International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group | Volume 5, Issue 2, 136-146, 2024
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X.IJERET-V5I2P114

Original Article

Early Diseases Diagnosis of Chropnic via Machine Learning Based Models in Big Data Health Records

Venkata Teja Nagumotu[1], Harsha Vardhan Reddy Kavuluri[2], Akhil Kumar Pathani[3], Ajay Dasari[4], Venkata Kishore Chilakapati[5], Srikanth Reddy Keshireddy[6]
[1]Sr Network Engineer, Techno-bytes Inc .
[2]Lead database administrator, Wissen infotech Inc.
[3]Network Engineer, Ebay.
[4]Senior Support Engineer, Microsoft.
[5]Support Escalation Engineer, Microsoft.
[6]Senior Software Engineer, .Keen Info Tek Inc.

Abstract - *Today's people suffer from a wide variety of diseases due to various influences and choices made at the community level. Thus, to prevent the occurrence of such illnesses, persistent identification and prediction are paramount. Manually determining the disorders is generally challenging for doctors to be accurate with the exact numbers. Using massive data extracted from EHRs, this research lays forth an effective machine learning (ML) approach for CKD early diagnosis. Data preparation steps (including outlier removal, missing value replacement and transforming categorical data) are done before using normalization and RFE to find the best features. ETC is used as the main classification model because it helps to improve prediction and reduces the chances of overfitting by splitting the data randomly. With an accuracy (ACC) of 99.5%, the model is very effective in diagnosing CKD. When evaluation measures include precision (PRE), recall (REC), F1-score (F1), and AUC-ROC, it shows that the approach performs well. They prove that using machine learning and big data together can enhance how early diagnosis and decisions are made in chronic disease cases.*

Keywords - *Chronic Kidney Disease Detection, Early Detection, Deep Neural Network, Chronic Kidney Disease (CKD) Dataset, Predictive Analytics, Health Informatics.*

1. Introduction

Worldwide healthcare systems are facing severe challenges because of people living with chronic diseases. The most common chronic diseases and deaths today are cardiovascular disease, diabetes, metabolic syndrome, hypertension, and heart failure [1]. The World Health Organisation (WHO) has indicated that chronic illnesses account for around 71% of all fatalities each year [2]. The increasing problem of sleep disorders shows how vital it is to find and use efficient data-based approaches to diagnosis and treatment.

Managing and analysing massive amounts of data kept in EHRs is one example of the current use of big data in healthcare, lab findings, personal information, medication records and past diagnoses [3]. These records support health professionals in seeing how diseases develop in a patient in the long run [4]. Since there is so much and so many types of data, ordinary statistical methods are usually not sufficient, so big data analytics must be used [5]. The process of predictive analytics helps detect high-risk individuals so that other health measures can be taken in advance and better results can be reached.

The earlier a condition is noticed, the better the outcomes are and the lower the healthcare costs, as many chronic diseases get worse without being noticed [6]. ML software can analyse large datasets to discover valuable patterns and insights into future risks, enabling early intervention, accurate diagnosis, and treatment tailored to the patient. There are fewer cases of illness, death and people in hospitals [7]. Because healthcare is now digital, there are now enormous databases containing EHRs, diagnostics, patient characteristics, medical backgrounds and lifestyle information [8]. Machines help overcome the issues that arise while working with these datasets. Alternatively, big data analytics makes it possible to discover key learnings and assist in building models that predict possible risks, the way a disease may advance and its possible consequences. They make medical decisions easier, result in better diagnoses and help doctors design correct treatment options for each person.

ML helps a lot in examining complicated medical records. Access to big collections of data allows it to understand disease causes better, identify high-risk patients and detect diseases early [9]. ML is also helpful in reviewing laboratory and clinical data, leading to accurate forecasts. This method mainly aims to diagnose chronic diseases in people with the help of ML. Many medical procedures now use these methods, such as those for the diagnosis, prognosis, and prediction of inflammatory bowel disease, MS, autoimmune kidney disease, and autoimmune rheumatic illnesses, among many others. ML also helps choose treatment plans, divide patients into groups, make medicines, recycle medicines and explain drug targets.

1.1. Motivation and Contribution of the Study

The rising levels of CKD across the world, along with its silent progression up to advanced stages, show that there is a great need to detect it early and accurately. Using old diagnostic methods usually takes a lot of time and cannot be used on a large scale. Because electronic health records are more extensive and ML is advancing, there is a strong reason to make automated diagnosis models that help doctors with efficient and error-free decisions and better patient outcomes. Several important results came from the study:

- Put forth a tough ML approach for identifying CKD early with data acquired from electronic health records.
- Carried out required data-cleaning operations such include filling in blanks, eliminating extreme cases, and encoding categories for improved data.
- Select the most relevant characteristics for improved efficiency and interpretation using Applied RFE.
- The ETC was implemented to gain high ACC and lower computations because it splits the data randomly.
- Checked the effectiveness of the model by evaluating using metrics like REC, ACC, PRE, F1, and AUC/ROC.

1.2. Novelty and Justification of the Study

In this research, a remarkable DNN was applied to recognize CKD from many clinical features. This strategy makes use of both numbers and categories in data, which are generally dismissed in normal diagnostics that only review split-up individual symptoms. This way of thinking is needed because there is a growing number of CKD cases, and healthcare needs modern tools that use data. The ability of DNNs to find complex features and connections in the data allows the proposed solution to solve the issues brought by class imbalance and boosts ACC. The framework makes early detection of CKD possible and encourages prompt action by medical teams which results in both better chances for patients and more efficient medical services.

1.3. Structure of the Paper

The structure of the paper is explained here: Section II covers the literature study about using ML techniques to identify Chronic Disease cases in the early stages. Section III lists the methods used, such as how to collect data, process it and implement the model. Section IV The findings and the outcomes of the experiment are presented at this stage. Finally, Section V ends by giving useful information and suggesting future research topics.

2. Literature Review

This section discusses recent advances in recognizing and classifying CKD, with particular attention to the benefits of ML. Advanced methods are examined to determine how they can improve the identification of CKD. These are some of the most important review works:

Almasoud and E (2019) This study aims to examine how well ML algorithms can predict chronic renal illness with minimal data. The ANOVA, Pearson's correlation, and Cramer's V tests were among the several statistical methods employed to weed out duplicate traits. The algorithms which were trained and verified using 10-fold cross-validation were LR, SVM, RF, and gradient boosting. Maximize the F1-measure of the Gradient Boosting classifier to a 99.1 ACC. Moreover, they found that when it comes to diagnosing CKD, haemoglobin is more relevant for both gradient boosting and RF [10].

Amirgaliyev, Shamiluulu and Serek (2018) This study sought to examine the potential consequences of utilizing clinical characteristics to classify individuals with chronic renal illness using the SVM algorithm. Clinical history, physical examinations, and laboratory results provide the basis of the chronic kidney disease demographic. The results showed that when using ACC, sensitivity, and specificity as performance metrics, the accuracy rate for identifying patients with renal diseases was above 93% [11].

S et al. (2018) proposed system, it provides ML methods for accurate prediction of different disease occurrences in societies that experience common diseases. It uses data from real hospitals to evaluate the updated estimation models. It studies cerebral infarction, a persistent, localized disease. It applies Map Reduce and ML Decision Tree to healthcare data in both organized and unstructured formats. No previous studies in medical big data analytics have tackled both types of data, as far as anybody can tell. In comparison to some popular estimate methods, the calculation ACC of the proposed algorithm reaches 94.8%, and its convergence speed is faster than that of the CNN-UDRP methodology [12].

Akben (2018) A new method was introduced to detect chronic kidney disease early on automatically. With the use of the results of the urine and blood tests as well as the patient's medical history, this method aims to facilitate medical diagnosis. Prior to submitting analytical data for pre-processing, data mining techniques based on classification algorithms were employed. The initial part of the study's approach was pre-processing CKD data. The method used for pre-processing was K-Means clustering. Classification algorithms (KNN, SVM, and Naïve Bayes) were used to the pre-processed data in order to diagnose CKD. Classification methods have a maximum success rate of 97.8% (98.2% for those 35 and over) [13].

Tekale, Shingavi and Wandhekar (2018) This study suggests that chronic kidney disease (CKD) is a disease that, in some instances, manifests with no symptoms whatsoever. There is little hope for the prediction, detection, and prevention of such a

sickness, which might cause irreversible harm to health, but ML offers promise as it excels at analysis and prediction. Apply several ML algorithms, including DT, SVM, and others, on data collected from 400 records of CKD patients with 14 features. Making an ACC-optimal model for determining the presence and severity of CKD [14].

Chen et al. (2017) An innovative CNN-MDRP technique is introduced, which utilizes both structured and unstructured hospital data for multimodal disease risk prediction. They were unaware of any prior work in medical big data analytics that combined the two kinds of data. With a convergence speed that surpasses that of the CNN-UDRP algorithm, their suggested method achieves a prediction ACC of 94.8%, outperforming most traditional prediction algorithms [15].

The analysis between studies that explore their names of who wrote it, how data was collected, information used, main points, drawbacks, and Table I contains the findings of the research conducted

Table 1: Summary of Reviewed Works on Chronic Diseases Using Big Data

Author(s)	Methodology	Datasets	Key Findings	Limitations & Future Work
Almasoud & E., et.al. (2019)	A few examples of feature selection methods include ANOVA and Pearson correlation, whereas LR, SVM, RF, and Gradient Boosting are ML algorithms. An evaluation method based on 10-fold cross-validation.	Chronic Kidney Disease dataset with clinical/lab features.	Gradient Boosting achieved the highest performance with 99.1% accuracy (F1-measure). Hemoglobin identified as the most important feature for CKD prediction in both RF and GB models.	Study limited to a relatively small feature set. Future work may test deep learning models or hybrid feature selection methods on larger real-world datasets.
Amirgaliyev, et.al. (2018)	Using the SVM algorithm for patient classification in chronic kidney disease (CKD). Evaluation of performance based on sensitivity, specificity, and accuracy.	CKD database created from medical records, x-rays, and other diagnostic procedures.	SVM achieved over 93% accuracy across evaluated metrics. Demonstrated strong capability of SVM in CKD detection.	Limited to one ML algorithm (SVM). Future research could compare multiple ML/DL models and include broader clinical attributes.
S et al. (2018)	Proposed a multimodal system using ML Decision Tree and MapReduce algorithm. Worked on structured and unstructured hospital data. Predictive analysis for cerebral infarction.	Real-life hospital data including both structured and unstructured records.	Achieved 94.8% accuracy with faster convergence compared to CNN-UDRP algorithm. This is the first research on medical big data analytics to concentrate on both forms of data.	Limited to a regional chronic illness case. Future studies should evaluate scalability across multiple diseases and multi-center datasets.
Akben (2018)	Proposed an early-stage CKD diagnosis method. Pre-processing using K-Means clustering. Applied classification algorithms: KNN, SVM, Naïve Bayes. Data mining approach combining clustering + classification.	CKD dataset containing urine test results, blood test values, and patient medical history.	Highest classification success rate: 97.8% overall. For age group ≥35 years, accuracy reached 98.2%. Demonstrated that clustering-based preprocessing improves classifier performance.	Limited evaluation across only three classifiers. Future work could use deep learning models, hybrid feature selection, and larger multi-center datasets.
Tekale, et.al. (2018)	Designed ML models to predict CKD and its severity. Algorithms used: Decision Tree, SVM, and others. Focused on predicting CKD presence and severity level.	CKD patient dataset with 14 attributes and 400 records.	Demonstrated ML effectiveness in predicting CKD where symptoms are unclear. Achieved high accuracy for CKD detection using multiple ML techniques.	Dataset size is relatively small, limiting generalization. Future research may include severity grading models, deep learning approaches, and real-time clinical validation.
Chen et al. (2017)	Constructed a convolutional neural network (CNN) disease risk prediction system (CNN-MDRP).	Hospital datasets with diverse data types.	Achieved 94.8% prediction accuracy, outperforming CNN-UDRP. Provided faster	Computationally intensive. Future work could optimize model efficiency and test on larger, heterogeneous

	Combined organized and unstructured healthcare records.		convergence and improved multimodal learning.	medical big data environments.

3. Methodology

Detail of the process for early detection of CKD using ML is shown in Figure. 1. Initially, the data is organized, where all missing values, strange points and categories are addressed and changed into numbers. After that, data normalization is done using a standard scaler to maintain equal scaling for all features and the most important characteristics are selected using RFE. Next, divide the dataset in half, with 75% serving as training data and 25% as test data. The Extra Tree Classifier (ETC) model is utilised in the classification section because of its efficiency and ability to minimise overfitting. ACC, PRE, REC, F1, and AUC/ROC are some of the performance indicators that are used to evaluate the model, which cover all important aspects of how accurately the model predicts.

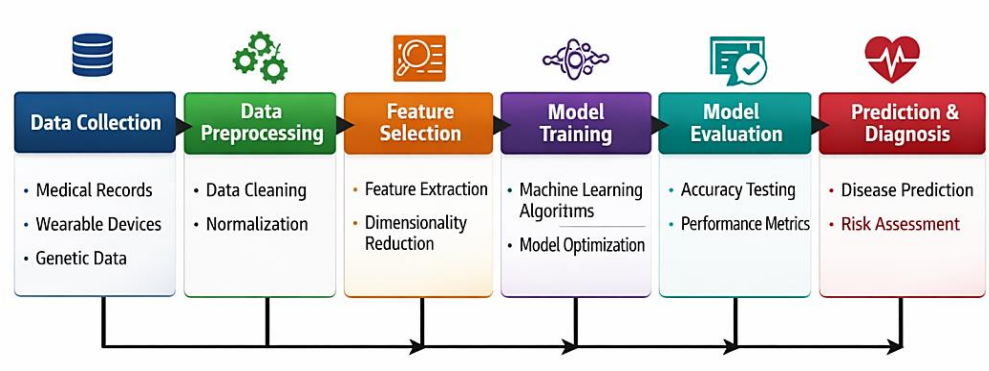

Fig 1: Data Flowchart Diagram for Chronic Diseases in Machine Learning

A brief discussion of each stage in the data flow diagram in Figure 1 is provided below.:

3.1. Data Collection and Visualization

The information needed for this research was drawn from the CKD dataset on this ML repository at UCI. Among the 400 patient records included in the record set, each one is tagged with 24 clinical features that relate to CKD, as well as whether the patient has the disease ("CKD") or does not have it ("notckd"). While there are 400 records in total, the label "CKD" affected 250 (62.5%) of them and "notckd" affected only 150 (37.5%). There are both numbers and categories in the dataset and sometimes values are missing, making it one of the main test datasets in medical ML. There are some visuals in this text:

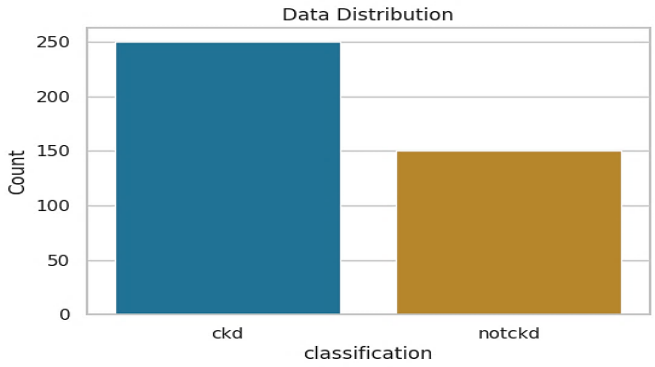

Fig 2: Class Distribution of CKD Dataset

Figure 2 clearly shows hence there are significantly more samples for CKD than for the not-chronic-kidney-disease (notckd) category. In all, 250 recent measures were classified as CKD, whereas 150 were classified as not CKD, showing that more people had CKD. Such class imbalance can negatively affect how machines learn and can decrease the reliability of the results, especially for sensitivity and specificity. Consequently, using strategies like resampling, making synthetic data or setting up class-weighted models is crucial to deal with prejudice and enhance the model's capacity to extrapolate in healthcare datasets that are unbalanced.

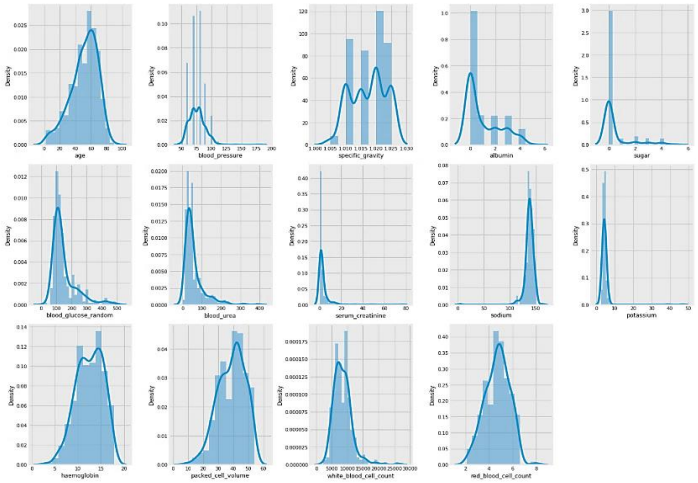

Fig 3: Distribution of Numerical Features in the CKD Dataset

Figure 3 presents the distribution plots of numerical characteristics from the CKD dataset, demonstrating that the majority of attributes exhibit significant levels of skewness and extensive variance. The values of blood pressure, serum creatinine and specific gravity are quite different from normal, since they have heavy tails and sharp peaks that might indicate unusual observations. In addition, blood measures such as haemoglobin, the lack of consistency in the clinical results is shown by the uneven patterns of packed cell volume and red blood cell count.

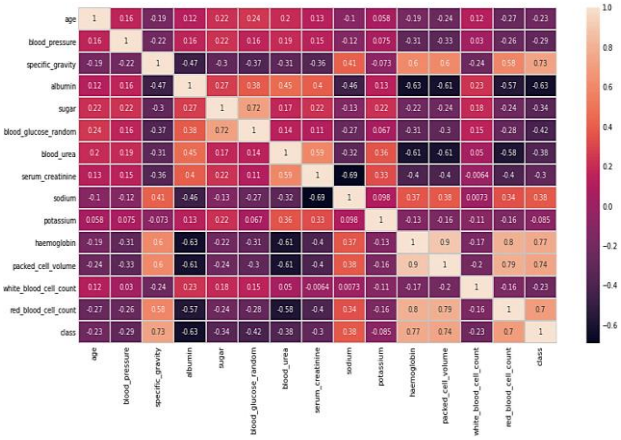

Fig 4: Correlation Heatmap of CKD Dataset

Figure 4 shows a heatmap of the association using Pearson's coefficient for the numerical characteristics in the CKD dataset. Positively, haemoglobin and serum creatinine show a strong negative connection with haemoglobin, although packed cell volume and red blood cell count display a solid positive correlation. There is a strong relationship between serum creatinine and haemoglobin levels and the class label, showing that these are important features for detecting CKD.

3.2. Data Preprocessing

Data preparation involves getting raw data ready to be analyzed and used to create models. To access the CKD dataset housed at the UCI Repository, this study carried out a sequence of pre-processing steps for handling missing data, organizing categorical variables, normalizing the values and finding outliers. The procedures are described one by one after this:

- Handling Missing Data: Missing data was handled by estimating the average for all numbers and the most common answer for all categories, which allowed us to fill in such values without changing the general structure of the data.
- The Interquartile Range (IQR): approach was used to eliminate outliers, since doing so improves the model's performance, cuts down on training time and keeps the data from being distorted.

3.3. Categorical Data Encoding

The purpose of Data Encoding is to turn types of data that are not numbers into numbers so that ML can work with them. Most ML algorithms require numeric input; therefore, categorical values must be converted into numerical form. Binary encoding is commonly used, where categories such as “no” and “yes” are represented as “0” and “1,” respectively.

3.4. Feature Importance

The key characteristics for CKD prediction were identified using RFE. It evaluates the model's performance and ranks the features by iteratively deleting the least significant ones. The most important characteristics were determined to be albumin, serum creatinine, blood pressure, and GFR. As a result of this choice, model ACC and dimensionality were both enhanced, which in turn helped to avoid overfitting.

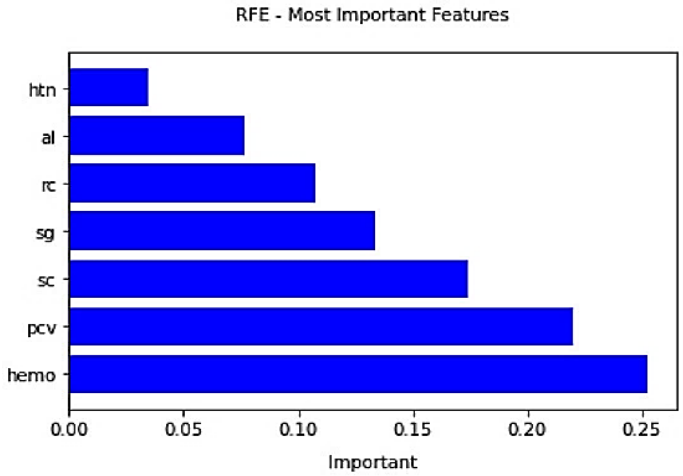

Fig 5: Important Features Selected By RFE

Figure 5 shows the top characteristics chosen by RFE for CKD prediction. Although albumin (al) and hypertension (hen) do not substantially impact the prediction process, hemoglobin (hem), pcv, and SC have the greatest significance ratings, suggesting a major effect on model performance.

3.5. Data Normalization with Standard Scaler

Data normalization scales the values of features to create a consistent distribution across all numerical features. In order to boost the ACC of ML models, this work applies data normalization. It changes values to fit in the range between −1 and +1. The average value of the converted set of data is 0, and its standard deviation is 1. The Equation for standardizing data (1) is shown below:

$$\omega = \frac{(X-\bar{x})}{\sigma} \qquad \square\square\square$$

where, w = Standardized score x = Observed value $\underline{x}$ = Mean σ = Standard deviation.

3.6. Data Splitting

Much of the data is utilized for training purposes, while half is put to use for testing. There are two sets of data used to train and evaluate the model: 75% and 25%, respectively.

3.7. Performance of Extra Tress Classifiers (ETC) Model

The ETC makes several random decision trees and then uses their predictions together to improve how well classifications are made [16]. While traditional decision tree ensembles are more rigid, ETC stands out by choosing random cutting points for all features and not applying bootstrapping on the data. Applying this approach makes the models more diverse, lowers how much the model changes and slows the chance of overfitting [17]. The classification outcome for a sample is decided X in the classification outcome for a sample is decided in Equation (2):

$$\hat{y} = mode\left(\{h_t(X)\}_{t\ =\ 1}^{T}\right)$$

Where: $\hat{y}$ is the predicted class label, T is the total number of trees, $h_t\ \ (X)$ The t-th tree gives a predicted class and mode selects the class that is most commonly selected by all the trees.

Each tree is grown by sampling some features and choosing random thresholds, instead of the ones that work best. While processing a node, each input feature f gets a random θ threshold drawn from its values and the best among those splits is picked. They define Equation (3) as:

$$\theta_f \sim u(\min(x_f), \max(x_f))$$

Where: θ_f is a randomly chosen threshold for feature θ_f, u denotes a uniform distribution over the observed range of f. This helps different trees in the forest to learn different features that increases the ACC and reliability of the classifier.

3.8. Performance Matrices

A collection of metrics used to evaluate ML model efficacy is called a Performance Matrix. TN, FN, FP and TP make up its components and they are used to form the confusion matrix. The matrix helps organize predictions, so it shows the ACC of predictions for every class.

The metrics like as REC, ACC, PRE, F1, and AUC-ROC to evaluate the ETC model's efficacy. The most important points are:

- True Positive (TP): Quantity of positive samples accurately predicted as positive by the classifier.
- True Negative (TN): The fraction of false negatives that the classifier properly labelled as false negatives.
- False Positive (FP): The amount of times a result is negative even if the classifier values it as positive.
- False Negative (FN): A measurement of how many positive samples the classifier believes are negative.

3.8.1. Accuracy

The term refers to the number of right guesses while considering all forecasts. ACC is the skill of making the right guess about what unfold. It in an Equation (4):

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

3.8.2. Precision

This helps in evaluating the model's performance when the cost of obtaining a FP is substantial. The value is found with the mathematical model known as Equation. (5):

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad (5)$$

3.8.3. Recall

Recall agrees with an observation only if the model predicts it positive, and its value is calculated by dividing the sum of all forecasts by the total number of observations in the class. Equation (6)

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

3.8.4. F1-Score

The F-measure joins PRE and REC, averaging them both with weighted values. Occasionally, the process leads to getting results that are not accurate. F-measure is used to refer to the Equation (7):

$$\text{F1} = \frac{2*(\text{precision}*\text{recall})}{\text{precision}+\text{recall}}$$

3.8.5. Auc-Roc

The AUC-ROC is critical for assessing the efficacy of the developed classification models. A ROC plot illustrates the association between specificity and REC as well as TP and FP. Besides, AUC shows how distinct the classes are separated by the classifier. From 0 to 1 is the range for AUC. This is why, if the AUC is high, the model recognizes minority and majority classes more effectively. See the Equations (8) and (9) in the picture.

$$FPR = \frac{FP}{TN+FP}$$
$$TPR = \frac{TP}{TP+FN}$$

They are used to compare the outcomes for the CKD dataset to assess the model's performance.

4. Results and Discussion

In this study, the DNN model is used to analyze its ability to predict CKD. The trial runs were conducted on a Windows 10 PC with 16 GB of RAM, an Intel i7 CPU (3.60 GHz, four-core), and Python 3, together with TensorFlow and Scikit-Learn. Table II presents the performance results for the DNN model, which achieves 99.9% ACC, guaranteeing reliable overall classification. Because the PRE is 99.9%, one may expect some inaccuracies, with positive results given to people who do not truly have CKD. An F1-score of 99.9% proves that both REC and ACC are adequately addressed by the model. The outcomes that the proposed model generates are explained in the section below:

Table 2: Results of ETC Model Performance on CKD Dataset for Big Data Health

Models	ETC Model
Accuracy	99.5
Precision	99.9
Recall	99.6
F1-score	99.4

Table III shows how the Extremely Randomized Trees Classifier (ETC) performs with the CKD data when looking at large health data. Compared to other models, ETC was superior in predicting, getting an F1 of 99.4%, a REC of 99.6%, a PRE of 99.9%, and an ACC of 99.5%. The model's reliability and usefulness in detecting CKD early on and in supporting healthcare analytics that include massive volumes of medical data are shown by these outcomes.

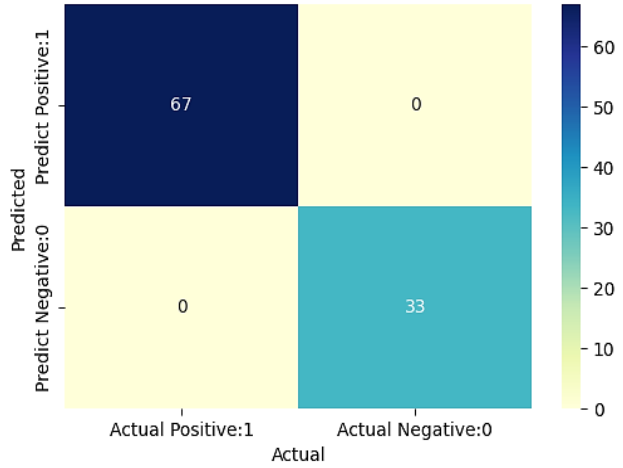

Fig 6: Performance of Confusion Matrices on ETC Model

The ETC model on the CKD dataset produces properly categorized results, as seen in the confusion matrix (Figure. 6). Every single positive and negative case was identified by the model, leading to zero cases of incorrect positives or negative results. The equal distribution on the confusion matrix honorably shows perfect ACC, REC and PRE, proving that the model is highly accurate for use with health big data.

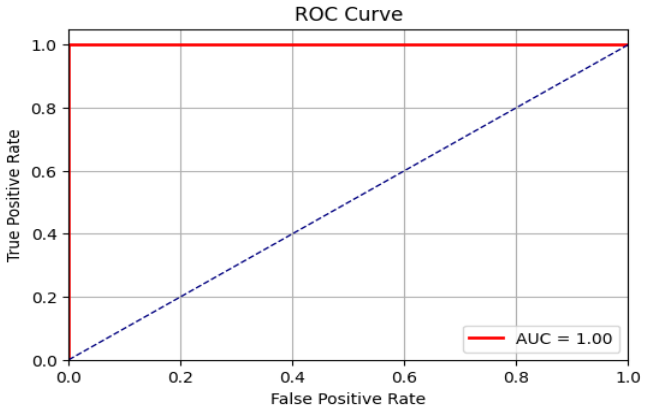

Fig 7: Performance on ROC Curve of ETC Model

Figure. 7 shows the ROC curve for the ETC model on the CKD dataset which demonstrates its outstanding results in classifying the disease. The curve rises quickly to have a TPR of 1.0 and an almost zero FPR, meaning classes are perfectly separated. The model performs at the top level and has no issue with choosing one over the other since its AUC is 1.00. The outcome proves that the ETC model works reliably for finding issues in healthcare data quickly and with high ACC.

4.1. Comparative Analysis

This part of the document compares the effectiveness of the proposed Extra Tree Classifiers (ETC) to several other ML models like SVM [18], LR [13]. Every model was taught and tested using the same dataset in the same way so that it could be fairly tested. As shown in Table III, the ETC model is more effective than the others in this case.

Table 3: ML and DL Models Comparison for Chronic Diseases Using Big Data Health

Models	SVM[18]	LR[19]	ETC
Accuracy	97.5	0.82	99.5
Precision	99.5	0.76	99.9
Recall	96.4	0.82	99.6
F1-score	97.9	0.79	99.4

Table III shows A big health data comparative analysis between ML and DL indicated that the Extra Trees Classifier (ETC) had the best performance among all models based on predetermined evaluation metrics. It achieved the highest ACC (99.5%), PRE (99.9%), REC (99.6%), and F1 (99.4%), which demonstrates its exceptional capacity to identify correctly the cases of disease with almost no mistakes. The SVM model is another one that has a strong performance as it provides 97.5% ACC and a balanced PRE, REC, and F1 making it a trustworthy alternative. Conversely, LR has a dramatically lower performance with an ACC of 0.82, PRE of 0.76, REC of 0.82, and an F1 of 0.79, which points out that linear models might not cater effectively to the challenges posed by the complexity and high-dimensional nature of big health datasets. To sum up, ensemble methods such as ETC provide more reliable predictive power for chronic disease classification.

The study underscores that predicting long-term illnesses using large healthcare datasets with modern ensemble learning models is more efficient than with conventional ML methods. Specifically, the Extra Trees Classifier has outperformed all other classifiers regarding different evaluation metrics, being more accurate than both SVM and LR in terms of ACC, PRE, REC, and F1. This remarkable performance is a reflection of the global generalization ability, toughness, and the high ACC of the ensemble-based methods even when handling large and complicated medical datasets. The results put forward that such ensemble techniques are superior in the capturing of minute health data patterns, thus increasing the trustworthiness and certainty of chronic disease diagnosis.

5. Conclusion and Future Direction

The chronicle disease application mainly involves applying both classification and prediction types of ML algorithms. Both DT, RF, and SVM are used to classify the data and determine the ACC of each algorithm using each disease's related data. For this study, rely on the datasets Heart diseases, Diabetes Mellitus dataset, and Liver dataset This study shows that an ML framework on big data helps detect CKD at an early stage. Using improved methods for preprocessing, choosing important features, and the Extra Tree Classifier, the model reached an ACC of 99.5%, demonstrating how useful it could be for clinical use. This finding agrees that using data in healthcare benefits the ACC of diagnosis and quick response time. For further improvements, this framework can be upgraded by using CNNs and LSTMs, which are advanced DL models, to help in predicting better and handling complicated data. Using IoT, patients can be closely monitored in real time for better health monitoring. Besides, having access to e-health records over time and developing models that predict the chance of several diseases at the same time helps the framework support more conditions and encourages better care decisions.

References

[1] A. Balasubramanian, "Intelligent Health Monitoring: Leveraging Machine Learning and Wearables for Chronic Disease Management and Prevention," *Int. J. Innov. Res. Eng. Multidiscip. Phys. Sci.*, vol. 7, no. 6, pp. 1–13, 2019, doi: 10.5281/zenodo.14535443.

[2] L. Segall, I. Nistor, and A. Covic, "Heart failure in patients with chronic kidney disease: A systematic integrative review," *Biomed Res. Int.*, vol. 2014, 2014, doi: 10.1155/2014/937398.

[3] R. Reynolds *et al.*, "A systematic review of chronic disease management interventions in primary care," *BMC Fam. Pract.*, vol. 19, no. 1, p. 11, Dec. 2018, doi: 10.1186/s12875-017-0692-3.

[4] W. Raghupathi and V. Raghupathi, "An Empirical Study of Chronic Diseases in the United States: A Visual Analytics Approach to Public Health," *Int. J. Environ. Res. Public Health*, vol. 15, no. 3, p. 431, Mar. 2018, doi: 10.3390/ijerph15030431.

[5] N. Bhardwaj, B. Wodajo, A. Spano, S. Neal, and A. Coustasse, "The Impact of Big Data on Chronic Disease Management," *Health Care Manag. (Frederick).*, vol. 37, no. 1, pp. 90–98, Jan. 2018, doi: 10.1097/HCM.0000000000000194.

[6] C. S. Kruse, R. Goswamy, Y. Raval, and S. Marawi, "Challenges and Opportunities of Big Data in Health Care: A Systematic Review," *JMIR Med. Informatics*, vol. 4, no. 4, p. e38, Nov. 2016, doi: 10.2196/medinform.5359.

[7] S. Achouche, U. B. Yalamanchi, and N. Raveendran, "Method, apparatus, and computer-readable medium for performing a data exchange on a data exchange framework," 2019

[8] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Brief. Bioinform.*, vol. 19, no. 6, pp. 1236–1246, Nov. 2018, doi: 10.1093/bib/bbx044.

[9] M. Elhoseny, K. Shankar, and J. Uthayakumar, "Intelligent Diagnostic Prediction and Classification System for Chronic Kidney Disease," *Sci. Rep.*, vol. 9, no. 1, p. 9583, Jul. 2019, doi: 10.1038/s41598-019-46074-2.

[10] M. Almasoud and T. E, "Detection of Chronic Kidney Disease using Machine Learning Algorithms with Least Number of Predictors," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 8, 2013, doi: 10.14569/IJACSA.2019.0100813.

[11] Y. Amirgaliyev, S. Shamiluulu, and A. Serek, "Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods," in *2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT)*, 2018, pp. 1–4. doi: 10.1109/ICAICT.2018.8747140.

[12] V. S, S. S, V. H, and S. S, "Disease Prediction Using Machine Learning Over Big Data," *Comput. Sci. Eng. An Int. J.*, vol. 8, no. 1, pp. 01–08, Feb. 2018, doi: 10.5121/cseij.2018.8101.

[13] S. B. Akben, "Early Stage Chronic Kidney Disease Diagnosis by Applying Data Mining Methods to Urinalysis, Blood Analysis and Disease History," *IRBM*, vol. 39, no. 5, pp. 353–358, Nov. 2018, doi: 10.1016/j.irbm.2018.09.004.

[14] S. Tekale, P. Shingavi, and S. Wandhekar, "Prediction of Chronic Kidney Disease Using Machine Learning Algorithm," *IJARCCE*, vol. 7, pp. 92–96, 2018, doi: 10.17148/IJARCCE.2018.71021.
[15] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017, doi: 10.1109/ACCESS.2017.2694446.
[16] A. V. Hazarika and Anju, "Extreme Gradient Boosting using Squared Logistics Loss function," *Int. J. Sci. Dev. Res.*, vol. 2, no. 8, pp. 54–61, 2017.
[17] R. Tarafdar and Y. Han, "Finding Majority for Integer Elements," *J. Comput. Sci. Coll.*, vol. 33, no. 5, pp. 187–191, 2018.
[18] M. Almasoud and T. E. Ward, "Detection of Chronic Kidney Disease using Machine Learning Algorithms with Least Number of Predictors," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 8, pp. 89–96, 2013, doi: 10.14569/IJACSA.2019.0100813.
[19] J. Xiao *et al.*, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," *J. Transl. Med.*, vol. 17, no. 1, pp. 1–13, 2019, doi: 10.1186/s12967-019-1860-0.
[20] Mamidala, J. V., Enokkaren, S. J., Attipalli, A., Bitkuri, V., Kendyala, R., & Kurma, J. (2023). Machine Learning Models Powered by Big Data for Health Insurance Expense Forecasting. *International Research Journal of Economics and Management Studies IRJEMS*, *2*(1).
[21] Nadella, V. M. (2023). Zero Trust Architecture for Telecom Operations. *International Journal of Emerging Research in Engineering and Technology*, *4*(3), 115-129.
[22] Bitkuri, V., Kendyala, R., Kurma, J., Enokkaren, S. J., & Mamidala, J. V. (2023). Forecasting Stock Price Movements With Deep Learning Models for time Series Data Analysis. *Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-531. DOI: doi. org/10.47363/JAICC/2023 (2)*, *489*, 2-9.
[23] Nadella, V. M. (2023). Anomaly Detection and Fault Prediction using ML in Telecom Operations. *International Journal of Emerging Trends in Computer Science and Information Technology*, *4*(3), 134-143.
[24] Kosaraju, P., & Nadella, V. M. (2022). Security and Privacy in IoT Ecosystems. *Universal Library of Engineering Technology*, (Issue).
[25] Singh, A. A. S. S., Mania, V., Kothamaram, R. R., Rajendran, D., Namburi, V. D. N., & Tamilmani, V. (2023). Exploration of Java-Based Big Data Frameworks: Architecture, Challenges, and Opportunities. Journal of Artificial Intelligence & Cloud Computing, 2(4), 1-8.
[26] Routhu, K. K. (2023). AI-driven succession planning in Oracle HCM Cloud: Building resilient leadership pipelines through predictive analytics. *International Journal of Science, Engineering and Technology*, *11*(5).
[27] Tamilmani, V., Namburi, V. D., Singh Singh, A. A., Maniar, V., Kothamaram, R. R., & Rajendran, D. (2023). Real-Time Identification of Phishing Websites Using Advanced Machine Learning Methods. *Available at SSRN 5837142*.
[28] Routhu, K. K. (2023). AI-driven succession planning in Oracle HCM Cloud: Building resilient leadership pipelines through predictive analytics. International Journal of Science, Engineering and Technology, 11(5). https://doi.org/10.5281/zenodo.17292018
[29] From Fragmentation to Focus: The Benefits of Centralizing Procurement. (2023). International Journal of Research and Applied Innovations, 6(6), 9820-9833. https://doi.org/10.15662/
[30] Routhu, K. K. (2023). Embedding fairness into the digital enterprise, data driven DEI strategies with Oracle HCM Analytics. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *9*(8), 266-274.
[31] Routhu, K. K. (2023). AI-driven skills forecasting in Oracle HCM Cloud: From static competencies to predictive workforce design. *International Journal of Science, Engineering and Technology*, *11*(1).
[32] Padur, S. K. R. (2023). AI-Augmented Enterprise ERP Modernization: Zero-Downtime Strategies for Oracle E-Business Suite R12. 2 and Beyond. *Available at SSRN 5605510*.
[33] Routhu, K. K. (2022). From Case Management to Conversational HR: Redefining Help Desks with Oracle's AI and NLP Framework. *International Journal of Science, Engineering and Technology*, *10*(6).
[34] Vattikonda, N., Gupta, A. K., Polu, A. R., Narra, B., Buddula, D. V. K. R., & Patchipulusu, H. H. S. (2022). Blockchain Technology in Supply Chain and Logistics: A Comprehensive Review of Applications, Challenges, and Innovations. *International Journal of Emerging Trends in Computer Science and Information Technology*, *3*(3), 72-80.
[35] Attipalli, A., BITKURI, V., Mamidala, J. V., Kendyala, R., & KURMA, J. (2022). Empowering Cloud Security with Artificial Intelligence: Detecting Threats Using Advanced Machine learning Technologies. *Available at SSRN 5741263*.
[36] Padur, S. K. R. (2022). Intelligent resource management: AI methods for predictive workload forecasting in cloud data centers. *J. Artif. Intell. Mach. Learn. & Data Sci*, *1*(1), 2936-2941.
[37] Nadella, V. M. (2022). Digital Twins for Predictive Network Management and System Simulation. *International Journal of AI, BigData, Computational and Management Studies*, *3*(3), 100-111.
[38] Routhu, K. K. (2022). From RFID to Geofencing: IoT-Enabled Smart Time Tracking in Oracle HCM Cloud. *International Journal of Science, Engineering and Technology*, *10*(4).
[39] Nadella, V. (2019). Extracting road traffic data through video analysis using automatic camera calibration and deep neural networks.

[40] Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., & Vangala, S. R. (2022). Data Security in Cloud Computing: Encryption, Zero Trust, and Homomorphic Encryption. *International Journal of Emerging Trends in Computer Science and Information Technology*, *3*(4), 31-41.

[41] Padur, S. K. R. (2022). AI augmented platform engineering, transforming developer experience through intelligent automation and self optimizing internal platforms. *International Journal of Science, Engineering and Technology*, *10*(5), 10-5281.

[42] Kosaraju, P. , & Nadella, V. M. (2021). Quality of Experience (QoE) and Network Performance Modelling for Multimedia Traffic. Journal of Artificial Intelligence and Big Data, 1(1), 1-13. https://doi.org/10.31586/jaibd.2021.1358.