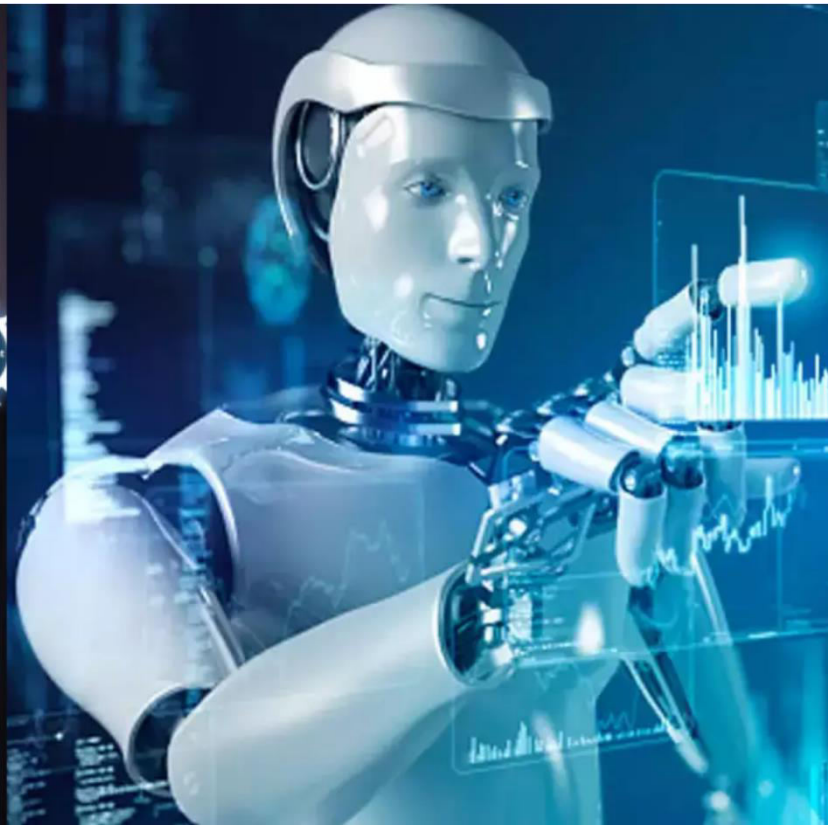
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404008|

A Robust Machine Learning Pipeline for Chronic Kidney Disease Prediction using Outlier Detection, Feature Selection, and Adam Optimization

Ravi Kumar Ravi

Florida League of Cities, Orlando, Florida, USA

ABSTRACT: Data mining technology for healthcare purposes transforms basic medical information into useful insights which helps doctors predict diseases along with diagnosing patients while tailoring personalized treatment plans. A successful healthcare analytics process requires a complete data pipeline which includes data preprocessing followed by extraction then selection after optimization until classification. Data preprocessing implements the Z-Score Method together with Isolation Forest along with Interquartile Range (IQR) Method to detect outliers which would potentially harm model performance. The Principal Component Analysis (PCA) serves as an extraction method to reduce dimensions alongside log transformation which addresses data skewness to establish data stability. Wrapper methods Recursive Feature Elimination (RFE) together with Genetic Algorithms build predictive accuracy by continuously refining the selected feature subsets. Model performance enhancement relies on the Adam optimizer together with other optimization methods that aid in achieving high accuracy during medical outcome predictions. Adam Optimization optimization achieved a maximum training accuracy of 1.0 while validation accuracy reached 0.9 as the optimal value. Improvements in loss functionality decreased it from 45 to almost zero levels proving successful learning takes place. Analysis showed that the model demonstrated a positive trend in both validation accuracy levels despite periodic accuracy difficulties because the pipeline functioned well for predicting CKD conditions. Future work will direct its attention towards adjusting model performance by fine-tuning along with regularization techniques and additional data acquisition for enhancing generalization capabilities.

KEYWORDS: Outlier Detection, Dimensionality Reduction, Feature Selection, Model Optimization, Gradient Boosting, Personalized Medicine.

I. INTRODUCTION

Building predictive models for modern data analysis and machine learning requires following three essential steps named feature extraction followed by feature selection and optimization respectively [1]. The implemented techniques sharpen the model's quality while minimizing overfitting issues to enhance overall generalization potential. Each procedure performs an essential part in transmuting raw data into significant insights and subsequently employs these findings for predictive or decision-making applications. The process of extracting beneficial data features from unprocessed information forms the basis of using features with machine learning algorithms. Feature selection allows experts to pick important features of extracted data which reduces system operations while improving model performance. Adam Optimization serves as an optimization method to enable efficient learning by finding optimal solutions to minimize the error function in the model training period. These different methods enhance their combined capability to develop sophisticated models which resolve intricate real-world tasks.

The first major operation for preparing data before machine learning applications involves the extraction of essential features. The procedure converts basic data into meaningful feature sets which models require for analyzing problems[2]. The basic definition presents data conversion as a process which transforms images and audio alongside text materials into numeric formats suitable for machine learning algorithm processing. Image processing obtains features such as edges and textures and patterns from raw pixels by implementing convolutional neural networks (CNNS). The processing of natural language texts into numeric forms can be achieved through word embeddings combined with bag-of-words models in natural language processing. Unstructured data requires feature extraction to simplify its complexity because the technique helps learning processes succeed while discarding useless data points. Through feature extraction the model gains capability to detect hidden patterns and data relationships that result in

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404008|

improved decision prediction and making performance. A machine learning model requires suitable feature extraction to grasp the data properly otherwise it will perform inadequately.

Next after extracting relevant features stands the procedure of selecting the most important features from existing collection. It is crucial to apply feature selection approaches because adding extra features beyond necessity leads to model overfitting that results in poor outcomes with new data sets[3]. The model absorbs both real relationships from the data and noise which causes it to become complex and non-generalizable in its ability to predict new observations. The removal of unneeded and weak associations in data through feature selection creates a brighter model performance with improved operational efficiency. The different methods used for feature selection include filter methods along with wrapper methods and embedded methods. The process of feature selection uses statistical testing to create a ranking of features through filters. The performance of variable subsets is investigated through model training implemented by wrapper methods while embedded methods perform feature selection during model training as a part of the regularized approach. Model accuracy improves alongside shorter training durations and better data understanding through feature selection because it identifies the most important variables. This procedure supports model overfitting prevention and develops general prediction capabilities for new data instances.

Adam optimization provides a solution to overcome two essential optimization problems in deep learning networks through its addressing of gradient explosion and gradient decay issues. The training process faces obstacles because gradients become either too small or too large through vanishing or exploding respectively. Adam executes gradient moving average calculations to optimize convergence while producing better results than traditional methods because it reduces noise in the gradient optimization path[4]. The learning rates of each parameter are controlled by Adam through automatic adjustment which makes it perform better than traditional gradient descent approaches. Speedup in convergence and stable training procedure are two essential benefits offered by this efficiency especially for deep neural network training operations. Many machine learning practitioners select Adam as their preferred optimizer because it works best with large datasets and complex models particularly for tasks like image recognition and natural language processing.

The integration of feature extraction methods with feature selection algorithms and optimization methods through Adam resolves multiple important issues that emerge in machine learning. The dimensional scale problem creates difficulties for processing high-dimensional datasets because it results in substantial computational expenses that lead to overfitting[5]. Feature extraction techniques together with selection methods decrease the number of features for analysis which maintains key information so models become more efficient and minimize the likelihood of overfitting. Overfitting stands as a significant machine learning challenge that happens when model complexity exceeds acceptable levels so it begins matching random noise instead of the actual data information. Feature selection together with optimization enable effective training of models to attain better generalization for new data by utilizing essential features alone [6]. Through Adam optimization the training speed significantly improves because it tackles slow convergence during times of large-scale operations or complex loss functions and deep neural training. Adam optimizes convergence speed through learning rate adjustments that also address gradient problems to establish stable convergence rates.

The proposed work serves machine learning critically because it delivers an organized solution to process data better and improve learning optimization. Work with large and complex datasets requires proper data extraction methods for selecting features because using raw data leads to insufficient model performance. Optimization algorithms that perform inefficiently create problems with both slow convergence rates and non-convergence situations thus preventing successful training of large datasets and deep architectures. Machine learning performance gets improved through the implementation of both extracted features and selected attributes alongside fast and efficient optimization techniques which include Adam. The proper utilization of these models becomes essential in healthcare because errors at diagnostic stages using sensitive medical data can produce critical results for patients.

Main Contribution of Proposed Work

- Advanced feature extraction methods and selection techniques from the proposed work help maximize model efficiency by both diminishing data dimensionality and enhancing data quality.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404008|

- The training speed gets improved by Adam Optimization which enables swift convergences and enhanced treatment of gradients during learning procedures.
- The proposed method fixes both data overfitting and generalization problems to generate stable machine learning models that deliver higher accuracy on complex database systems.

The review in Section II examines contemporary machine learning tools with an analysis of feature extraction methods and selection algorithms and optimization frameworks alongside their predictive modeling characteristics. In Section III the proposed system architecture shows how data preparation connects to feature choice processes that lead to model development utilizing Adam optimization. The experimental section of the report examines the proposed methodology through performance benchmarks that show comparison against conventional modeling approaches. Section V serves as the study conclusion by presenting a summary of obtained findings alongside recommendations for future research in prediction task pipeline optimization.

II. LITERATURE SURVEY

Mondal et al. (2024) present a study about AI-driven big data analytics for personalized medicine through the combination of federated learning with blockchain and quantum computing technology to boost healthcare systems[8]. Post-Talib Soroushmehr and Kasra Najarian (2016) study how big data evolves into medical and healthcare computational frameworks[9]. The researchers emphasis big data analytics for developing predictive models which produce health outcome predictions through extensive clinical dataset evaluation.

Alyass et al. (2015) analyze the process of shifting big data research toward personalized medicine [10]. The research implementation focuses on uniting different big data technologies which enable the development of customized healthcare solutions. The research by Ahmed et al. (2020) describes a multiple-function platform based on machine learning which seeks to enhance healthcare and precision medicine processes[11]. The methodology implements artificial intelligence together with machine learning algorithms to build programs which support healthcare providers deliver custom-made and exact care to patients.

Ghanem et al. (2024) examine how artificial intelligence transforms personalized medicine through predictive models and AI technologies which enhance care delivery [12]. Through their studies these researchers merge clinical information with AI-analyzed data to establish individualized treatment plans. According to Hu (2019) machine learning applications for personalized medicine have been evaluated for three specific diseases which include coronary heart disease and substance use disorder along with Alzheimer's disease[13]. A machine learning approach to process medical patient information enables the development of disease prediction models according to the dissertation.

Goswami and Sharma (2024) investigate AI and quantum computing and cloud computing integration in personalized medicine fields of next-generation sequencing bioinformatics according to their research[14]. Through integration these technologies create more efficient and accurate analysis of genomic data. The paper by Valero-Ramon et al. (2020) introduces dynamic models based on healthcare sensors along with interactive process mining for personalized chronic disease management[15]. Real-time sensor data processing enables the development of dynamic models to boost the management of chronic diseases through their methodological approach.

III. PROPOSED WORK

The figure 1 displays the multi-stage machine learning pipeline from initial data preparation to feature extraction then data selection before optimization and model evaluation. The initial step of Data Preprocessing includes outlier detection through the implementation of Z-Score Method, Isolation Forest and Interquartile Range (IQR) Method which enables data cleansing before modeling begins. At this point the system implements Feature Extraction through the combination of Principal Component Analysis (PCA) and Log Transformation to minimize data dimensions and fix distribution irregularities. Model building takes place in the Data Selection phase after applying Recursive Feature Elimination (RFE) and Genetic Algorithms to determine the most important features. After model development Adam Optimization serves as the optimization method to achieve efficient training convergence. The model evaluation phase uses different metrics for performance assessment to verify its capacity to work with untested data. The workflow provides a methodical procedure for designing reliable machines that perform effectively as learning systems.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404008|

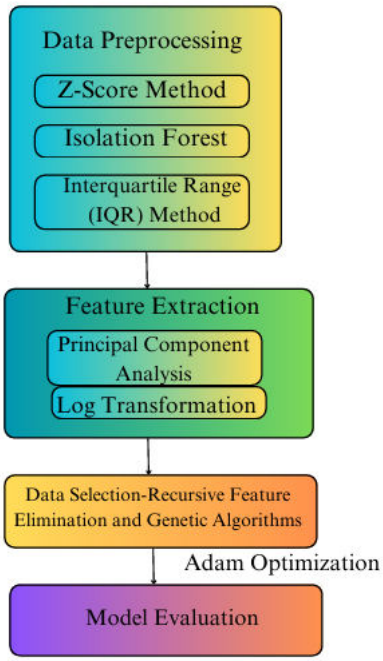


Figure 1: Workflow of Machine Learning Pipeline

A. Data Preprocessing

The preprocessing of data stands as an essential requirement in machine learning procedures plus data analysis work for big datasets containing outlier data points. Outliers represent data points which differ substantially from typical values so they negatively affect both statistical processes and explanations and degrade model functioning. The process of finding and handling these outliers proves essential to improving both predictive model accuracy and reliability rates. The detection of outliers occurs using three popular methods including Z-Score Method together with Isolation Forest and the Interquartile Range (IQR) method.

- **Z-Score Method**

The Z-Score method represents an easy statistical tool that identifies outliers in well-distributed normal datasets. The Z-score function determines the standard deviation distance of a point from its mean value. The calculation of Z-score for individual data points utilizes the following mathematical formula:

$$z(x_i) = \frac{x_i - \mu}{\sigma} \qquad (1)$$

Where:

- x_i is the data point.
- μ is the mean of the data.
- σ is the standard deviation of the data.

A Z-score greater than a certain threshold (commonly $|Z|>3$) indicates that the data point x_i is far from the mean and is considered an outlier. This method is effective when the data is approximately normally distributed, as it assumes that data points far from the mean (in terms of standard deviations) are outliers.

- **Isolation Forest**

The unsupervised learning algorithm called Isolation Forest serves as a detection method specifically designed for large datasets. The core principle of Isolation Forest involves separating outlier elements instead of targeting standard elements. The technique builds Isolation Trees from multiple algorithms then labels any point as an outlier which requires the least splits for its isolation in these trees. The algorithm creates new trees by selecting

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404008|

features randomly followed by random split value choices between minimum to maximum feature values. Points receive an isolation score which equals to:

$$\text{Score}(x) = 2^{-\frac{h(x)}{c(n)}} \qquad (2)$$

Where:

- h(x) is the number of edges required to isolate the point x.
- c(n) is a normalization constant related to the number of data points n.

The method detects outliers through data points which need fewer splits to separate them. Lifetime performance of the Isolation Forest method excels when working with extensive datasets because it successfully identifies outliers which conventional methods like Z-score cannot detect well in high-dimensional environments.

Interquartile Range (IQR) Method

Non-parametric Interquartile Range (IQR) stands as a popular method for outlier detection specifically designed for datasets without normal distribution patterns. The IQR value results from subtracting Q1 (the 25th percentile) from Q3 (the 75th percentile) in a dataset. The IQR calculation provides information about middle 50% spread and enables detection of outliers appearing below Q1−1.5×IQR. The required formula for IQR determination is:

$$IQR = Q3 - Q1 \qquad (3)$$

Where:

- Q1 is the 25th percentile (lower quartile).
- Q3 is the 75th percentile (upper quartile).

Outliers are defined as data points that lie outside the range:

$$\text{Lower Bound} = Q1 - 1.5 \times IQR$$
$$\text{Upper Bound} = Q3 + 1.5 \times IQR$$

Points lying outside the defined upper and lower boundaries get classified as outliers. This data analysis technique proves especially useful on datasets with distribution skews because it lacks requirements about the data distribution model. The system can easily be implemented using efficient computing methods.

B. Data Extraction

PCA functions as a statistical method which transforms high-dimensional data into lower dimensions while keeping original variance levels intact. The technique converts original features into unrelated principal components which become the basis of the new feature set. The principal components represent data dimensions that arrange from high to low variance values to let researchers concentrate on essential elements. PCA starts by deducing the covariance matrix from the data which displays variable pairs interrelations. Eigenvalues along with eigenvectors can be derived from the covariance matrix to obtain principal components. The main step of PCA consists of applying data projection methods using eigenvectors as calculation basis:

$$X' = X.V \qquad (4)$$

Where:

- X is the original data matrix (with nnn samples and ppp features).
- V is the matrix of eigenvectors (principal components).
- X′ is the transformed data matrix in the new space of principal components.

The selection of top k components for dimensional reduction considers the cumulative variance explained up to that point where selecting more components ensures an adequate retention of information. PCA serves tasks which include data visualization while also reducing noise and optimizing model performance through elimination of features presenting redundancy.

Log Transformation

The log transformation provides a solution for identifying deviations in data distributions through its application of logarithmic functions to the existing data points. This procedure yields ideal results in data distributions that range over multiple orders because it decreases high values while spreading out lower values thus transforming data to normal distribution patterns. Log transformation applies to each element of X through the following operation:

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404008|

$$X' = \log(X + 1) \qquad (5)$$

Where:

- X is the original dataset with potentially skewed values.
- X′ is the transformed dataset after applying the log transformation.
- The addition of 1 ensures that zero values are handled appropriately, as the logarithm of zero is undefined.

The log transformation process transforms the data while reducing outlier impact and steering it toward normal distributions that benefit many machine learning algorithms using assumptions of normal data. The technique proves effective for data sets that show exponential growth or present heavy-tailed distributions especially when dealing with financial information and biological information (for instance, gene expression levels).

Combining PCA and Log Transformation

Data preparation for machine learning models includes frequently combining the use of PCA with log transformation. The most important aspects of data reduction through PCA work in synergy with log transformation which improves the data quality by rectifying distributional skewness and normalizing feature distributions. The joint implementation of these techniques delivers robust and efficient models which perform better due to decreased computational complexity and increased generalization ability and data distribution alignment with learning algorithm assumptions. The combination of log transformation with PCA application produces effective results when working with high-dimensional and skewed datasets and improves predictive performance.

C. Data Selection

Data selection stands as a vital procedure for constructing machine learning models. Selecting the most appropriate set of features from a model determines operational efficiency while preventing model overfitting. The evaluation of feature subsets becomes possible through wrapper methods because these evaluate different feature combinations against predictive models to discover valuable inputs. Two effective wrapper methods used for feature selection are Recursive Feature Elimination (RFE) and Genetic Algorithms (GA). Such methods determine an optimal feature selection that optimizes model performance while boosting accuracy rates together with simplifying computational workloads and enhancing interpretability of the models.

Recursive Feature Elimination (RFE)

The iterative feature selection technique RFE removes features in successive steps to construct models for measuring their importance levels. The method begins with training a model across all features and progressively removes each weakest feature in line with its ranking criteria based on model coefficient values or feature importance scores. The process stops when the specified number of features is achieved. RFE adopts the following mathematical structure when implemented:

$$F_k = \mathrm{RFE}(X, y, k) \qquad (6)$$

Where:

- F_k is the set of the k most important features after the elimination process.
- X represents the feature matrix (with all features initially).
- y is the target vector.
- k is the desired number of features to retain.

The procedure includes several rounds where RFE examines a model's evaluation with cross-validation to determine the most important features in order. Each round removes the least essential feature from the analysis leading to kkk features remaining at the end of the process. When used to identify the model-performance related features it works effectively yet proves costly in terms of computation especially with large datasets containing multiple features.

Genetic Algorithms (GA)

Genetic Algorithms (GA) function as optimization methods which draw ideas from natural selection sequences. The implementation of GA within feature selection uses evolutionary simulation methods. A population of distinct feature subset candidates exists while each proposed set receives assessment through a fitness function that normally gauges the model performance during training. The breeding of best-performing genetic pool combinations produces new subsets through crossover operations that are later subjected to mutation then selected for the upcoming generation. The solution proceeds through multiple interations before locating the best feature combination. Mathematically the process appears as:

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404008|

$$S_{t+1} = GA(S_f, f) \quad (7)$$

Where:

- S_f is the set of feature subsets at generation t.
- f is the fitness function evaluating the subsets.
- S_{t+1} is the new set of feature subsets after crossover and mutation.

The genetic algorithm uses the feature subsets within its population to form chromosomes. Each feature subset population receives evaluation through a fitness function to determine its performance with those selected elements. By exploring extensive search areas GA effectively finds optimal solutions for complicated problems requiring evaluation of high-dimensional data. GA demands significant computational resources when applying the technique to large problem sets combining numerous generations as well as considerable population sizes. Proper adjustment of population size along with modification of mutation rate and crossover rate parameters becomes necessary for genetic algorithm success.

D. Optimization

Adam represents an optimization method that serves as one of the leading and effective approaches to train machine learning models especially deep learning models. Adam represents an optimization technique which assimilates Momentum together with RMSProp functionality. Each parameter's learning rate adjustment under this algorithm considers first moment means and second moment variances of gradient data. In consequence of its adaptive properties Adam operates optimally in different tasks thus eliminating the requirement to manually adjust learning rates. Adam performs gradient scaling through running average computation on gradient values and squared values which improves parameter learning rate efficiency and stability. The Adam optimizer uses following steps to update its parameters:

$$m_t = \beta_1 . m_{t-1} + (1 - \beta_1) . g_t \quad (8)$$

$$v_t = \beta_2 . v_{t-1} + (1 - \beta_2) . g_t 2 \quad (9)$$

$$m'_t = \frac{m_t}{1 - \beta_1^t} \quad (10)$$

$$v'_t = \frac{v_t}{1 - \beta_2^t} \quad (11)$$

Where:

- m_t is the first moment estimate (mean of the gradients).
- v_t is the second moment estimate (variance of the gradients).
- β_1 and β_2 are the exponential decay rates for the first and second moments, typically set to 0.9 and 0.999, respectively.
- g_t is the gradient of the loss function at time step ttt.
- m'_t and v'_t are the bias-corrected estimates of the first and second moments.

Adam's key advantage lies in its ability to adapt the learning rate for each parameter, considering both the mean and the variance of past gradients. This results in more stable and faster convergence during training, especially in high-dimensional parameter spaces.

IV. RESULT & DISCUSSION

The Chronic Kidney Disease (CKD) dataset combines hospital data from two months and contains 400 instances and 24 attributes that include both numerical and categorical data points. The dataset functions similarly to a classification model by using age measurements alongside blood pressure data and specific gravity results and albumin readings and red blood cells counts and pus cell results and blood glucose values and several more health factors as evaluation inputs to determine patient CKD status. The features exhibit several continuous variables (blood glucose and urea levels) and categorical variables (showing bacteria results and anemia status and diabetes condition). The dataset carries missing data points and three nominal features such as hypertension and diabetes mellitus and coronary artery disease that require preprocessing methods before successful training of the model becomes possible. A binary classification makes up the target variable because the class label reveals the presence or absence of CKD in patientsckd[7].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404008|

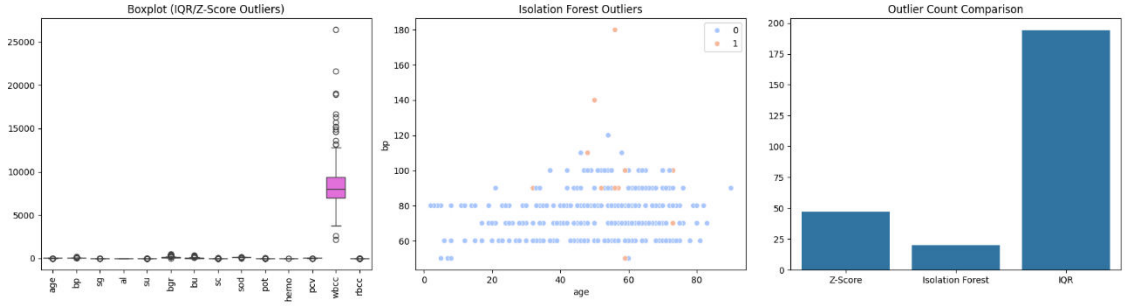


Figure 2: Outlier Detection Comparison

The Chronic Kidney Disease (CKD) dataset merges medical records from two months and uses 400 observations that encompass 24 attributes with numerical and categorical factors throughout the data. The dataset functions like a classification model through the analysis of age measurements with blood pressure data and specific gravity results and albumin readings alongside red blood cells counts and pus cell results and blood glucose values and several more health factors for determining patient CKD status. The presence of blood glucose and urea levels serves as continuous variables while bacteria test results and anemia status and diabetes condition categorize the variables. Most of the dataset consists of empty values while three features with nominal types including hypertension and diabetes mellitus and coronary artery disease demand special preprocessing procedures to allow successful model training. The target variable consists of binary classification which indicates whether patients have CKD or notckd[7].

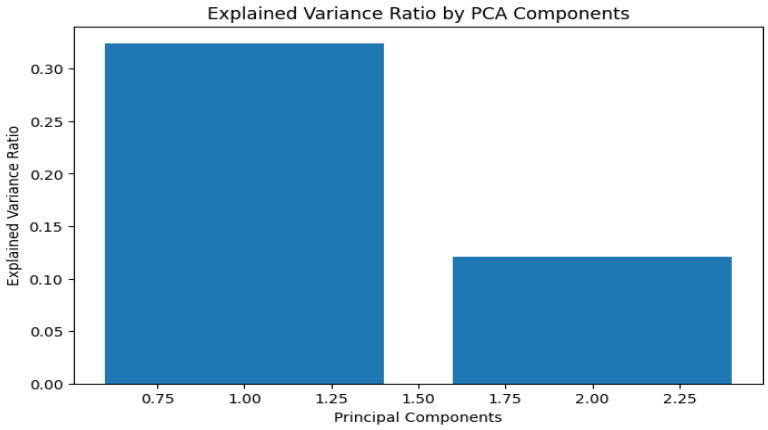


Figure 3: Explained Variance Ratio by PCA Components

The explained variance ratios appear in figure 3 as a result of PCA on the data. The first bar indicates the first principal component maintains about 30% of the overall variance followed by rapid decreases in subsequent bars which represent the following components. The analysis indicates that data variability gets most effectively captured by initial principal components while additional components maintain smaller and smaller proportions. The graphical presentation enables evaluation of retained information per component thus supporting proper component selection for analysis or modeling purposes.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404008|

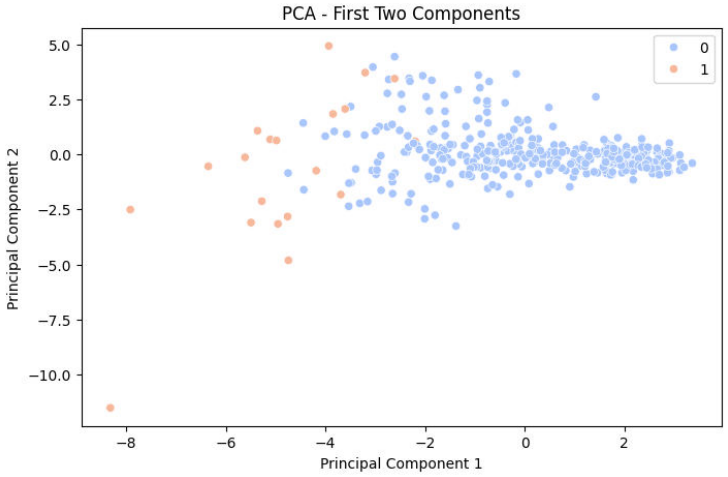


Figure 4: PCA - First Two Components

A scatter plot based on the first two components derived from PCA displays the data according to Figure 4. Observations appear as points in the graph which correspond to Principal Component 1 and Principal Component 2 values at their respective axes. Data labels determine spot colors since blue dots (0) and orange dots (1) fill the dimensional space. The visual depiction shows how PCA separates different classes within a reduced-dimensional framework as Principal Component 1 and Principal Component 2 explain most of the data variance. The scatter plot assists in evaluating the effectiveness of PCA to separate classes through its reduced features.

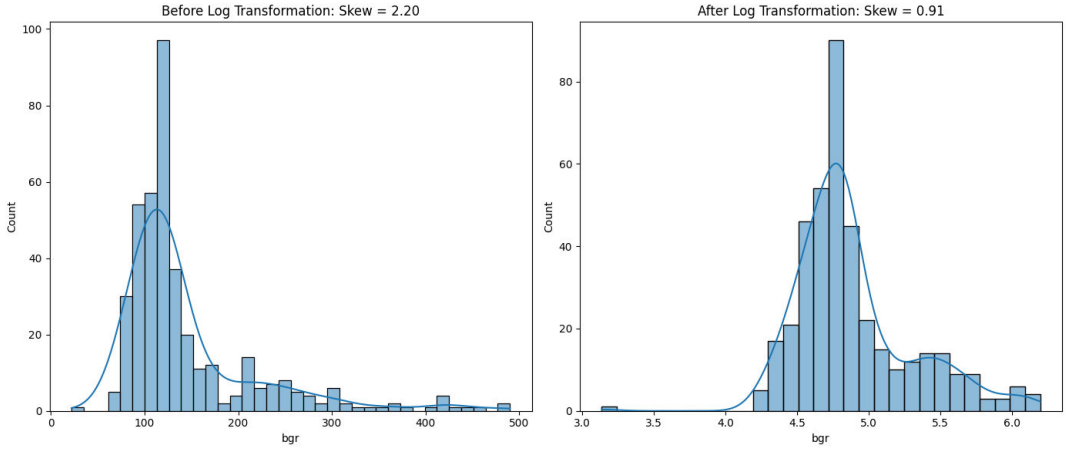


Figure 5: Effect of Log Transformation on Skewed Data

The figure 5 demonstrates how log transformation affects changes in the blood glucose level distribution. The left graph presents an original histogram displaying clear skewness because the value of skewness reaches 2.20. The distribution appears right-skewed or long-tailed. The log transformation made the distribution symmetric and decreased its skewness value to 0.91 in the right histogram. A log transformation both reduces the size of high values and expands the smaller values resulting in normalized distribution in the data. The data transformation leads to more stable variance since machine learning models rely on normal data distribution.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404008|

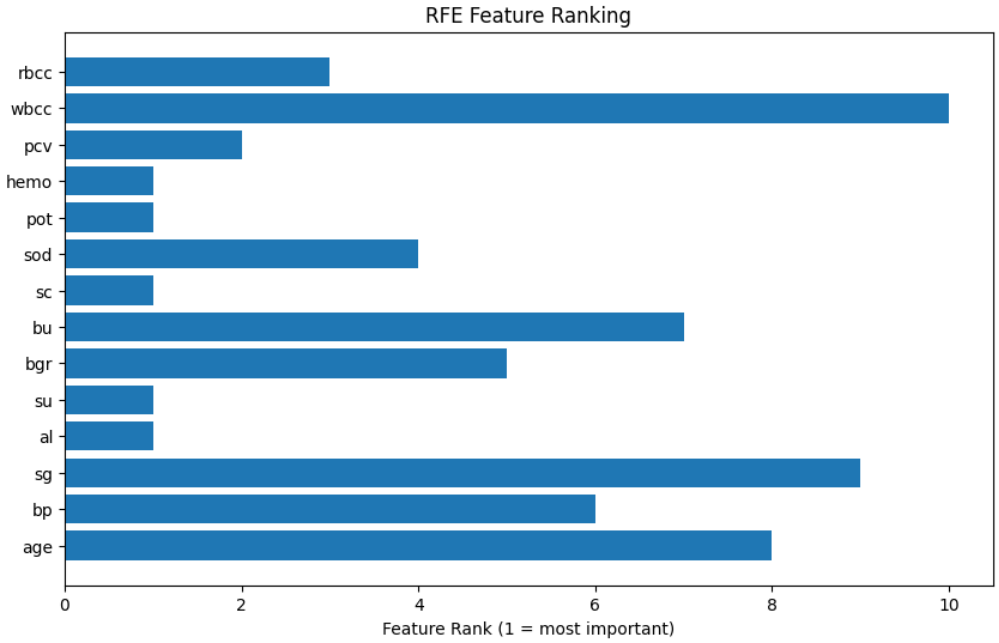


Figure 6: RFE Feature Ranking

The figure 6 demonstrates Recursive Feature Elimination (RFE)'s feature ranking process which ranks features according to their model contribution. The top two features according to the horizontal bar chart include rbcc (red blood cell count) along with wbcc (white blood cell count) that received the highest ranking scores proximate to 1 and following features include pcv (packed cell volume) and hemo (hemoglobin). The predictive model receives less contribution from the features age, bp (blood pressure), and sg (specific gravity) since these elements receive lower ranking positions. RFE method selects important features while discarding unneeded ones to enhance model performance along with simplification of computational requirements.

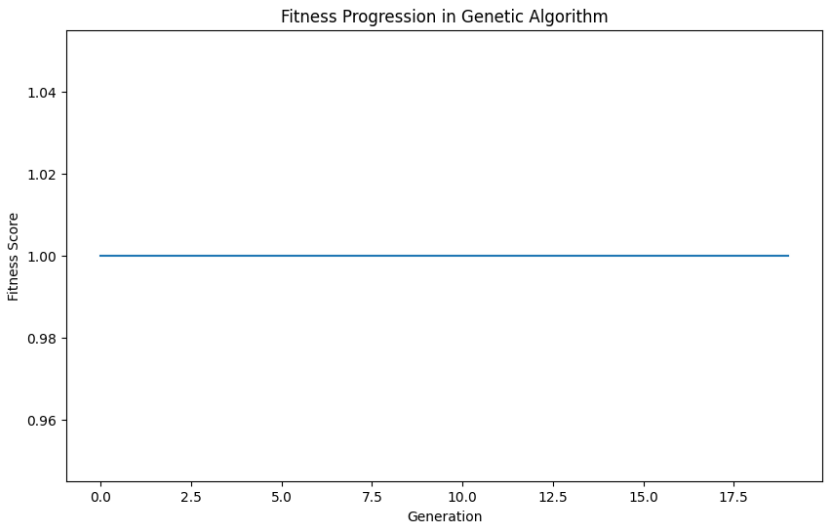


Figure 7: Fitness Progression in Genetic Algorithm

A Genetic Algorithm (GA) fitness evolution appears in figure 7 through its x-axis generation count and its y-axis fitness scoring metric. The 18 generation period shows that the fitness score stays at 1.00 which suggests that the algorithm

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404008|

discovered one optimal solution during its early execution phases. Whether the genetic algorithm reached its solution too fast or the population lacked diversity became a factor in generating insufficient changes across generations. A static fitness score trend is normal in dynamic genetic algorithms when the population approaches genetic equilibrium. The flat fitness level found during this experiment suggests no meaningful alterations in fitness that might result from problems such as diverse population deficiencies or insubstantial crossover and mutation elements.

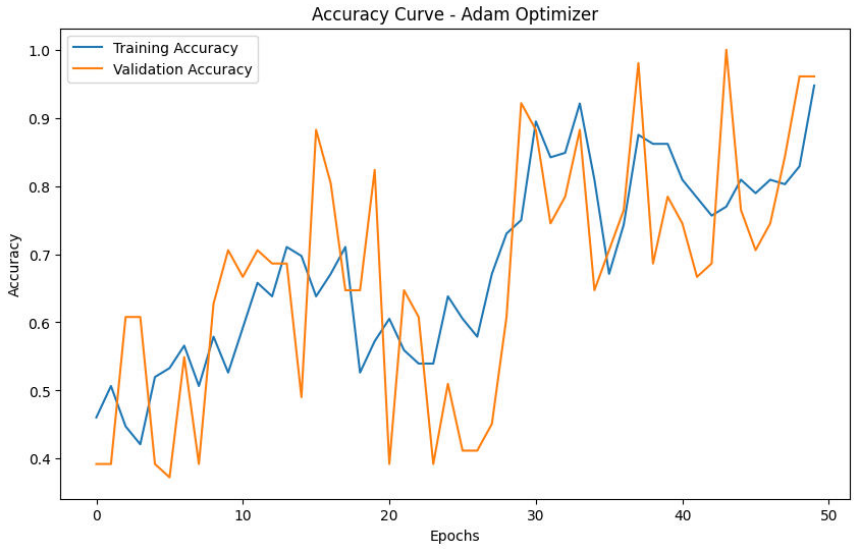


Figure 8: Accuracy Curve - Adam Optimizer

The accuracy curve during training with Adam optimizer is displayed in figure 8 over a 50 epoch period. The training accuracy (blue line) begins at 0.4 before continuing an upward progression until it attains 1.0 accuracy during the 50 epochs. Validation accuracy (orange line) maintains an initial level of 0.5 higher than training accuracy before it descends and reaches a maximum point of 0.9 near the end of the 50 epochs. The model demonstrates varied capability in generalization because its validation accuracy shows multiple peaks and valleys throughout the process. The upward slope in both curves demonstrates that Adam optimizer successfully improves model performance levels. Standard overfitting appears through the training accuracy maintaining higher performance than validation accuracy but this discrepancy reduces toward the end of training.

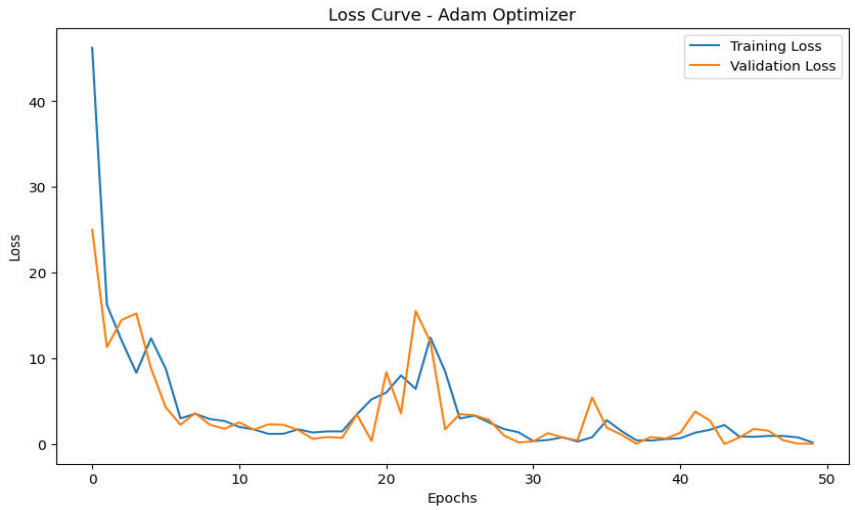


Figure 9: Loss Curve - Adam Optimizer

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404008|

The right figure demonstrates loss changes throughout training during 50 epochs when using Adam optimization. During the first epoch the training loss attains an initial value of 45 before a swift descent that brings it near 0 within the first handful of epochs. Loss values exhibit occasional spikes specifically in epochs 20 and 35 before experiencing periods of decreased values. During the training process the validation loss (orange line) starts at a higher value compared to training loss while it decreases consistently although it maintains slightly elevated levels. The curves experience numerous fluctuations especially during late epochs potentially due to overfitting when the model aligns too closely with training data thus creating higher values of validation loss. The model demonstrates progress during training as both losses decrease and the training loss approaches zero indicating effective learning although more training could enhance model performance.

V. CONCLUSION

The proposed machine learning pipeline enables effective preprocessing together with feature extraction as well as selection and optimization and evaluation for building robust CKD prediction models based on clinical data. The combination of Z-Score and Isolation Forest and IQR outlier detection approaches produces clean and reliable data followed by PCA and log transformation which enhances model capability to detect meaningful patterns in highly skewed datasets. RFE with Genetic Algorithms together with the Adam optimizer enable the selection of essential predictors to improve model performance. The data learning effectiveness of the model is supported by results from accuracy, loss and fitness progression metrics although additional adjustments to the model framework will potentially lead to better performance. Additional predictive power could be achieved through studying alternative optimization methods and adopting regularization strategies as well as including demographic and genetic data sources for future model development. The practical use of the model in medical practice requires investigation of unbalanced data class techniques along with studies about deploying the model into real-time clinic environments.

REFERENCES

1. Hassan, M., Awan, F. M., Naz, A., deAndrés-Galiana, E. J., Alvarez, O., Cernea, A., ... & Kloczkowski, A. (2022). Innovations in genomics and big data analytics for personalized medicine and health care: a review. International journal of molecular Sciences, 23(9), 4645.
2. Gupta, N. S., & Kumar, P. (2023). Perspective of artificial intelligence in healthcare data management: A journey towards precision medicine. Computers in biology and medicine, 162, 107051.Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., & Hua, L. (2012). Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems, 36, 2431-2448.
3. Melnykova, N., Shakhovska, N., Gregus, M., Melnykov, V., Zakharchuk, M., & Vovk, O. (2020). Data-driven analytics for personalized medical decision making. Mathematics, 8(8), 1211.
4. Clim, A., Zota, R. D., & Tinica, G. (2019). Big Data in home healthcare: A new frontier in personalized medicine. Medical emergency services and prediction of hypertension risks. International Journal of Healthcare Management, 12(3), 241-249.
5. Johnson, K. B., Wei, W. Q., Weeraratne, D., Frisse, M. E., Misulis, K., Rhee, K., ... & Snowdon, J. L. (2021). Precision medicine, AI, and the future of personalized health care. Clinical and translational science, 14(1), 86-93.
6. Alyass, A., Turcotte, M., & Meyre, D. (2015). From big data analysis to personalized medicine for all: challenges and opportunities. BMC medical genomics, 8, 1-12.
7. Dataset Repository: UCI Chronic Kidney Disease Dataset (Link: https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease),
8. Mondal, S., Das, S., Golder, S. S., Bose, R., Sutradhar, S., & Mondal, H. (2024, December). AI-Driven Big Data Analytics for Personalized Medicine in Healthcare: Integrating Federated Learning, Blockchain, and Quantum Computing. In 2024 International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA) (pp. 1-6). IEEE.
9. Reza Soroushmehr, S. M., & Najarian, K. (2016). Transforming big data into computational models for personalized medicine and health care. Dialogues in clinical neuroscience, 18(3), 339-343.
10. Alyass, A., Turcotte, M., & Meyre, D. (2015). From big data analysis to personalized medicine for all: challenges and opportunities. BMC medical genomics, 8, 1-12.
11. Ahmed, Z., Mohamed, K., Zeeshan, S., & Dong, X. (2020). Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database, 2020, baaa010.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404008|

12. Ghanem, M., Ghaith, A. K., & Bydon, M. (2024). Artificial intelligence and personalized medicine: transforming patient care. In The New Era of Precision Medicine (pp. 131-142). Academic Press.
13. Hu, Z. (2019). Achieving personalized medicine using machine learning: clinical data mining studies on coronary heart disease, substance use disorder, and Alzheimer's disease (Doctoral dissertation, University of Pittsburgh).
14. Goswami, S., & Sharma, S. (2024, March). Artificial intelligence, quantum computing and cloud computing enabled personalized medicine in next generation sequencing bioinformatics. In 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI) (Vol. 2, pp. 1-5). IEEE.
15. Valero-Ramon, Z., Fernandez-Llatas, C., Valdivieso, B., & Traver, V. (2020). Dynamic models supporting personalised chronic disease management through healthcare sensors with interactive process mining. Sensors, 20(18), 5330.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details