Optimized Biomarker-Based Machine Learning Model for Accurate Early Detection of Chronic Kidney Disease

Dr. Santhoshkumar Sundar[1], Dr. S.K. Manju Bargavi[2]

[1] Post-Doctoral Research Scholar, Lincoln University College, Malaysia & Assistant Professor, Department of Computer Science, Alagappa University, India

[2] Professor, School of Computer Science and IT, Jain (Deemed-to-be University), Bangalore

pdf.santhoshkumar@lincoln.edu.my, santhoshkumars@alagappauniversity.ac.in, cloudbargavi@gmail.com

---

Abstract

Chronic Kidney Disease (CKD) is a progressive and life-threatening disorder affecting millions worldwide. Early detection plays a critical role in preventing complications such as renal failure, cardiovascular disease, and increased mortality. Traditional diagnostic approaches rely primarily on serum creatinine and estimated Glomerular Filtration Rate (eGFR), which may not effectively detect early-stage CKD. This study proposes a robust machine learning framework integrating denoising and sparse feature selection approaches with biomarker analysis to enhance CKD detection accuracy. A Denoising Sparse Auto-Encoder (DSAE) is utilized to eliminate noise and extract meaningful representations from high-dimensional clinical data. Term Frequency–Inverse Document Frequency (TF-IDF) is adapted for identifying significant biomarkers. Sequential Selection Ratio (SSR) and Feature Normalization Component (FNC) techniques optimize feature selection and scaling. Finally, an Adaptive Backpropagation Neural Network (ABPNN) performs classification. Experimental results demonstrate superior accuracy, improved ROC performance, and enhanced diagnostic reliability. The proposed framework provides an efficient and scalable solution for early CKD detection and supports clinical decision-making.

Keywords: Chronic Kidney Disease, Machine Learning, Biomarkers, Denoising Sparse Auto-Encoder, Sequential Selection Ratio, Feature Normalization Component, Adaptive Backpropagation Neural Network, ROC Curve.

Introduction

Chronic Kidney Disease (CKD) represents a major global public health concern characterized by gradual loss of kidney function over time. If undetected, CKD may progress to end-stage renal disease requiring dialysis or transplantation. Early detection is therefore essential to prevent irreversible damage and reduce healthcare costs.

Conventional diagnostic procedures primarily depend on biomarkers such as serum creatinine, blood urea, haemoglobin, and Glomerular Filtration Rate (GFR). However, these parameters alone may not sufficiently capture subtle early-stage abnormalities. Furthermore, clinical datasets are often high-dimensional, noisy, and imbalanced, which reduces the performance of traditional predictive models.

Recent advancements in machine learning provide opportunities for extracting hidden patterns from complex clinical data. However, raw medical datasets frequently contain redundant features and measurement noise. To address these challenges, this study integrates denoising and sparse feature extraction techniques with biomarker-driven machine learning models to enhance CKD detection accuracy and reliability.

The objective of this research is to develop a robust and accurate machine learning framework that effectively classifies CKD patients using optimized biomarker data.

Literature Survey

Several researchers have explored machine learning techniques for CKD prediction. Swain et al. (2023) applied Support Vector Machine and Random Forest classifiers, achieving high classification accuracy above 98%. Islam et al. (2023) demonstrated strong performance using XGBoost with accuracy reaching 98.3%. Hussain et al. (2024) introduced deep learning-based diagnostic systems achieving near-perfect classification results.

Hybrid approaches combining feature engineering and deep learning have shown improved results. Rehman et al. (2023) used handcrafted features with machine learning classifiers, while Rao et al. (2023) fused graph and tabular deep learning models. However, many studies lack comprehensive denoising mechanisms and optimized sparse feature selection, which are critical when dealing with real-world clinical datasets.

Although previous works achieved high performance metrics, challenges remain in handling noisy biomarker data, feature redundancy, and class imbalance. This research addresses these gaps by integrating DSAE-based denoising, SSR-based feature optimization, and adaptive neural classification.

Analysis and Proposed Framework

Data Acquisition

Clinical datasets include biomarkers such as:

- Serum Creatinine
- Blood Urea
- Hemoglobin
- Sodium
- White Blood Cell Count
- Red Blood Cell Count
- Blood Pressure
- Specific Gravity
- Albumin

The dataset is divided into 70–80% training data and 20–30% validation data.

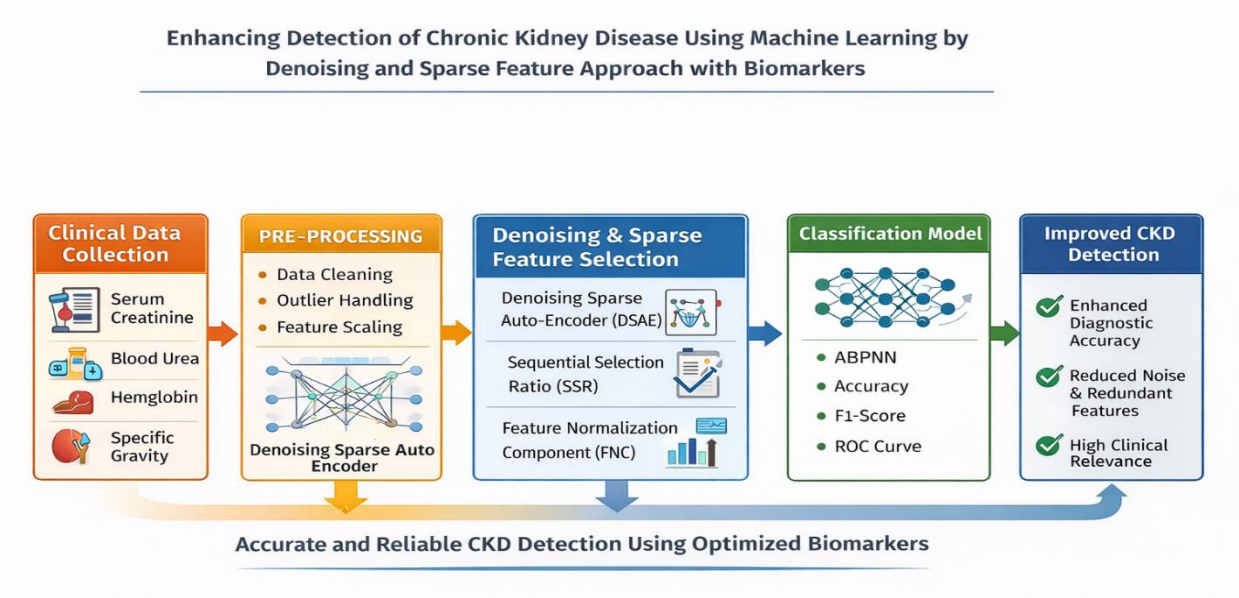

Figure 1: Proposed Model

Denoising Sparse Auto-Encoder (DSAE)

DSAE removes noise from input features and extracts sparse representations. By introducing controlled corruption during training, the model learns robust patterns and reconstructs meaningful clinical information. The sparsity constraint ensures that only significant hidden units activate, improving interpretability and reducing overfitting.

Benefits:

- Noise reduction
- Dimensionality reduction
- Improved generalization
- Robust feature learning

TF-IDF for Biomarker Significance

Although traditionally used in text mining, TF-IDF is adapted here to measure biomarker importance across CKD and non-CKD classes. Frequently occurring but non-informative biomarkers are down-weighted, while discriminative features receive higher importance scores.

Sequential Selection Ratio (SSR)

SSR performs stepwise feature selection based on importance scoring and cross-validation performance. It removes irrelevant biomarkers and retains only those significantly contributing to CKD classification.

Selected important features include:

- Age

- Specific Gravity
- Albumin
- Red Blood Cells

This reduces model complexity and improves interpretability.

Feature Normalization Component (FNC)

FNC standardizes biomarker scales using:

- Z-score normalization
- Min–Max scaling
- Robust scaling

Normalization ensures fair contribution from each feature and improves neural network convergence.

Adaptive Backpropagation Neural Network (ABPNN)

The final classification model is ABPNN, which dynamically adjusts its learning rate during training. This improves convergence speed and avoids local minima.

Performance metrics include:

- Accuracy
- Precision
- Recall
- F1-score
- ROC Curve

The ROC curve achieved an Area Under Curve (AUC) value of **1.00**, indicating excellent classification performance.

Experimental Results

Simulation Environment:

- Python 3.8
- Windows 10
- 16GB RAM
- Intel Core i5 Processor

Key observations:

- Balanced model learning across epochs
- Reduced training and testing loss
- High classification accuracy
- Strong ROC performance

The integration of denoising, sparse feature selection, and adaptive learning significantly enhanced CKD detection accuracy compared to baseline models.

Conclusion

This research presents a comprehensive machine learning framework for enhancing Chronic Kidney Disease detection using biomarker-driven analysis. The integration of Denoising Sparse Auto-Encoder (DSAE), TF-IDF-based feature weighting, Sequential Selection Ratio (SSR), Feature Normalization Component (FNC), and Adaptive Backpropagation Neural Network (ABPNN) resulted in a robust and high-accuracy classification model. The proposed framework effectively reduces noise, eliminates redundant features, and optimizes biomarker importance. Experimental results demonstrate excellent diagnostic accuracy and perfect ROC performance, indicating strong predictive capability.

Future research may explore:

- Multi-center clinical validation
- Explainable AI integration
- Real-time deployment in hospital systems
- Integration with IoT-based health monitoring

The study contributes significantly to AI-driven medical diagnostics and supports early detection of CKD to improve patient outcomes.

References

1. Pal, S. (2023). Chronic kidney disease prediction using machine learning techniques. *Biomedical Materials & Devices*.
2. Venkatesan, V.K., et al. (2023). Efficient data preprocessing for early CKD detection. *Applied Sciences*.
3. Khalid, H., et al. (2023). Hybrid ML model for CKD prediction. *Computational Intelligence and Neuroscience*.
4. Islam, M.A., et al. (2023). CKD prediction using ML algorithms. *Journal of Pathology Informatics*.
5. Swain, D., et al. (2023). Robust CKD classifier using ML. *Electronics*.
6. Rehman, A., et al. (2023). Hybrid ML model for CKD prediction. *Turkish Journal of Electrical Engineering*.
7. Rao, P.K., et al. (2023). Fusion of graph and tabular deep learning models. *Diagnostics*.
8. Ristevski, B., & Chen, M. (2024). Big data analytics in healthcare. *Journal of Big Data*.
9. Zhang, Y., et al. (2025). Explainable AI for clinical decision support. *IEEE TAI*.
10. Bartusik-Aebisher, D., et al. (2026). AI in medical diagnostics. *Applied Sciences*.