Mining Techniques*", Procedia Computer Science*, vol. 132, pp. 1068–1076.

[9]. Subas Neupane*et.al.*, "Overweight and obesity among women: analysis of demographic and health survey data from 32 Sub-Saharan African Countries" DOI 10.1186/s12889-016-2698-5, 2016.

[10]. Simi M.S.*et.al* ,2017, "Exploring Female Infertility Using PredictiveAnalytic", *IEEE*.

[11]. Cheong Kim *et.al* 2019,"Predicting Factors Affecting Adolescent Obesity Using General Bayesian Network and What-If Analysis", *Int. J. Environ. Res. Public Health*, vol. 16, pp. 4684..

[12]. P. Suresh Kumar, S. Pranavi 2017 ", "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics",*International Conference on Infocom Technologies and Unmanned Systems* (ICTUS'2017),Dec. 18-20, ADET.

[13]. Sajida Perveen, Muhammad Shahbaz, Karim Keshavjee 2018, "Metabolic Syndrome and Development of Diabetes Mellitus: Predictive Modeling Based on MachineLearning Techniques", pp. 2169-3536, IEEE.

[14]. SundusAbrar, Chu KiongLoo *et.al.* ACCESS.2021, "A Multi-Agent Approach for Personalized Hypertension Risk Prediction",*Digital Object Identifier* 10.1109/.3074791

[15]. Dinu A.J., Ganesan R., Felix Joseph and Balaji V,2017,"A study on Deep Machine Learning Algorithms for diagnosis of diseases", *International Journal of Applied Engineering Research*, ISSN 0973-4562, Volume 12, Number 17.

[16]. Ajad Patel, Sonali Gandhi, SwethaShetty,Prof. BhanuTekwani Jan -2017, "Heart Disease Prediction Using Data Mining", *IRJET*, Vol. 4, Issue1.

[17]. https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/

[18]. (https://towardsdatascience.com/logistic regression, ANN

[19]. https://en.wikipedia.org/wiki/Exploratory_data_analysis

[20]. DrSaravanakumar, Eswari, Sampath, Lavanya "Predictive Methodology for Diabetic Data Analysis in Big Data," *Elsevier, ISBCC*, 2015.

[21]. Stephanie Revels, Sathish A.P. Kumar and Ofir Ben-Assuli, "Predicting Obesity Rate and Obesity-Related Healthcare Costs using Data Analytics", *Health Policy & Tech.*, http://dx.doi.org/10.1016/j.hlpt, 2017.

[22]. Vijayalakshmi N, UmaMaheswari M., "Data mining to elicit predominant factors causing infertility in women", *IJCSMC*, Vol. 5, Issue. 8, August2016.

[23]. Min Chen, YixueHao, Kai Hwang, Lu Wang and LigWang, "Disease prediction by machine learning over big data from Healthcare communities", *IEEE Access*,2017.

[24]. Yu, C.Y.; Li, X.X.; Yang, H.; Li, Y.H.; Xue, W.W.; Chen, Y.Z.; Tao, L.; Zhu, F., "Assessing the Performances of Protein Function Prediction Algorithms from the Perspectives of Identification Accuracy and False Discovery Rate",*Int. J. Mol. Sci*., 2018.

[25]. Menzies, N.A.; Wolf, E.; Connors, D., "Progression from latent infection to active disease in dynamic tuberculosis transmission models: A systematic review of the validity of modelling assumptions", *LancetInfect. Dis*., 2018.

[26]. D.Sindhuja, R. JeminaPriyadarsini, "A Survey on Classification Techniques in Data Mining for Analyzing Liver Disease Disorder", *International Journal of Computer Science and Mobile Computing*, Vol.5, Issue.5, ISSN 2320–088X, May 2016.

[27]. RifatHossaina, *et.al.*, "PRMT: Predicting Risk Factor of Obesity among Middle-Aged People Using Data Mining Techniques*", Procedia Computer Science*, vol. 132, pp. 1068–1076, 2018.

[28]. SubasNeupane*et.al.*, "Overweight and obesity among women: analysis of demographic and health survey data from 32 Sub-Saharan African Countries" DOI 10.1186/s12889-016-2698-5, 2016.

[29]. Simi M.S.*et.al.*, "Exploring Female Infertility Using PredictiveAnalytic", *IEEE*, 2017.

[30]. Cheong Kim *et.al.*, "Predicting Factors Affecting Adolescent Obesity Using General Bayesian Network and What-If Analysis", *Int. J. Environ. Res. Public Health*, vol. 16, pp. 4684, 2019.

[31]. P. Suresh Kumar, S. Pranavi", "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics",*International Conference on Infocom Technologies and Unmanned Systems* (ICTUS'2017),Dec. 18-20, 2017, ADET.

[32]. Sajida Perveen, Muhammad Shahbaz, Karim Keshavjee, "Metabolic Syndrome and Development of Diabetes Mellitus: Predictive Modeling Based on MachineLearning Techniques", pp. 2169-3536, IEEE, 2018.

[33]. SundusAbrar, Chu KiongLoo *et.al.*, "A Multi-Agent Approach for Personalized Hypertension Risk Prediction",*Digital Object Identifier* 10.1109/ACCESS.2021.3074791

[34]. Dinu A.J., Ganesan R., Felix Joseph and Balaji V,"A study on Deep Machine Learning Algorithms for diagnosis of diseases", *International Journal of Applied Engineering Research*, ISSN 0973-4562, Volume 12, Number 17. 2017.

[35]. Ajad Patel, Sonali Gandhi, SwethaShetty,Prof. BhanuTekwani, "Heart Disease Prediction Using Data Mining", *IRJET*, Vol. 4, Issue1, Jan -2017.

[36]. Institute of Medicine (US) Forum on microbial threats. vector-borne diseases: understanding the environmental, human health, and ecological connections, workshop summary. Washington, DC, 2008. (doi:10.17226/11950)

[37]. National Academies of Sciences Engineering and Medicine. Global health impacts of vector-borne diseases: workshop summary. Washington, DC, 2016. (doi:10.17226/21792)

[38]. James SL et al. 2018 Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 392, 1789–1858. (doi:10.1016/S0140-6736(18)32279-7)

[39]. Gubler DJ. 1998 Resurgent vector-borne diseases as a global health problem. Emerg. Infect. Dis. 4, 442–450. (doi:10.3201/eid0403.980326)

[40]. Caminade C, McIntyre KM, Jones AE. 2019 Impact of recent and future climate change on vector borne diseases. Ann. N. Y. Acad. Sci. 1436, 157–173. (doi:10.1111/nyas.13950)

[41]. Gray JS, Dautel H, Estrada-Peña A, Kahl O, Lindgren E. 2009 Effects of climate change on ticks and tickborne diseases in Europe. Interdiscip. Perspect. Infect. Dis. 2009, 593232. (doi:10.1155/2009/ 593232)

[42]. Bouchard C, Dibernardo A, Koffi J, Wood H, Leighton PA, Lindsay LR. 2019 Increased risk of tickborne diseases with climate and environmental changes. Can. Commun. Dis. Rep. 45, 83–89. (doi:10.14745/ccdr.v45i04a02)

[43]. Campbell-Lendrum D, Manga L, Bagayoko M, Sommerfeld J. 2015 Climate change and vectorborne diseases: what are the implications for public health research and policy? Phil. Trans. R. Soc. B 370, 1–8. (doi:10.1098/rstb.2013. 0552)

[44]. Semenza J, Lindgren E, Balkanyi L, Espinosa L, Almqvist M, Penttinen P, Rocklöv J. 2016 Determinants and drivers of infectious disease threat events in Europe. Emerg. Infect. Dis. J. 22, 581. (doi:10.3201/eid2204. 151073)

[45]. Rocklöv J, Dubrow R. 2020 Climate change: an enduring challenge for vector-borne disease prevention and control. Nat. Immunol. 21, 479–483. (doi:10.1038/s41590-020-0648-y)

[46]. Sumilo D, Asokliene L, Bormane A, Vasilenko V, Golovljova I, Randolph SE. 2007 Climate change cannot explain the upsurge of tickborne encephalitis in the Baltics. PLoS ONE 2, e500–e500. (doi:10.1371/journal.pone. 0000500)

[47]. Semenza JC, Suk JE. 2018 Vector-borne diseases and climate change: a European perspective. FEMS Microbiol. Lett. 365, fnx244. (doi:10.1093/femsle/ fnx244)

[48]. Fouque F, Reeder JC. 2019 Impact of past and ongoing changes on climate and weather on vector borne diseases transmission: a look at the evidence. Infect. Dis. Poverty 8, 51. (doi:10.1186/s40249-019- 0565-1)

[49]. Negev M, Paz S, Clermont A, Pri-Or NG, Shalom U, Yeger T, Green MS. 2015 Impacts of climate change on vector borne diseases in the mediterranean basin—implications for preparedness and adaptation policy. Int. J. Environ. Res. Public Health 12, 6745–6770. (doi:10.3390/ ijerph120606745)

[50]. Sadeghieh T, Waddell LA, Ng V, Hall A, Sargeant J. 2020 A scoping review of importation and predictive models related to vector-borne diseases, pathogens, reservoirs, or vectors (1999–2016). PLoS ONE 15, e0227678. (doi:10.1371/journal.pone.0227678)

[51]. Kearney M, Porter W. 2009 Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. Ecol. Lett. 12, 334–350. (doi:10.1111/j.1461-0248.2008. 01277.x)

[52]. Paz S. 2015 Climate change impacts on West Nile virus transmission in a global context. Phil. Trans. R. Soc. B 370, 1–11. (doi:10.1098/rstb.2013. 0561)

[53]. Estrada-Peña A, Ayllóon N, Fuente J de la. 2012 Impact of climate trends on tick-borne pathogen transmission. Front. Physiol. 3, 1–12. (doi:10.3389/ fphys.2012.00064)

[54]. Tjaden NB, Caminade C, Beierkuhnlein C, Thomas SM. 2018 Mosquito-borne diseases: advances in modelling climate-change impacts. Trends Parasitol. 34, 227–245. (doi:10.1016/j.pt.2017.11.006)

[55]. Reiner RC et al. 2013 A systematic review of mathematical models of mosquito-borne pathogen transmission: 1970–2010. J. R. Soc. Interface 10, 20120921. (doi:10.1098/rsif.2012.0921)

[56]. Altizer S, Dobson A, Hosseini P, Hudson P, Pascual M, Rohani P. 2006 Seasonality and the dynamics of infectious diseases. Ecol. Lett. 9, 467–484. (doi:10. 1111/j.1461-0248. 2005.00879.x)

[57]. Vogels CBF, Hartemink N, Koenraadt CJM. 2017 Modelling West Nile virus transmission risk in Europe: effect of temperature and mosquito biotypes on the basic reproduction number. Sci. Rep. 7, 5022. (doi:10.1038/s41598-017-05185-4)

[58]. Ziegler U et al. 2019 West Nile virus epizootic in Germany, 2018. Antiviral Res. 162, 39–43. (doi:10. 1016/j.antiviral.2018.12.005)

[59]. Vlaskamp DR et al. 2020 First autochthonous human West Nile virus infections in the Netherlands, July to August 2020. Euro Surveill. 25, 2001904. (doi:10.2807/1560-7917.ES.2020.25.46. 2001904)

[60]. Chancey C, Grinev A, Volkova E, Rios M. 2015 The global ecology and epidemiology of West Nile virus. Biomed Res. Int. 2015, 376230. (doi:10.1155/2015/ 376230)