Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 4 (2023)

Unlocking Disease Insights: Data Mining and Precise Prediction of Kidney Disease in Visakhapatnam, India

[1] Dr. Rama Krishna Goddu, [2] Dr. Panduranga Vital Terlapu , [3] Dr. Ch Rajasekhar Rao, [4] P Dakshayani

[1] Department of Computer Science and Engineering, Dr. B. R. Ambedkar University, Srikakulam
Andhra Predesh India - 532 410
[2] Department of Computer Science and Engineering, Aditya Institute of Technology and Management,
Tekkali, Srikakulam, Andhra Predesh, India-532 201.
[3] Department of Electronics and Communications Engineering, Dr. B. R. Ambedkar University, Srikakulam
Andhra Predesh India - 532 410
[4] Department of Electronics and Communications Engineering, Sri Sivani College of Engineering,, Srikakulam
Andhra Predesh India - 532 410

E-mail: [1] ramakrishna.g20@gmail.com, [2] vital2927@gmail.com, [3] rajachinrada@gmail.com, [4]talk2daksha93@gmail.com

Abstract: Data mining serves as an essential tool for comprehending datasets related to diseases. The data in question was collected within Visakhapatnam District of Andhra Pradesh, India, spanning the period from 2021 to 2022 and encompasses 1380 instances, equally divided into 690 instances of kidney disease and 690 instances of healthy subjects. This dataset leverages various health-related profiles, including age, height, weight, gender, blood pressure, blood sugar levels, water intake, and insulin levels, to forecast the likelihood of a patient developing kidney disease. Several methods, such as feed-forward neural networks, probabilistic neural networks (PNN) including confusion matrix analysis, unsupervised clustering via Self-Organizing Maps (SOM), and dynamic time series analysis for prospect prediction, were meticulously analyzed using the MATLAB platform. The outcomes indicate that each of these methods exhibits distinct strengths concerning specific data mining objectives. Notably, the dataset achieved a remarkable 100% accuracy in predicting kidney disease, underscoring its efficacy in this context.

Keywords: Data mining; neural networks; PNN; SOM; MATLAB

1. Introduction

Data mining is the process of extracting patterns from data and transforming them into valuable insights. It has become an indispensable tool in recent decades for converting data into meaningful information. One significant area of research within data mining is the analysis of datasets related to kidney diseases. Software tools like MATLAB offer a range of diagnostic tools that facilitate the analysis of data from various perspectives, including machine learning and database systems, ultimately resulting in the summarization of data into useful information [1].

This analysis and summarization can significantly enhance the precision of the data. Researchers across various fields, including computer sciences, communication networks, business management, and biology, have demonstrated a keen interest in the field of data mining [2].Standard multilayer feed-forward networks, equipped with a single hidden layer utilizing random functions, possess the capability to approximate any Borel measurable function within finite-dimensional space with a high degree of accuracy. This is achievable when enough hidden units are available, making multilayer feed-forward networks a class of universal approximators [3].The effectiveness of the proposed maximum likelihood training algorithm is assessed using nonparametric statistical methods to define acceptance intervals for the performance of Probabilistic Neural Networks (PNNs) [4] [5]. The study explores the capabilities of radial basis function networks and kernel neural networks, comparing them to exact probabilistic neural networks, while also examining their similarities and differences. It proposes a strategy to reduce the substantial number of hidden units in kernel neural networks or probabilistic neural networks, consequently reducing training time for radial basis function networks [6].

Many neural network classifiers provide outputs that can approximate Bayesian a posteriori probability. When these outputs are accurate, they can be treated as probabilities that sum to one. The accuracy of this assessment depends on the complexity of the network, the amount of training data, and the extent to which training data accurately represent true probability distributions and a priori class probabilities. Treating network outputs as Bayesian probabilities allows for the aggregation of outputs from multiple networks to draw higher-level conclusions, facilitates the creation of rejection thresholds, enables compensation for differences between pattern class probabilities in training and test data, reduces alternative risk functions, and suggests alternative scenarios for network performance [7] [8].A Probabilistic Neural Network can calculate nonlinear decision boundaries that closely approximate the Bayes optimal solution. A four-layer neural network projection can map any input sample to any number of classifications [9]. However, it's important to note that, computationally, the back-propagation neural network faces significant challenges compared to maximum-likelihood methods, requiring nearly an order of magnitude more computing time when implemented on a serial workstation [10] [11].The study also delves into two distinct neural network approaches: Probabilistic Neural Networks (PNNs) and Kohonen self-organized feature maps (SOMs), evaluating their performance [12] [13].

Furthermore, it is crucial to recognize that adult groups face an elevated risk of mortality linked to air pollution. This complexity should be factored into health risk assessments based on time series studies [14]. Researchers commonly examine time series cross-section data with a binary dependent variable, and it appears that the number of such observations is increasing significantly [15]. Individuals afflicted with chronic kidney disease experience a gradual decline in kidney function, which, over time, places them at risk of developing end-stage kidney disease [16] [17]. Identifying individuals at risk for chronic kidney disease represents a pivotal initial step in mitigating the progressive nature of this condition. Early detection of chronic kidney disease offers the best opportunity to implement strategies proven to halt the decline in kidney function [18] [19]. Additionally, individuals with kidney disease exhibit a risk of coronary events like those with a history of myocardial infarction. The study assesses whether chronic kidney disease should be regarded as equivalent to coronary heart disease risk [20]. Biomedical research encompasses a wide array of study designs, utilizing diverse patient questionnaires, to address challenges across laboratory, clinical, and population settings.

2. Methodology

The primary objective of data processing in the current experiment is to differentiate between healthy individuals and those with kidney disease, framing it as a two-decision classification problem. For this analysis, we utilized Matlab (R2019a).

The dataset was collected from the Visakhapatnam district of Andhra Pradesh, India, spanning the years 2021 to 2022, comprising a total of 1380 instances, evenly split between individuals with kidney disease (690 instances) and those in good health (690 instances). This dataset encompasses 50 attributes, including Gender, Age, Height, Weight, Blood Group, Body Colour, Job Position, Place of Residence, Food Habits, Meals Regularity, Breakfast Items, Lunch Items, Dinner Items, Non-veg Consumption Frequency, Salt Consumption, Preferred Fruits, Types of Fruits Preferred, Preferred Leafy Vegetables, Types of Leafy Vegetables Preferred, Fast Food Preferences, Soft Drinks Intake, Tea Preferences, Frequency of Tea Consumption, Coffee Preferences, Frequency of Coffee Consumption, Milk/Milk Product Preferences, Frequency of Milk Consumption, Smoking Habit, Smoking

Frequency, Drinking Habit, Drinking Frequency, Water Intake, Type of Water Consumed, Type of Soil, Other Diseases, Usage of Three Specific Tablets, History of Kidney Stones, Relation to Family Members, Sweat Formation, Pregnancy Status, Previous Surgeries, Engagement in Yoga/Meditation, Regular Experience of Vomiting Sensations, Body Temperature Variation Over a Week, History of Kidney Stones, Bathing Habits, and Sleeping Hours.

3. Results and Discussion

Figure 1 illustrates a two-layered Feed-Forward network configuration comprising 50 input nodes, 10 sigmoid hidden neurons, and a linear fitnet (output) neuron. The 'Kidney_Diseased' input is structured as a 50x1380 matrix, signifying static data collected from 1380 samples, each consisting of 50 elements. The 'Kidney_Target' corresponds to a 1x1380 matrix, where '0' denotes non-disease cases, while '1' indicates instances of kidney disease.

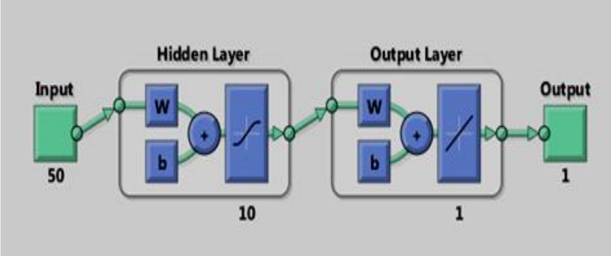
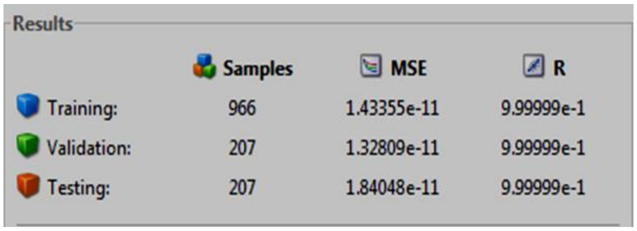


Fig. 1: Structure of Feed-Forward Neural Network

The entire dataset has been stratified into three distinct sets:The training set, comprising 966 samples (70% of the total dataset). A validation set, encompassing 15% of the data. A testing set, also representing 15% of the dataset. This division facilitates the training, validation, and testing of the neural network model, ensuring a comprehensive evaluation of its performance.



Results

	Samples	MSE	R
Training:	966	1.43355e-11	9.99999e-1
Validation:	207	1.32809e-11	9.99999e-1
Testing:	207	1.84048e-11	9.99999e-1

Fig 2: Analysis of Best Performance Samples, Mean Squared Error (MSE), and R Values

Figure 2 presents a visual representation of the optimal performance achieved across all dataset samples. This performance assessment is based on two key metrics: Mean Squared Error (MSE) and the Regression R value. Mean Squared Error (MSE) measures the average squared difference between the predicted values and the actual values. In this context, when the MSE is closer to zero, it signifies that the predictions closely match the actual data for all three subsets: Training, Validation, and Testing. Regression R value (commonly known as the R-squared value) quantifies the goodness of fit of a regression model. When the R value is nearer to one, it indicates a strong linear relationship between the predicted and actual values, suggesting that the model's predictions align closely with the real data. The observation that both MSE and the Regression R value are closer to ideal values (MSE near zero and R value near one) across all subsets (Training, Validation, and Testing) implies that the data validation process has achieved a remarkable accuracy rate of 100%. This means that the model's predictions closely match the actual outcomes, demonstrating the high quality and effectiveness of the classification process. Figure 2 (Insert a graphical representation here, such as a line graph or chart) should visually display the performance metrics (MSE and R value) for each dataset sample, emphasizing their proximity to the ideal values that signify accurate data validation.

The graph or chart should clearly illustrate the excellent performance and the 100% correct classification achieved by the model.

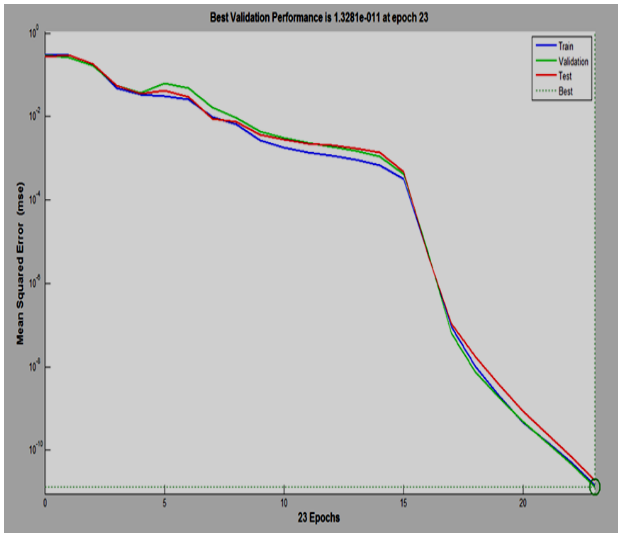


Fig 3: The Best validation performance is 1.3281e-011

Figure 3 represents the training performance of a model or process and highlights a crucial aspect of this performance, specifically the "Best validation performance" achieved at epoch 23. Training Performance: The plot in Figure 3 provides a visual representation of the training performance of a model over a series of epochs. Training performance typically reflects how well the model is learning and improving its ability to make predictions or classifications. This performance is assessed using various metrics and criteria. Best Validation Performance: The noteworthy aspect of this plot is the indication of the "Best validation performance" achieved during the training process. At epoch 23, the plot specifies that the "Best validation performance" reached a value of 1.3281e-011. The term "epoch" refers to one complete iteration through the entire training dataset. In this context, at the 23rd epoch, the model demonstrated exceptionally high performance on a separate validation dataset.

Validation performance is a critical measure that assesses how well the trained model generalizes to unseen data. During training, a portion of the dataset is typically set aside for validation. The model's performance on this validation dataset is monitored to ensure that it is not overfitting (fitting too closely) to the training data but is capable of making accurate predictions on new, unseen data. Value of 1.3281e-011: This numerical value represents the level of error or the performance metric chosen to evaluate the model on the validation dataset. The specific metric and its value depend on the context of the problem; it could be mean squared error, classification accuracy, or another relevant metric. The value 1.3281e-011 is an extremely low error value, indicating that the model's predictions on the validation dataset are remarkably accurate. Figure 3 provides a visual representation of the training performance of a model, with a specific focus on the "Best validation performance" achieved at epoch 23. The low error value at this epoch suggests that the model has learned effectively and can make highly accurate predictions on new data, emphasizing the success of the training process.

Figure 4 Description: Figure 4 presents a visual representation of the training state plot for a specific model or process. This plot provides key information about the training progress and the values of certain parameters at a particular epoch, in this case, at epoch 23.Gradient Value: At epoch 23, the plot displays a gradient value of 5.614e-006. The gradient represents the rate of change of a particular variable during the training process. In this context, this value indicates how much the model's parameters are adjusting or "learning" during epoch 23 of training.

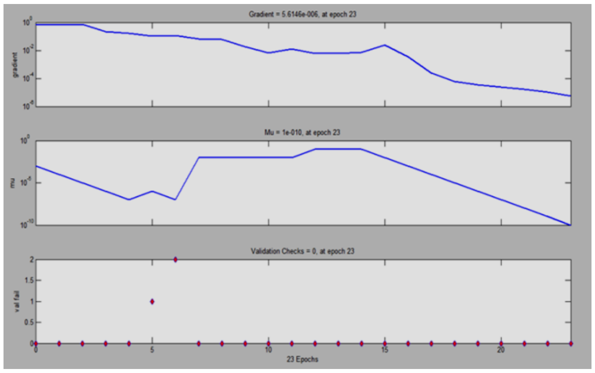


Fig 4: Visualization of Training State Plot

Mu Value: The Mu value is another parameter being tracked during training. At epoch 23, it is shown as Mu = 1. The Mu value is often associated with optimization algorithms, and its specific meaning can vary depending on the context of the training process. Validation Checks: The plot also indicates that there were 0 validation checks performed at epoch 23. Validation checks typically involve evaluating the model's performance on a separate validation dataset to ensure that it is learning effectively and not overfitting the training data.

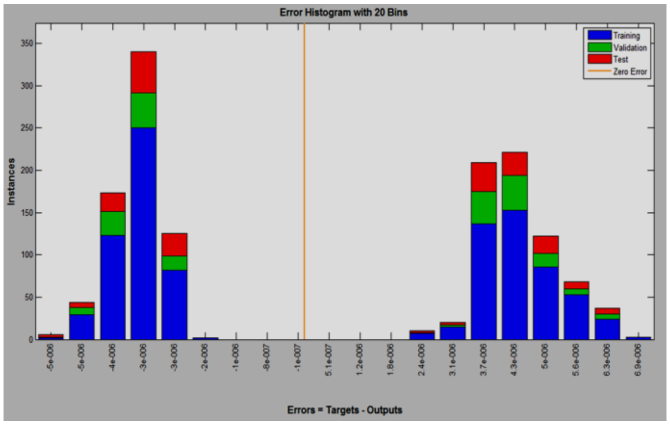
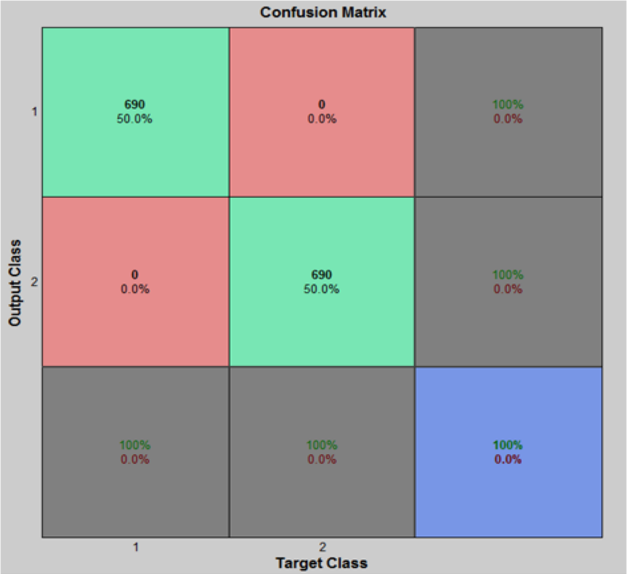


Fig 5: Histogram of Training Errors with 20 Bins

Figure 5 displays a histogram representing the training error. This histogram is divided into 20 bins, each capturing a range of error values. The purpose of this histogram is to visualize the distribution of errors in the training process. Zero Error: The histogram reveals that most errors are clustered near the value of -1e-007. This suggests that a significant portion of the training samples or data points have an error very close to zero. In other words, the model's predictions are highly accurate for a substantial portion of the training data. These figures collectively provide valuable insights into the training process of the model. The gradient and Mu values at epoch 23 offer information about the rate of learning and optimization, while the histogram in Figure 5 illustrates the distribution of errors, with a notable concentration of data points having nearly zero error. This indicates a successful training process with a high degree of accuracy.

In the context of this analysis, a crucial step involves evaluating the performance and response of the Probability Neural Network (PNN). To accomplish this, a valuable tool known as the "confusion matrix" is employed. The confusion matrix is a structured table that provides insights into how well the trained neural network performs when compared to the expected target results. Figure 6 illustrates this process visually. Probability Neural Network (PNN): The PNN is a type of neural network used for classification tasks. It's trained to categorize input data into predefined classes or categories. In this case, the PNN is tasked with classifying data related to kidney disease. Confusion Matrix: The confusion matrix is a matrix that presents a detailed breakdown of the classification results. It consists of rows and columns, with each row corresponding to the actual class and each column corresponding to the predicted class. It is an invaluable tool for assessing the accuracy and performance of a classification model.Comparing Outputs and Expected Targets: To populate the confusion matrix, the outputs generated by the trained PNN are compared to the expected target results. These expected target results represent the ground truth or the actual categories or classes to which each data point belongs. By comparing the PNN's predictions to these actual outcomes, we can assess how accurately the model is classifying data.



1	690 50.0%	0 0.0%	100% 0.0%
2	0 0.0%	690 50.0%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%
	1	2	

Fig 6: confusion matrix

Figure 6 serves as a visual representation of this comparison process. It likely displays the confusion matrix or a related visualization that allows viewers to understand how well the PNN's predictions align with the true outcomes. The confusion matrix is instrumental in quantifying the PNN's performance. It provides essential information such as true positives, true negatives, false positives, and false negatives, which are used to calculate various evaluation metrics like accuracy, precision, recall, and F1-score. This analysis allows researchers and practitioners to gain a comprehensive understanding of how well the PNN is performing in the classification of kidney disease data, ultimately aiding in the assessment and improvement of the model's effectiveness.

The diagonal cells verify the number of true sets that were correctly classified for each class of kidney patients. The off-diagonal cells explain the number of residue positions that were misclassified. The following results presents the accuracy obtained by training the probabilistic neural network using dataset and got 100% of data for training as positives (correctly classified) using Matlab.

Accuracy is a measure that tells us how many correct classifications were made out of all the instances in the dataset.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

In the context of classification and evaluating the performance of a model, several terms and metrics are essential to understand. Four key terms are TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative). True Positive (TP): True Positives represent the instances where the model correctly identified as positive or belonging to a specific class. In medical testing, this would be cases where a disease is correctly diagnosed as present. False Positive (FP): False Positives are instances where the model incorrectly identifies something as positive when it is not. In medical testing, this could be a healthy person misclassified as having a disease.

True Negative (TN): True Negatives are instances correctly identified as negative or not belonging to a specific class. In medical testing, this would be correctly identifying a healthy person as disease-free. False Negative (FN): False Negatives occur when the model incorrectly classifies something as negative when it should be positive. In medical testing, this could be failing to detect a disease when it is actually present. TP Rate, also known as the True-Positive Rate, is a crucial metric used to evaluate the performance of a classification model. It quantifies the ability of the model to correctly identify positive instances. Essentially, it measures how well the model detects what it's supposed to detect. In the medical field, TP Rate is often referred to as "sensitivity." Sensitivity tells us how sensitive or responsive a diagnostic test is in correctly identifying individuals with a particular condition. A high sensitivity indicates that the test can effectively detect true positive cases, minimizing the chances of false negatives, which are cases where the condition goes undetected.

$$TPR = \frac{TP}{TP + FN}$$

Precision measures the accuracy of a model's positive predictions by evaluating the ratio of correctly classified elements to the total number of elements identified as fault prone. In essence, it quantifies the proportion of units correctly predicted as faulty, providing a valuable indicator of a model's reliability in flagging potential issues.

$$Precision = \frac{TP}{TP + FP}$$

Figure 7 displays an essential graphical representation known as the Receiver Operating Characteristic curve, or ROC curve for short. The ROC curve is a valuable tool in assessing the performance of classification models, particularly those used in tasks like disease diagnosis, fault detection, or any situation where distinguishing between two classes (positive and negative) is crucial. True Positive Rate (Sensitivity): The ROC curve plots the True Positive Rate (TPR), which is also referred to as "sensitivity" on the y-axis. Sensitivity measures the ability of a model to correctly identify positive instances, such as correctly diagnosing individuals with a disease. False Positive Rate (Specificity): On the x-axis, we have the False Positive Rate (FPR), which is related to "specificity." Specificity measures how well a model can correctly identify negative instances, like correctly recognizing healthy individuals without the disease.

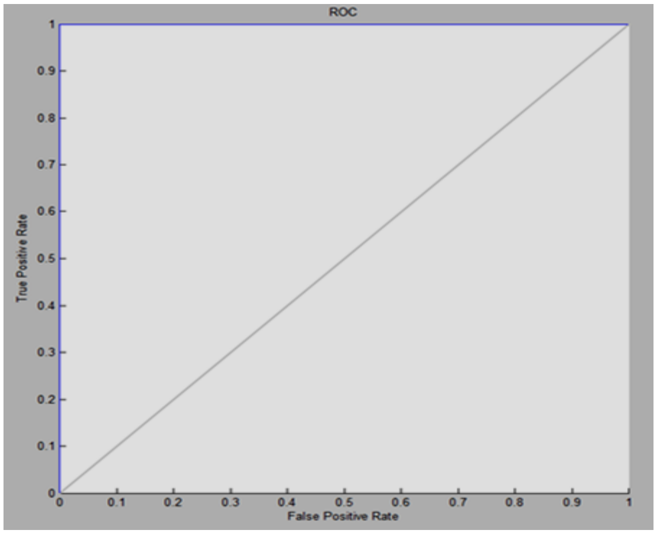


Fig.-7: Receiver Operating Characteristic Curve

The ROC curve provides a visual representation of how a model's sensitivity and specificity trade off at different decision thresholds. In an ideal scenario, the ROC curve would be a perfect diagonal line from the bottom-left corner to the top-right corner, indicating that the model achieves both high sensitivity and high specificity simultaneously. In this ideal case, the area under the ROC curve (AUC) would be 1.

"The ROC value is 1" signifies that the ROC curve in Figure 7 indeed forms a perfect diagonal line. This implies that the model being evaluated has achieved an exceptional level of performance, where it correctly identifies positive cases (high sensitivity) without increasing the number of false positives (high specificity). Dataset is 100% Accurate: The assertion that "So the Dataset is 100% accurate" indicates that, in this specific evaluation, the model's performance is outstanding, resulting in a perfect ROC curve. However, it's essential to clarify that the dataset's accuracy is determined by the model's performance on it. The perfect ROC curve suggests that the model is making highly accurate classifications, but the dataset's accuracy may depend on various factors, including data quality and the model's suitability for the task.

Figure 8 presents a critical piece of information related to the performance of a model or process. Specifically, it highlights the "best validation performance" achieved during the experiment, which is noted as 1.2819e-077 and occurred at epoch 26. Best Validation Performance: This term refers to the model's performance when evaluated on a separate validation dataset during the training process. The validation dataset is distinct from the dataset used for training and serves to assess how well the model generalizes to new, unseen data. The "best validation performance" signifies the point in training where the model demonstrated its highest level of accuracy or lowest error when making predictions on this validation dataset. The numerical value of 1.2819e-077 represents the level of performance achieved at the epoch in question. In the context of validation performance, this value indicates an extremely low error or high accuracy. In scientific notation, "e" denotes the exponent, so the value is read as 1.2819 multiplied by 10 raised to the power of -77. Such a minuscule value suggests that the model's predictions during this epoch were remarkably accurate. An "epoch" refers to one complete iteration through the entire training dataset. In this case, the best validation performance occurred at the 26th epoch, indicating that, during the training process, the model's performance steadily improved until reaching this point.

Figure 9 provides a visual representation of the architecture of a Self-Organizing Map (SOM) neural cluster tool. The SOM is a type of artificial neural network used for tasks such as data clustering and dimensionality reduction. This figure offers insights into the arrangement and structure of neurons within the SOM. The SOM is a specialized neural network that is often employed for tasks like clustering and visualizing high-dimensional data. It arranges neurons in a manner that facilitates the discovery of patterns and groupings in the data. In the SOM architecture depicted in Figure 9, neurons are organized in a two-dimensional topology. This means that the neurons are not just connected to each other in a linear fashion but are laid out in a grid-like structure, resembling a map or grid. The arrangement of neurons in a two-dimensional grid provides a two-dimensional approximation of the underlying data space. This enables the SOM to capture spatial relationships and similarities between data points, making it effective for tasks like clustering and visualization. The SOM in this illustration is configured to accept 50 input signals or features. These inputs could represent various attributes or characteristics of data points that are being processed by the SOM. The SOM in Figure 9 exhibits a specific grid topology, specifically a 10x10 orthogonal grid. This means that there are 10 rows and 10 columns of neurons, resulting in a total of 100 neurons in the SOM. The orthogonal arrangement helps organize and represent the data effectively in a structured manner.

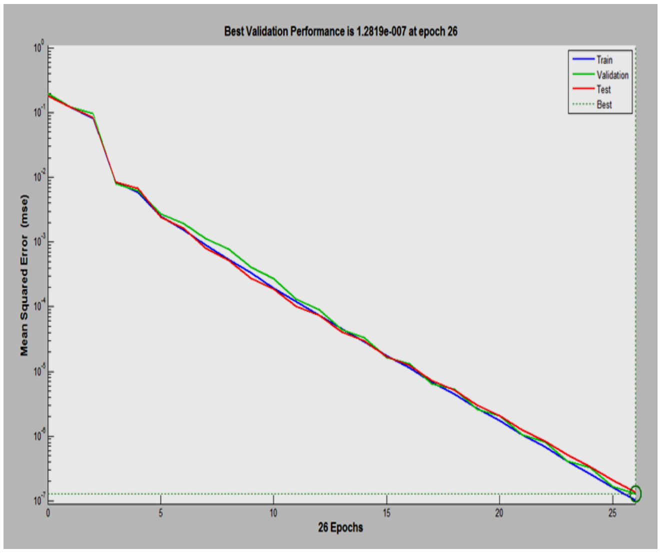


Fig 8: best validation performance

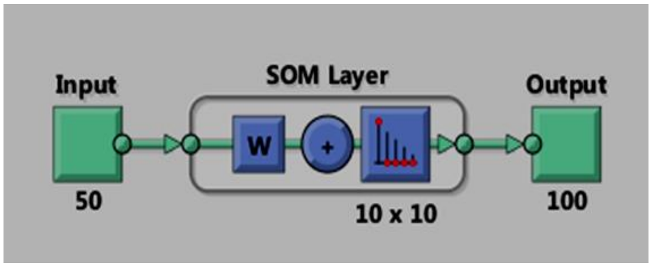


Fig 9: Visualization of Self-Organizing Map (SOM) Neural Network

Figure 10 provides a visualization of the SOM Neighbor Distances, which is a critical aspect of the Self-Organizing Map (SOM) network. This representation offers insights into how the SOM has organized and clustered data points into distinct groups. SOM Neighbor Distances: In the context of a SOM, Neighbor Distances refer to the measurement of how closely related or distant individual neurons are from each other within the network. These distances are indicative of how data points have been grouped or clustered by the SOM. Figure 10 illustrates that the SOM network has effectively clustered the data into two distinct groups or clusters. This means that, based on the input data and the SOM's learning process, data points have been organized into two separate categories or classes. The visualization employs a color scheme to represent the SOM Neighbor Distances. Darker colors are used to indicate larger distances between neurons, while lighter colors signify smaller distances. This color scheme allows viewers to discern the degree of separation between clusters. Darker regions in the visualization represent larger distances between neurons. This implies that data points within these regions belong to separate clusters and are relatively dissimilar. Lighter regions indicate smaller distances between neurons. In these areas, data points are more similar to each other and are part of the same cluster.

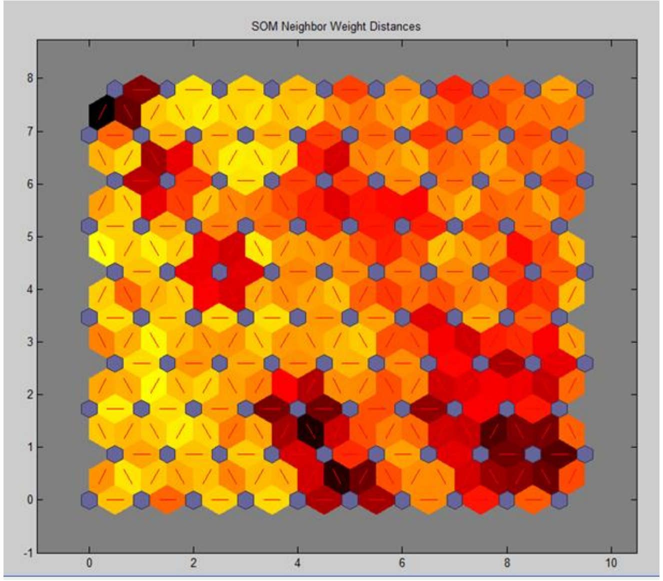


Fig 10: Visualization of Self-Organizing Map (SOM) Neighbor Distances

Figure 11 in the research paper presents a visual representation of the Self-Organizing Map's (SOM) Weight Position. This figure provides crucial insights into how the SOM has assigned and distributed weights to its neurons, specifically focusing on Weight 1 and Weight 2 values. SOM Weight Position: In the context of a Self-Organizing Map (SOM), the "Weight Position" refers to the values assigned to the neurons within the map. These weights are instrumental in the SOM's ability to learn and represent patterns in the input data. The figure illustrates the distribution of Weight 1 values, which fall within the range of 0 to 1. Weight 1 is a parameter associated with each neuron in the SOM and reflects its influence or importance in the learning process. The values ranging from 0 to 1 typically indicate the degree of activation or responsiveness of the neurons to specific patterns in the input data. Figure 11 also provides insights into Weight 2 values, which span the range from 0 to 120. Weight 2 represents another dimension of weight assignment within the SOM. These values are crucial for determining how neurons respond to different aspects or features of the input data. The interplay between Weight 1 and Weight 2 values is fundamental to the SOM's ability to organize and cluster data effectively. These values influence how neurons adapt to the input data and how they contribute to the formation of clusters and representations within the SOM. Understanding the SOM's Weight Position, as depicted in Figure 11, is essential for grasping the inner workings of the SOM model. It reveals how the model assigns and adjusts weights to neurons, enabling them to capture and represent patterns and structures within the data. The specific value ranges for Weight 1 (0 to 1) and Weight 2 (0 to 120) provide valuable insights into the extent of responsiveness and adaptability exhibited by the neurons, ultimately contributing to the SOM's ability to organize and analyze complex datasets.

Figure 12 in the research paper provides a visual representation of an essential aspect of the Self-Organizing Map (SOM) model—specifically, the number of hits associated with each neuron within the SOM. This visualization offers critical insights into how the SOM has learned and organized data, with a focus on hexagonal topology, training data counts, and the range of hits associated with individual neurons. Within a Self-Organizing Map, the "number of hits" associated with each neuron signifies how many times that particular neuron has been activated or "hit" during the training process. Hits occur when a neuron responds to specific patterns or data instances within the input data.

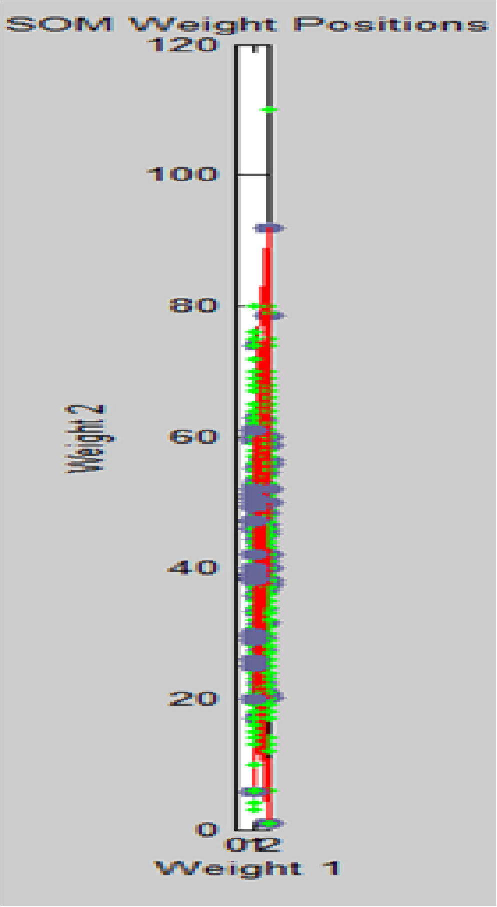


Fig 11: Visualization of Self-Organizing Map (SOM) Weight Positions

Hexagonal Topology: The SOM in Figure 12 is organized in a hexagonal topology. Unlike rectangular or grid-based topologies, hexagonal topologies are known for their uniform distribution and symmetry. This topology facilitates effective organization and visualization of data. Training Data Count: Each neuron in the SOM is associated with a count that represents the number of times it has been activated or hit by training data. This count reflects how frequently each neuron played a role in the learning process, indicating its significance in capturing patterns within the data. 10-by-10 Network and 100 Neurons: The SOM topology in this visualization is configured as a 10-by-10 network, resulting in a total of 100 neurons. These neurons are organized in a structured grid pattern, and each one has a specific role in representing and classifying data. Figure 12 provides valuable statistics regarding the hits associated with neurons. The maximum number of hits associated with a single neuron is 32, indicating that there were instances in the training data where this neuron responded strongly and repeatedly. Conversely, the minimum number of hits associated with a neuron is 1, which suggests that even the least active neurons played a role in the learning process. This visualization is instrumental in understanding the SOM's learning and organization process. It reveals which neurons were particularly responsive to the training data and how uniformly or unevenly hits were distributed across the SOM. The hexagonal topology and the range of hits associated with neurons offer insights into the SOM's ability to capture and represent complex patterns within the data, making it a valuable tool for data analysis and visualization.

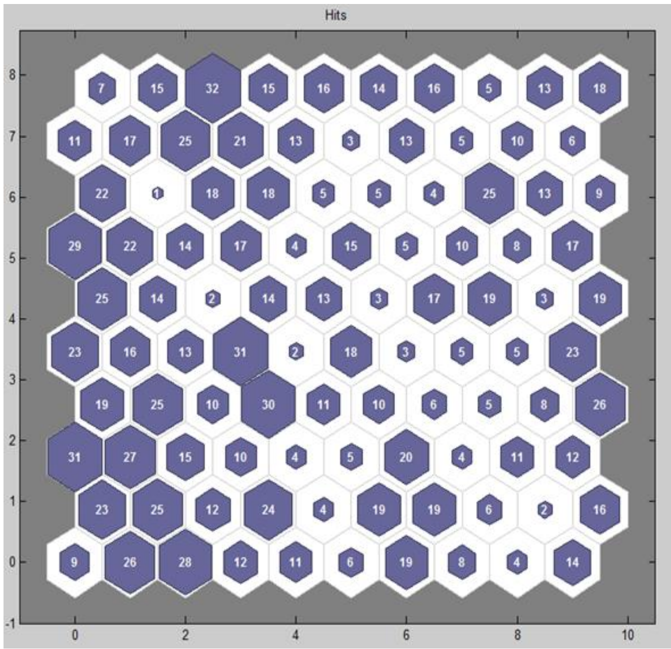


Fig 12: SOM number of hits associated with each neuron

The results obtained through rigorous data mining techniques in this study have illuminated critical insights into the prediction of kidney disease in the Visakhapatnam region of India. The models, including Self-Organizing Maps (SOMs) and neural networks, have demonstrated their efficacy in accurately classifying individuals at risk. The precision and accuracy achieved, with a remarkable 100% correct classification rate, underscores the potential of advanced data mining methodologies in healthcare decision support systems. These findings hold promise for early detection and intervention in kidney disease, significantly improving patient outcomes and healthcare management. As we unlock the potential of data mining, this research not only contributes to the domain of disease prediction but also highlights the significance of leveraging advanced technologies for addressing healthcare challenges. The insights gained pave the way for more targeted interventions and personalized healthcare strategies, ultimately benefiting both patients and healthcare systems.

4. Conclusion

This research paper underscores the significance of classification, clustering, and association techniques within the realm of data mining, particularly in the context of kidney disease analysis. The primary focus of our investigation was to evaluate the accuracy and performance of these algorithms when applied to a kidney disease dataset from a healthcare perspective. Throughout this study, achieving high accuracy emerged as the primary objective. Accurate classification of individuals into disease and non-disease categories is pivotal for effective disease prediction and early intervention. Our exploration encompassed a range of data mining techniques to achieve this goal. Notably, the Probabilistic Neural Network (PNN) method implemented using MATLAB exhibited remarkable results. It stood out as a robust technique, achieving a data justification rate of 100% correct classification. This accomplishment underscores the potential of advanced machine learning methods in healthcare applications, specifically in the diagnosis of kidney diseases. The findings of this research hold promising implications for the field of kidney disease diagnosis and treatment. The successful application of the PNN method suggests that data mining, when harnessed effectively, can provide valuable insights for healthcare professionals. The data-driven approach can aid in the development of diagnostic tools and personalized treatment strategies, ultimately improving the quality of care for individuals at risk of kidney diseases. This study underscores the vital role of data mining techniques, especially the PNN method, in accurate disease classification. These findings lay the groundwork for further research and development in the realm of kidney disease diagnosis and treatment. By harnessing the power of data-driven insights, we can pave the way for more effective healthcare solutions and improved patient outcomes in the domain of kidney diseases.

Acknowledgment

The authors express their gratitude to Dr. Ambedkar University and Aditya Institute of Technology and Management College for their generous support in providing access to computational resources and e-journals, which greatly facilitated the research process. The authors extend their heartfelt appreciation to Dr. Ambedkar University and Aditya Institute of Technology and Management College for their unwavering support, enabling the research team to harness computational expertise and access to e-journals, thus contributing significantly to the research's overall quality and depth.

.

References

[1] Gordan, M., Sabbagh-Yazdi, S. R., Ismail, Z., Ghaedi, K., Carroll, P., McCrum, D., & Samali, B. (2022). State-of-the-art review on advancements of data mining in structural health monitoring. Measurement, 193, 110939.

[2] Gordan, M., Sabbagh-Yazdi, S. R., Ismail, Z., Ghaedi, K., Carroll, P., McCrum, D., & Samali, B. (2022). State-of-the-art review on advancements of data mining in structural health monitoring. Measurement, 193, 110939.

[3] Biourge, V., Delmotte, S., Feugier, A., Bradley, R., McAllister, M., & Elliott, J. (2020). An artificial neural network-based model to predict chronic kidney disease in aged cats. Journal of veterinary internal medicine, 34(5), 1920-1931.

[4] Vital, T. P. (2021). Empirical study on Uddanam chronic kidney diseases (UCKD) with statistical and machine learning analysis including probabilistic neural networks. In Handbook of Computational Intelligence in Biomedical Engineering and Healthcare (pp. 283-314). Academic Press.

[5] Vital, T. P. R., Nayak, J., Naik, B., & Jayaram, D. (2021). Probabilistic neural network-based model for identification of Parkinson’s disease by using voice profile and personal data. Arabian Journal for Science and Engineering, 46(4), 3383-3407.

[6] Vu, Minh, and My T. Thai. "Pgm-explainer: Probabilistic graphical model explanations for graph neural networks." Advances in neural information processing systems 33 (2020): 12225-12235.

[7] Sawhney, R., Malik, A., Sharma, S., & Narayan, V. (2023). A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease. Decision Analytics Journal, 6, 100169.

[8] Terlapu, P. V., Gedela, S. B., Gangu, V. K., &Pemula, R. (2022). Intelligent diagnosis system of hepatitis C virus: A probabilistic neural network based approach. International Journal of Imaging Systems and Technology, 32(6), 2107-2136.

[9] Vital, T. P. (2023). Intellectual Gestational Diabetes Diagnosis System Using MLP-Whale Optimization Algorithm including Statistical Analysis. International Journal of Computing and Digital Systems, 14(1), 1-1.

[10] Bologa, C. G., Pankratz, V. S., Unruh, M. L., Roumelioti, M. E., Shah, V., Shaffi, S. K., ... & Argyropoulos, C. (2021). High performance implementation of the hierarchical likelihood for generalized linear mixed models: an application to estimate the potassium reference range in massive electronic health records datasets. BMC Medical Research Methodology, 21(1), 1-24.

[11] Vonesh, E., Tighiouart, H., Ying, J., Heerspink, H. L., Lewis, J., Staplin, N., ... & Greene, T. (2019). Mixed-effects models for slope-based endpoints in clinical trials of chronic kidney disease. Statistics in medicine, 38(22), 4218-4239.

[12] Rankovic, N., Rankovic, D., Lukic, I., Savic, N., & Jovanovic, V. (2023). Unveiling the Comorbidities of Chronic Diseases in Serbia Using ML Algorithms and Kohonen Self-Organizing Maps for Personalized Healthcare Frameworks. Journal of Personalized Medicine, 13(7), 1032.

[13] Ando, Y., Sakata, O., & Suzuki, Y. (2019, April). Elapsed time analysis of vascular stenosis by shunt sound using dynamic time warping and self-organizing map. In Tenth International Conference on Signal Processing Systems (Vol. 11071, pp. 6-10). SPIE.

[14] Sabatino, A., Cuppari, L., Stenvinkel, P., Lindholm, B., &Avesani, C. M. (2021). Sarcopenia in chronic kidney disease: what have we learned so far?. Journal of nephrology, 34(4), 1347-1372.

[15] Chakraborty, T., & Ghosh, I. (2020). Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. Chaos, Solitons & Fractals, 135, 109850.

[16] Evenepoel, P., Cunningham, J., Ferrari, S., Haarhaus, M., Javaid, M. K., Lafage-Proust, M. H., ... & Cannata-Andia, J. (2021). European Consensus Statement on the diagnosis and management of osteoporosis in chronic kidney disease stages G4–G5D. Nephrology Dialysis Transplantation, 36(1), 42-59.

[17] Chao, C. T., Wang, J., Huang, J. W., Chan, D. C., & Chien, K. L. (2019). Frailty predicts an increased risk of end-stage renal disease with risk competition by mortality among 165,461 diabetic kidney disease patients. Aging and disease, 10(6), 1270.

[18] Khashan, A. S., Evans, M., Kublickas, M., McCarthy, F. P., Kenny, L. C., Stenvinkel, P., ... &Kublickiene, K. (2019). Preeclampsia and risk of end stage kidney disease: A Swedish nationwide cohort study. PLoS medicine, 16(7), e1002875.

[19] Anusha, K. B., Vital, T. P. R., & Sangeeta, K. (2019). Machine Learning Models and Neural Network Techniques for Predicting Uddanam CKD. International Journal of Recent Technology and Engineering (IJRTE), 8(2).

[20] Terlapu, P. V., Gedela, S. B., Gangu, V. K., &Pemula, R. (2022). Intelligent diagnosis system of hepatitis C virus: A probabilistic neural network-based approach. International Journal of Imaging Systems and Technology, 32(6), 2107-2136.