The Eastasouth Journal of Information System and Computer Science
Vol. 2, No. 01, August 2024, pp. 162-172
ISSN: 3025-566X, DOI: 10.58812/esiscs.v2i01

Data-Driven Predictive Modeling for Early Diagnosis and Risk Stratification of Chronic Diseases

Md Jubayar Hossain[1], Mohammed Majid Bakhsh[2], Muhammad Adnan[3], Evha Rozario[4], Sraboni Clara Mohonta[5], Tushar Roy[6], Md Imtiaz Faruk[7], Emran Hossain[8]
[1] St. Francis College, Brooklyn, New York, USA
[2] Washington University of Science and Technology, Alexandria, Virginia, USA
[3] Trine University, Detroit, Michigan, USA
[4] International American University, Los Angeles, California, USA
[5] Baruch College, The City University of New York, New York, USA
[6] St. Francis College, Brooklyn, New York, USA
[7] St. Francis College, Brooklyn, New York, USA
[8] Humphreys University, Stockton, California, USA

Article Info

Article history:

Received, 10 June, 2024
Revised, 25 July, 2024
Accepted, 15 August, 2024

Keywords:

Chronic disease;
Chronic Kidney Disease (CKD) dataset;
Data-Driven Predictive Modeling;
Early Disease Diagnosis;
Machine Learning;
ResNet50–SVM Classification

ABSTRACT

Chronic kidney disease (CKD) often remains undetected until advanced stages, limiting opportunities for timely intervention and increasing the burden of complications and treatment. This study proposes a data-driven hybrid ResNet50-SVM framework for the early diagnosis and risk stratification of CKD. Clinical data were prepared through missing-value treatment, data cleaning, Min-Max normalization, reduction of redundant information, and optimized feature selection. The processed dataset was divided into training and testing subsets using an 80:20 ratio. ResNet50 was employed to extract high-level representations, while a support vector machine performed the final classification. Model performance was evaluated using accuracy, precision, recall, F1-score, confusion-matrix analysis, and training and validation behavior. The proposed framework achieved an accuracy of 98.76%, precision of 97.65%, recall of 99.21%, and F1-score of 98.34%. It also outperformed logistic regression, a convolutional neural network, and a decision tree across the reported evaluation metrics. The high recall indicates a strong capacity to identify patients with disease while reducing missed positive cases. These findings demonstrate the potential of hybrid deep-learning and machine-learning methods to support accurate, scalable, and timely CKD screening. Further validation using larger, more diverse clinical datasets and explainable AI methods is needed before deployment in real-world clinical decision-support systems.

Corresponding Author:

Name: Md Jubayar Hossain
Institution: St. Francis College, Brooklyn, New York, USA
Email: jubayar.sfc@gmail.com

1. INTRODUCTION

Chronic diseases are a significant global healthcare issue that directly impacts death rates. Reducing risk factors associated with chronic diseases and the likelihood of deaths is crucial, since these illnesses often have substantial treatment costs that severely deplete patients' income [1], [2], [3]. A significant worldwide health problem is the incidence of chronic diseases. About seven of the world's ten leading causes of mortality in 2019 were chronic diseases, according to the WHO. More than 63% of all deaths globally are caused by chronic diseases [4]. Hypertension, diabetes, and heart disease are prevalent chronic diseases that mostly stem from people's own unhealthy lifestyle choices. A variety of vital organs, including those in the eyes, brain, heart, kidneys, and others, are vulnerable to damage from several chronic diseases. This can cause a chain reaction of serious problems that affect both one's personal and professional life. Patients already dealing with chronic conditions are at a much higher risk of complications from infectious diseases like the 2019 coronavirus disease (COVID-19). Severe symptoms are more common in over 48% of COVID-19 individuals with a background of chronic diseases. The cost of medical care will also increase with chronic diseases. The Centers for Disease Control and Prevention state that chronic diseases account for the majority of the nation's 3.8 trillion annual health care costs [5], [6], [7].

The key to successful treatment and better health outcomes from chronic diseases is early identification. When chronic diseases are recognised on time, it may lead to improved care and fewer associated complications [8], [9], [10], therefore minimizing the related health expenses. Risk stratification assists healthcare providers in prioritizing high-risk patients, instituting preventive interventions, and optimizing treatment strategies. Predictive models can estimate the severity and risk of disease progression by concurrently measuring multiple clinical indicators, thus enhancing the management of patients and aiding personalized healthcare delivery. Data-driven predictive modeling is particularly important in the diagnosis of chronic diseases and risk stratification [11], [12], [13]. Predictive models ML and DL methods can uncover relevant associations between clinical variables, which can better predict the occurrence and progression of a disease. Predictive frameworks by combining innovative algorithms and health records can identify latent trends, increase the diagnostic database, and facilitate the timely intervention. Early detection of high-risk patients and support of evidence-based therapeutic decisions are two measures in which such methods can contribute to more efficient healthcare systems. As a result, predictive modeling methods have become a fundamental part of the process of creating intelligent healthcare systems aimed at managing chronic diseases. This study's significance stems from the rising incidence and severity of CKD, especially in the elderly, where early diagnosis is essential for efficient treatment and better patient outcomes. The need for automated, accurate, and scalable diagnostic solutions is driven by the fact that traditional procedures are often labour-intensive, expensive, and susceptible to human error. A strong possibility to improve diagnosis precision using real-world clinical data has arisen with the progress of AI and DL, especially CNN like the hybrid model. The purpose of this study is to use these technologies to help medical personnel make timely and accurate predictions, which will eventually improve patient care and disease management. The following are the study's contributions:

a. The research paper presents an innovative use of the hybrid model in early detection of CKD, which allows efficient and automated recognition of chronic kidney disease (CKD) datasets.
b. Adopts an end-to-end system, which gets rid of manual feature engineering using deep convolutional layers to do automatic and hierarchical feature extraction.
c. Provides high-quality model input through a highly developed preprocessing pipeline, which consists of optimized feature selection, Min-Max scaling normalization, and missing value treatment.
d. Compare the proposed model with other more traditional ML models

regarding accuracy, precision, recall, F1score; concludes that the latter two are more effective and reliable to be used in clinical diagnostic predictions.

Organization of The Paper

The paper's framework is as follows: Section II offers a literature overview on chronic disease detection; Section III delves into the dataset and methods; The experimental data and their interpretation are presented in Section IV, and Section V concludes with the most relevant discoveries and suggestions for further research.

2. LITERATURE REVIEW

This section will provide a literature review on the available research on predictive modeling of chronic diseases using data, specifically chronic kidney disease (CKD). Table I provides a summary of the reviewed studies by mentioning their methodology, challenges and future research directions.

Clinical Notes, Demographic Data, Medical Records, test results, and the whole dataset were used to construct and evaluate experimental models. Elements of DT, RF, SVM, and GB are used as ML techniques. The testing findings demonstrated that all algorithms had excellent prediction accuracy, with gradient boosting reaching a maximum accuracy of 90%. Evaluations of the AUC-ROC ranged from 0.82 to 0.92. Blood pressure, age, glomerular filtration rate, serum creatinine, and urine protein levels were shown to be linked with the development of CKD in the variable significance analysis [12].

Implementing ML techniques is essential for accurately classifying individuals with chronic diseases. Data mining is a powerful tool for disease prediction and broad prognosis based on patient symptoms. CNN and other ML algorithms automate feature extraction and disease prediction, while KNN uses distance calculations to find the exact match between the dataset and the final disease prediction result. A wide range of disease symptoms and the individual's lifestyle choices were used to compile the data set with the purpose of making a broad disease prediction [8]. Study accepts CKD is common, complicated, and expensive. The need for renal care is on the rise, one solution is a finite-horizon MDP model that uses data and evidence to determine when patients should have their follow-up appointments. It is parameterized using information from 68,513 CKD patients across 11 US Department of Veterans Affairs facilities. It provides an ideal monitoring strategy based on a number of approximations and hypotheses. It was discovered that the severity of CKD, comorbidities, age, and distance to a nephrologist all influence how patients need treatment [13].

Integrated hypertension and diabetes risk assessment algorithms powered by ML, specifically designed for the vulnerable population screened in community-based initiatives in areas with limited resources. The RF model, which included data from 22,78 patients gathered by community health workers, outperformed the US and UK risk scores for diabetes by 35.5% in terms of AUC and had the best prediction accuracy for both diseases [14]. A solid dataset allows a ML system to diagnose the stage of this deadly disease considerably more quickly. To get extremely accurate prediction results, four dependable algorithms were employed, including Support Vector Machine (henceforth SVM) and AdaBoost (henceforth AB). The online dataset of the UCI ML repository is used to develop these methods. Performance assessment measures have also been given to illustrate suitable results, with GB classifiers obtaining the greatest predicted accuracy of roughly 99.80% [15]. The analytical accuracy of the prediction for early disease detection is correlated with the state of the medical data, therefore a bad state of the medical data results in a lower prediction accuracy. The RNN algorithm provides the highest prediction accuracy of approximately 97.6% and the convergence speed. The prediction procedure is carried out utilizing the dataset that is made available online from certain hospitals [16].

Table 1. Summary of The Related Work on Early Detection Og Chronic Disease Using Machine Learning

Author & Year	Dataset	Key Findings	Challenges	Limitations	Future Work
[12]	CKD dataset containing demographic data, medical history	Gradient Boosting achieved the highest accuracy of 90% with AUC values between 0.82–0.92	Managing large clinical datasets and identifying significant medical features	Limited dataset diversity and reliance on traditional ML methods	Explore deep learning techniques and larger medical datasets
[8]	Dataset containing disease symptoms	Demonstrated improved disease prediction using symptom-based data analysis	Complexity in handling heterogeneous medical data	Model evaluated mainly on symptom-based datasets	Integrate more clinical indicators and real-time healthcare data
[13]	Data from 68,513 CKD patients	Identified key factors such as CKD severity, comorbidities, age, and distance to healthcare providers affecting patient care	Managing large healthcare datasets and designing optimal monitoring strategies	Model focused primarily on scheduling rather than prediction	Extend models for predictive risk assessment and personalized treatment planning
[4]	Electronic health records, medical imaging	AI systems identified high-risk individuals with AUC between 0.80–0.88 across multiple chronic diseases	Integrating diverse healthcare data sources	Predictive performance varies across demographic groups	Develop more generalized AI models and improve fairness across populations
[14]	Community-based health screening dataset	Random Forest achieved the highest prediction performance and improved AUC by 35.5% compared to US and UK risk scores	Data collection challenges in low-resource healthcare settings	Limited sample size and population diversity	Expand models to larger populations and integrate mobile health data
[15]	UCI Machine Learning Repository dataset	Gradient Boosting achieved 99.80% prediction accuracy, demonstrating strong classification performance	Model generalization across different datasets	Evaluation mainly based on a single dataset	Test models on real-world clinical datasets

3. METHODS

Figure 1 shows the systematic and controlled approach applied in early detection of CKD. It starts with the gathering of real-world clinical information, which is followed by thorough preprocessing. This incorporates the cleaning of data, to deal with missing or inconsistent data, and also, data reduction to remove irrelevant or redundant data. Min-Max normalization is used to normalize the preprocessed data such that the feature scales are comparable, and feature selection is optimized to increase the accuracy and efficiency of the model. When the data is ready it is split into training and testing sets to develop a model. The processed data are then classified using a hybrid ML model which is a combination of ResNet50 and SVM. Deep feature extraction is done by ResNet50, whereas the final classification is done by SVM. Accuracy, precision, recall, and F1-score are typical classification metrics used to evaluate the efficacy of every model. The step-by-step methodology provides a well-developed and precise prediction system that will guarantee the timely diagnosis and improved healthcare outcomes through the early detection of CKD.

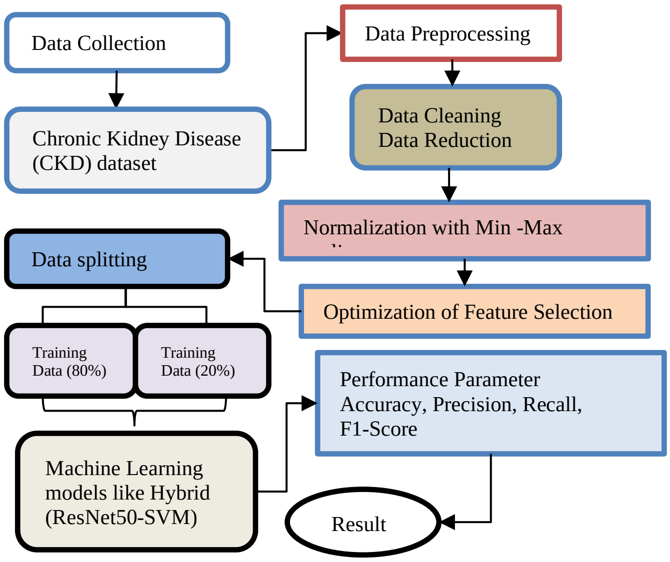

Figure 1. Flowchart for Chronic Disease Detection

Each step of proposed flowchart of methodology is briefly explained below:

3.1 Data Collection

In this study on chronic disease prediction, the CKD dataset was utilized to determine the disease status of patients. This real-world dataset, collected from Apollo Hospital in Tamil Nadu, India, spans a two-month period and is publicly available through the UCI ML Repository. Specifically curated for CKD research, it contains records of 400 patients, each described by 25 relevant medical attributes. The dataset includes a mix of numerical, decimal, and categorical values, reflecting diverse clinical measurements and patient information essential for accurate disease prediction.

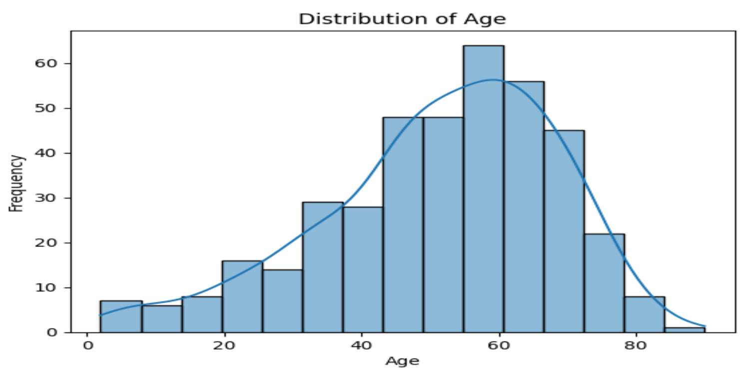

Figure 2. Distribution of Age

The age distribution of people included in the dataset with the help of a density curve in a histogram is shown in Figure 2. The y-axis denotes frequency with the x-axis being the age. The majority of people are aged between (45) 70 years with the highest concentration being at the ages of 55-65 years. There are fewer samples at old age and at younger ages.

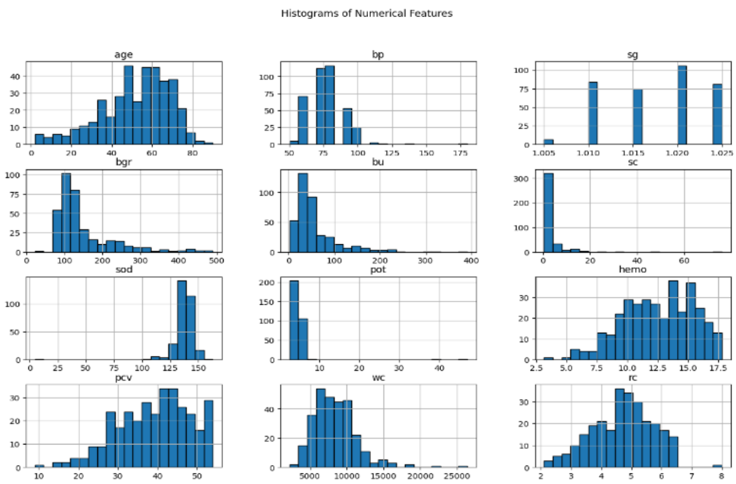

Figure 3. Histogram of Numerical Feature

Histograms of multiple numerical features from the CKD dataset, including age, BloodPressure (bp), SpecificGravity (sg), blood glucose random (bgr), BloodUrea (bu), Serum Creatinine (sc), sodium (sod), potassium (pot), hemoglobin (hemo), packed cell volume (pcv), White Blood Cell count (wc), and Red Blood Cell count (rc) is shown in Figure 3. There are trends and variances in patient health indicators throughout the dataset, and each histogram shows the frequency distribution of the corresponding variable.

3.2 Data Preprocessing

The data preprocessing stage ensures proper preparation of the CKD dataset through data cleaning and normalization. These steps improve data quality and prepare the dataset for effective model training. Additional pre-processing procedures are implemented in the subsequent stages.

a. Data Cleaning: Data cleaning is performed to improve the quality and reliability of the CKD dataset by removing inconsistencies, handling missing values, and correcting incorrect entries.
b. Data reduction: Data reduction involves removing duplicates in addition to reducing the data's dimensions. The goal of data reduction in CKD research is to minimize the dimensionality of the dataset by eliminating characteristics that are either unnecessary, redundant, or noisy. This lowers computing costs while improving the efficiency of ML models.

3.3 Data normalization with Min-Max scaling

Feature value rescaling's data normalization approach. Raw data often has a wide range of values with varying scales. Data normalization, specifically min-max normalization, is essential for the efficient operation of objective functions in some ML algorithms. This approach is essentially a straightforward way to scale the feature value range to either [0, 1] or [−1, 1]. Data characteristics dictate the choice of the target range. The general formula for a minmax of [0, 1] is given in Equation (1):

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}, \qquad (1)$$

Where x is an Original Value and x′ is the Normalized Value.

3.4 Optimization of Feature Selection

Feature Selection is an optimization method that uses a subset of relevant features to increase classification accuracy by removing unnecessary features from the original feature space. This study employed four ranking-based FSTs to rate the attributes. The ranking-based Feature Selection method is used to determine the importance of characteristics based on their ranking.

3.5 Data Splitting

The dataset is divided into two parts: the training set and the testing set, using an 80/20 split. This random split ensured that 80% of the data was used for training the model, while 20% was used to assess its performance.

3.6 Classification with hybrid (ResNet50-SVM) Models

ResNet-50 is a CNN form of Residual Network (ResNet) with 50 layers. Additionally, it has 48 Convolution Layers, 1 MaxPool layer, and 1 Average Pool Layer. A hybrid model combining ResNet-50 and SVM is widely used for chronic disease classification because it integrates powerful deep feature extraction with a robust machine-learning classifier. In this approach, ResNet50 acts as a feature extractor that processes the input medical image III and generates a high-level feature representation. These deep characteristics describe significant shapes on diseases like tumor, cyst, or other abnormalities. The feature extraction process can be expressed as Equation (2):

$$F = f_{ResNet50}(I) \quad (2)$$

Where I represents the input image and F denotes the extracted deep feature vector. ResNet50 avoided the vanishing gradient problem through skip connections, which is achieved through residual learning. The residual mapping used in the network is defined as Equation (3):

$$y = F(x) + x \quad (3)$$

Where x is the input feature map and F(x) represents the learned residual function. After feature extraction, the resulting feature vector is passed to the SVM classifier for disease prediction. By maximizing the margin between various disease groups, the SVM creates an ideal hyperplane that divides them. The decision function of the SVM classifier is defined as Equation (4).

$$f(x) = w^T x + b \quad (4)$$

Where w denotes the WeightVector, x is the FeatureVector obtained from ResNet50, and b is the bias term. The final classification of chronic disease is obtained by applying the sign function Equation (5).

$$y = \text{sign}(w^T f_{ResNet50}(I) + b) \quad (5)$$

Where y represents the predicted disease class (e.g., Normal, Cyst, Tumor, or Stone). This hybrid ResNet50–SVM framework improves classification.

3.7 Model Evaluation

The assessment measure is critical during classification training since it determines the best classifier. Thus, selecting a proper assessment measure is critical for differentiating and finding the best classifier. The five assessment measures used were accuracy, precision, recall, F1 Score, and Confusion Matrix. True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) are four crucial elements in the equation (Equations 6 to 9).

a. When the model successfully detects a positive example, it is called a true positive.
b. A false positive occurs when the model improperly assigns a positive label to a negative instance.
c. Correctly detected negative situations are referred to as true negatives.
d. When a positive example is wrongly classified as negative by the model, this is known as a false negative.

Accuracy: Accuracy is a crucial parameter in classification models, providing a measure of overall performance. This may be expressed numerically as Equation (6).

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad (6)$$

Precision: Precision, a popular metric known as the positive predictive value, indicates how well the model makes positive predictions. The equation below illustrates this mathematically (7):

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (7)$$

Recall: Recall, often referred to as sensitivity or TPR, provides an evaluation of the model's ability to correctly identify positive cases relative to the real positive examples found in the dataset. The following equation is used to define this connection (8):

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (8)$$

F1 Score: The F1score gives a reasonable evaluation of the two measures as it is the harmonic mean of rec and prec. In mathematics, the F1score is calculated using Equation (9):

$$\text{F1} - \text{Score} = 2 * \frac{(\text{Precision}*\text{Recall})}{(\text{Precision}+\text{Recall})} \quad (9)$$

Accuracy, precision, recall, and F1 score are all evaluated using these four parameters. A number between zero and one is assigned to each of these criteria.

4. RESULTS AND DISCUSSION

The machine used for all testing has an IntelCore i7 CPU running at 3.4 GHz, 16GB of RAM, and a 4GB NVIDIA GeForceGTX 1650GPU. The Scikit-learn package was used in the Python implementation. This paper implements the proposed model using Python in both Jupyter Notebook and Visual Studio Code, with the help of the following libraries: Pandas, NumPy, Scikit-learn, Matplotlib and Seaborn, used to process data, train the model, and visualize it. The suggested Hybrid (ResNet50SVM) model is evaluated using experimental data derived from CKD in Table 2. The data of the experiment indicate that the classification is very good with the acc of 98.76, the prec of 97.65, the rec of 99.21, and the F1score of 98.34. The results indicate that the hybrid model may improve the overall categorization of chronic diseases and can correctly detect all instances of disease.

Table 2. Experiment Results of Proposed Models for Chronic Disease Detection

Classification	Hybrid (ResNet50+SVM)
Accuracy	98.76
Precision	97.65
Recall	99.21
F1-score	98.34

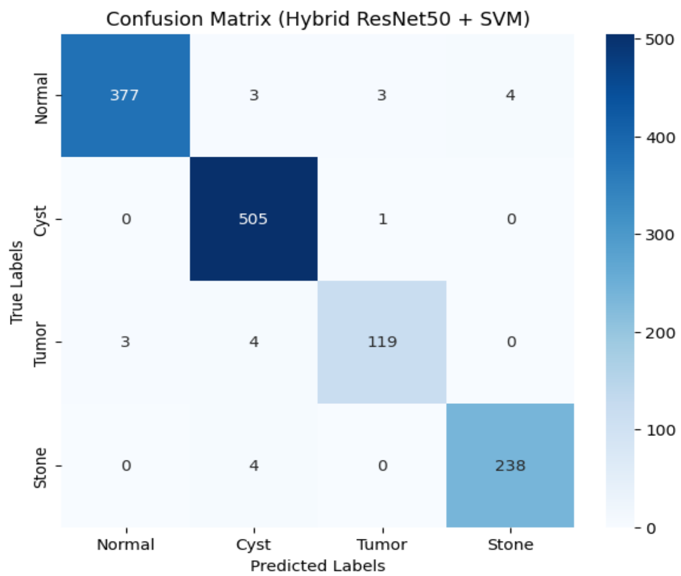

Figure 4. Confusion Matrix of Hybrid Model

Chronic diseases are classified using the hybrid ResNet50SVM model's confusion table. The four predicted and actual labels of four classes, namely Normal, Cyst, Tumor, and Stone is shown in Figure 4. The correct classification is shown along the diagonal with 377, 505, 119, 238 respectively whereas the number of misclassifications is limited to a few, some being 3 and 4 between related classes.

Training and validation accuracy of the hybrid ResNet50-SVM model with varying epochs in Figure 5. The training accuracy levels off at approximately 0.88 up to almost 0.98, but validation accuracy remains comparatively steady with slight variations. The curves are approaching the later epochs, which signifies the stable learning behavior and high model generalization performance.

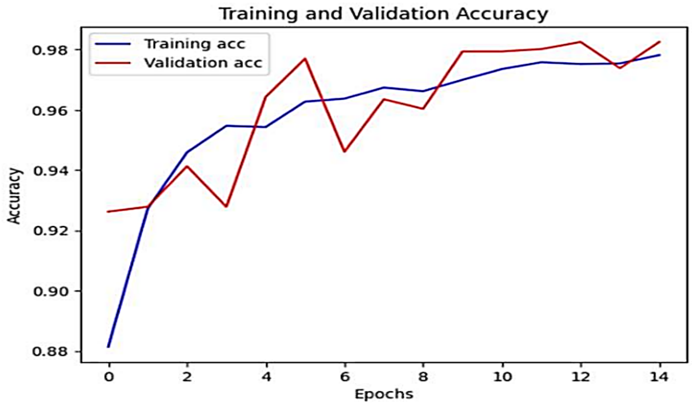

Figure 5. Training and Validation Accuracy Curve of Proposed Model

Figure 6. Training and Validation Loss Curve of Proposed Model

The training and validation error of the hybrid ResNet50-SVS model with various epochs in Figure 6. Training loss reduces gradually, with the range being about 0.33 to about 0.06, and validation loss also falls with minimal variation. The two curves meet in later epochs, which shows that learning, less error, and stability of the model during training.

4.1 Comparative Analysis

The performance comparison of various ML and DL models for chronic disease classification based on key evaluation metrics is shown in Table III. The evaluated models include LR, CNN, Decision Tree, and the proposed hybrid model. The LR model achieves about 96% acc with moderate prec and rec values, while the CNN model improves classification performance with high rec and F1score. The Decision Tree model shows comparatively lower 91.75% accuracy with balanced precision–recall performance. Among all models, the proposed hybrid model achieves the best performance with an acc of 98.76%, prec of 97.65%, rec of 99.21%, and an F1score of 98.34%, demonstrating its effectiveness and reliability for chronic disease classification.

Table 3. Model Performance Comparison of Different Evaluation Metrics on Chronic Kidney Disease Dataset

Model	Accuracy	Precision	Recall	F1 Score
LR [17]	96.02	0.887	0.857	0.871
CNN [18]	95	94	99	98
Decision Tree [19]	91.75	85.02	94.66	89.58
Proposed model	98.76	97.65	99.21	98.34

The suggested framework DL and ML approaches to the classification of chronic diseases through the hybrid ResNet50, SVM model to combine the high accuracy and reliable prediction. The model combines ResNet50 to extract deep features and SVM as effective disease-categories classifiers. The suggested method has an accuracy of 98.76% showing a good ability to detect various chronic disease conditions. By combining a powerful deep feature representation with a powerful margin-based classifier, this hybrid system improves classification performance and produces more accurate and secure medical diagnoses.

4.2 *Justification and Novelty*

The paper's originality comes from its use of a DL-based method to identify early-stage CKD using actual clinical data. Deep hierarchical learning and autonomous feature extraction are the mainstays of the suggested model, which can discover intricate patterns in medical data without the human intervention required by conventional ML algorithms. The practical utility of the study is further supported by the use of a standardised and widely recognised healthcare dataset. By offering a scalable, data-driven model that may be used for comparable medical categorization tasks, this strategy represents a significant advancement in predictive analytics for early chronic disease detection.

5. CONCLUSION

The early detection of chronic disease is essential for the improvement of patient outcomes and the reduction of medical complications. Nevertheless, people do not pay regular attention to their health due to their busy schedules, and thus, most often, they visit the doctor after symptoms become critical. In this research, a predictive modeling model is proposed which is a data-driven predictive model that is founded on a hybrid ResNet50SVM model that is used to predict and stratify the risk of CKD on the CKD dataset. The given strategy is an integration of deep feature extraction (ResNet50) and the prediction ability of SVM to enhance the accuracy of prediction. The experimental findings prove that the hybrid model has high performance with an accuracy of 98.76, data-based structure in the discovery of complicated clinical patterns and predicting CKD reliably, which can help health care specialists timely to diagnose.

The robustness and dependability of medical models may be improved via future research by using more and more varied datasets. Additionally, other advanced sophisticated deep learning networks like Efficient Net or DenseNet which can be further improved to enhance the level of prediction. Transparency and trust in clinical decision-making would be improved by the integration of XAI techniques, too. Moreover, the creation of a real-time clinical decision support system, which will be created on the basis of the suggested predictive framework, will allow for practical application in healthcare facilities and developing better management of chronic diseases.

REFERENCES

[1] S. S. Shambharkar, P. S. Moon, P. A. Bainalwar, and S. M. Boarkar, "Machine Learning-Based Approach for Early Detection and Prediction of Chronic Diseases," in *2023 1st DMIHER International Conference on Artificial Intelligence in Education and Industry 4.0 (IDICAIEI)*, IEEE, Nov. 2023, pp. 1–8. doi: 10.1109/IDICAIEI58380.2023.10406914.

[2] S. Singamsetty, "Neurofusion: Advancing Alzheimer's Diagnosis with Deep Learning and Multimodal Feature Integration," *Int. J. Educ. Appl. Sci. Res.*, vol. 8, no. 1, pp. 23–32, 2021, doi: 10.5281/zenodo.14889013.

[3] B. Khan, R. Naseem, F. Muhammad, G. Abbas, and S. Kim, "An empirical evaluation of machine learning techniques for chronic kidney disease prophecy," *IEEE access*, vol. 8, pp. 55012–55022, 2020, doi: 10.1109/ACCESS.2020.2981689.

[4] M. F. Alam and M. F. Alam, "Predictive Artificial Intelligence Models for Early Detection and Management of Chronic Diseases To Strengthen National Healthcare Resilience in the United States," *Am. J. Interdiscip. Stud.*, vol. 03, no. 04, pp. 268–293, 2022, doi: 10.63125/9t5ar104.

[5] X. Yuan, S. Chen, C. Sun, and L. Yuwen, "A novel early diagnostic framework for chronic diseases with class imbalance.," *Sci. Rep.*, vol. 12, no. 1, p. 8614, May 2022, doi: 10.1038/s41598-022-12574-x.

[6] S. Nusinovici *et al.*, "Logistic regression was as good as machine learning for predicting major chronic diseases," *J. Clin. Epidemiol.*, vol. 122, pp. 56–69, 2020, doi: 10.1016/j.jclinepi.2020.03.002.

[7] S. Agarwal, C. Prabha, and M. Gupta, "Chronic diseases prediction using machine learning – a review," *Ann. Rom. Soc. Cell Biol.*, vol. 25, no. 1, pp. 3495–3511, 2021.

[8] R. Alanazi, "Identification and Prediction of Chronic Diseases Using Machine Learning Approach," *J. Healthc. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/2826127.

[9] J. Antony, P. Palsuk, S. Gupta, D. Mishra, and P. Barach, "Six Sigma in healthcare: a systematic review of the literature," *Int. J. Qual. Reliab. Manag.*, vol. 35, no. 5, pp. 1075–1092, 2018, doi: 10.1108/IJQRM-02-2017-0027.

[10] F. M. Delpino, Â. K. Costa, S. R. Farias, A. D. P. Chiavegatto Filho, R. A. Arcêncio, and B. P. Nunes, "Machine learning for predicting chronic diseases: a systematic review," *Public Health*, vol. 205, pp. 14–25, Apr. 2022, doi: 10.1016/j.puhe.2022.01.007.

[11] V. Varma, "Data Analytics for Predictive Maintenance for Business Intelligence for Operational Efficiency," *Asian J. Comput. Sci. Eng.*, vol. 7, no. 4, pp. 1–7, 2022.

[12] S. P. Abhang *et al.*, "Machine Learning-Based Monitoring and Prognosis of Chronic Kidney Disease Patients," in *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)*, 2023, pp. 1–6. doi: 10.1109/ICAIIHI57871.2023.10489422.

[13] Z. Nenova and J. Shang, "Personalized Chronic Disease Follow-Up Appointments: Risk-Stratified Care Through Big Data," *Prod. Oper. Manag.*, vol. 31, no. 2, pp. 583–606, 2022, doi: 10.1111/poms.13568.

[14] J. J. Boutilier, T. C. Y. Chan, M. Ranjan, and S. Deo, "Risk Stratification for Early Detection of Diabetes and Hypertension in Resource-Limited Settings: Machine Learning Analysis," *J. Med. Internet Res.*, vol. 23, no. 1, 2021, doi: 10.2196/20123.

[15] P. Ghosh, F. M. Javed Mehedi Shamrat, S. Shultana, S. Afrin, A. A. Anjum, and A. A. Khan, "Optimization of Prediction Method of Chronic Kidney Disease Using Machine Learning Algorithm," in *2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 2020, pp. 1–6. doi: 10.1109/iSAI-NLP51646.2020.9376787.

[16] P. Anandajayam, S. Aravindkumar, P. Arun, and A. Ajith, "Prediction of Chronic Disease by Machine Learning," in *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, 2019, pp. 1–6. doi: 10.1109/ICSCAN.2019.8878724.

[17] Z. Ullah *et al.*, "Detecting High-Risk Factors and Early Diagnosis of Diabetes Using Machine Learning Methods," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/2557795.

[18] M. Pal, S. Parija, and G. Panda, "Improved prediction of diabetes mellitus using machine learning based approach," in *2nd International Conference on Range Technology, ICORT 2021*, 2021. doi: 10.1109/ICORT52730.2021.9581774.

[19] M. R. Narasinga Rao, T. Sajana, N. Bhavana, M. Sai Ram, and C. Nikhil Krishna, "Prediction of chronic kidney disease using machine learning technique," *J. Adv. Res. Dyn. Control Syst.*, vol. 10, no. 10, pp. 328–332, 2018, doi: 10.17148/IJARCCE.2018.71021.