Volume 1 No 6, 2025
Journal of Global Research in Multidisciplinary Studies (JGRMS)
Research Paper
Available online at https://saanvipublications.com/journals/index.php/jgrms/index

Efficient Machine Learning Tree-Based Models for Recognition of Chronic Diseases Using Big Data Health Records

Mrs. Neha Upadhyay,
Assistant Professor, Department of Computer Applications
IIS University,Bhopal(M.P.)
Email: neha.upadhyay887@gmail.com

***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 of 99.5%, the model is very effective in diagnosing CKD. When evaluation measures include precision, recall, F1-score, 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.**

I. INTRODUCTION

Worldwide healthcare systems are facing severe challenges because of people living with chronic diseases. CKD, the metabolic syndrome, high blood pressure, and heart failure are now some one of the primary reasons for chronic problems and death [1]. It has been revealed by WHO that chronic diseases make up about 71% of all deaths in a 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 [3].

Managing and analyzing massive amounts of data kept in EHRs is one example. Regarding big data's present use in healthcare, lab findings, personal information, medication records and past diagnoses [4]. These records support health professionals in seeing how diseases develop in a patient in the long run[5]. 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 [6]. The process of predictive analytics helps detect high-risk individuals so that other health measures can be taken in advance and better results are reached [7].

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 [8]. ML software can analyze big data records to discover valuable patterns and tips about future risks [9], making it possible for intervention to happen early, diagnosis to be accurate and treatment to match the patient. As a result, there are fewer cases of illness, death and people in hospitals [10]. Because healthcare is now digital, there are now enormous databases containing EHRs, diagnostics, patient characteristics, medical backgrounds and lifestyle information. 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 [11]. It makes medical decisions easier, result in better diagnoses and help doctors design correct treatment options for each person [12].

ML helps a lot in examining complicated medical records [13]. Access to big collections of data allows it to understand disease causes better, identify high-risk patients and detect diseases early [5][14]. 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 [15]. These approaches are now part of many medical uses, for instance in the diagnosis, prognosis, and prediction of illnesses such inflammatory bowel disease, multiple sclerosis, autoimmune kidney disease, and autoimmune rheumatic disorders [16]. ML also helps choose treatment plans, divide patients into groups, make medicines, recycle medicines and explain drug targets [17].

A. Motivation and Contribution of 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 including filling in blanks, eliminating extreme cases, and encoding categories for improved data.

- Select the most relevant characteristics for improved efficiency and interpretation using Applied Recursive Feature Elimination.
- The ETC was implemented to gain high accuracy and lower computations because it splits data randomly.
- I evaluated how the model was working by calculating F1-score, AUC/ROC, recall, accuracy, and precision.

B. 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. DNNs' ability to discern complex data patterns and correlations allows the proposed solution to rectify issues brought about by class imbalance and enhance precision. 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.

C. 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.

II. Literature Review

This part discusses the recent advances in recognizing and classifying CKD, with attention given to the benefits of ML. Advanced methods are examined to see how they can improve the way CKD is identified. These are some of the most important review works.:

M and N (2025) examine if ML can be used, such as the Light Gradient Boosting Machine (LightGBM) algorithm, to build models for predicting the risk of CKD from clinical and demographic factors. As it handles both big datasets and a significant number of features well, LightGBM is appropriate for this task. Using details such as age, blood pressure and results from lab tests, the model can foresee the risk of CKD and help enforce early screening and specific treatment. Running lightGBM against KNN and Decision Trees proves that lightGBM achieves better results. Reaching 96% accuracy, 93% recall and an F1-score of 89%, the model shows it can be relied on for predicting CKD [18].

Yogalakshmi and M. (2025) uses a method that mixes CNNs for identifying features and SVMs for classification to help detect CKD more precisely. After being trained and verified on a CKD dataset, the model had an accuracy of 96.18%. The SVM creates distinct borders to identify the distinctions between CKD and non-CKD instances, while the CNN extracts valuable information. To achieve good results and ensure data quality, two approaches called normalization and augmentation were used in the study. AI progress related to CKD diagnosis is well covered in the literature and highlights that hybrid architectures and multi-task learning help improve how well the model performs [19].

Lee et al. (2024) consider studying walking patterns to find CKD by analyzing movement signs gathered by IMU sensors. For this research, deep learning is used on gait measurement data taken from 276 individuals who have CKD and 217 healthy controls supplied by Hallym University Chuncheon Sacred Heart Hospital. They suggested a method that involves using CNN and BiLSTM together to find CKD by analyzing normalized gait data. On segmenting the data, method A reached an 84.98% accuracy rate, while the voting approach resulted in 79.61% accuracy. This suggests that using information from someone's walk could detect CKD, which is a new idea for early diagnosis and better care of the disease [20].

Kumar S et al. (2024) recommends a 32-layer ResNet32 architecture for use in a Lightweight multi-attention convolutional neural Network. To generate a feature map with lower kernel filter settings, the attention module is included in the residual block of the network. This block captures the relevant dataset features. Applying the CKD dataset, they conduct experimental analysis of the suggested technique. The proposed model outperforms state-of-the-art approaches in a cross-fold validation performance analysis, with an accuracy of 98.89% [21].

Bhuria (2024) introduces a highly accurate and sophisticated ML approach for predicting CKD. The model's performance, as indicated by a confusion matrix, includes 45 occurrences accurately classified as positive, 72 instances accurately classified as negative, and 0 instances misclassified as positive. The outcome yields an accuracy rate of 97.5%. It guarantees complete accuracy and thorough coverage for non-CKD, making certain that none of the cases are put into the CKD group by mistake. Even so, this model only manages to detect 96.25% of the CKD cases, so it misses three of them [22].

Jeyalakshmi et al. (2024) Advanced neural networks used in the suggested DL method to examine various patient records, together with genetic, clinical and imaging records which helps in understanding disease patterns. It is evident that the recommended Deep Learning system, developed for identifying the biomarkers associated with CKD, is more effective compared to present methods, delivering higher sensitivity (94.50/0), specificity (92.2), accuracy (90.8%) and optimal ROC (0.94), thus enabling effective prediction of CKD development and making the model ideal for early identification and custom treatment [23].

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 I. Summary of Reviewed Works on Chronic Diseases Using Big data

Author(s)	Methodology	Datasets	Key Findings	Limitations & Future Work
M and N (2025)	Applied LightGBM algorithm for CKD risk prediction using clinical and demographic indicators; compared performance with KNN and Decision Trees.	Clinical dataset with features like age, blood pressure, and lab findings.	LightGBM achieved 96% accuracy, 93% recall, and 89% F1-score; it outperformed KNN and Decision Trees in handling high-dimensional data.	Limited scope of patient features; lacks real-time or longitudinal data integration.
Yogalakshmi and M. (2025)	Built a CNN-SVM hybrid DL model for feature extraction and classification; preprocessed data using normalization and augmentation methods.	CKD dataset; specifics not detailed.	Achieved 96.18% accuracy; CNN effectively extracted high-dimensional features, and SVM provided robust classification.	Dataset details and size not specified; model complexity may limit deployment.
Lee et al., (2024)	CNN + BiLSTM on normalized gait data	Gait data from 276 CKD patients and 217 healthy controls (Hallym Univ. Hospital)	Binary classification: 84.98% (segment), 79.61% (voting); gait useful for CKD detection	Small sample size; needs multimodal data and real-world testing
Kumar S et al., (2024)	Lightweight Multi-Attention CNN using ResNet32	CKD dataset with multi-stage classification	Achieved 98.89% accuracy; lightweight with attention-based residual blocks	Limited generalization; requires validation across multiple datasets.
Bhuria (2024)	ML model analyzed with confusion matrix	Custom dataset with CKD and non-CKD classes	Accuracy: 97.5%, perfect non-CKD precision, recall: 93.75%	Lower CKD recall; improvement needed in sensitivity
Jeyalakshmi et al., (2024)	DL on multimodal data (genetic, clinical, imaging)	Integrated patient dataset with various modalities	Accuracy: 90.8%; Sensitivity: 94.5%; Specificity: 92.2%; AUC: 0.94	Complex data integration; further validation on diverse populations

III. 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, it divides the dataset in half, with 75% serving as training data and 25% as test data. The Extra Tree Classifier (ETC) model is used in the classification portion because of its efficiency and ability to avoid overfitting. Various measures, for example, accuracy, precision, recall, F1-score, and AUC/ROC, are applied to judge the model's prediction accuracy.

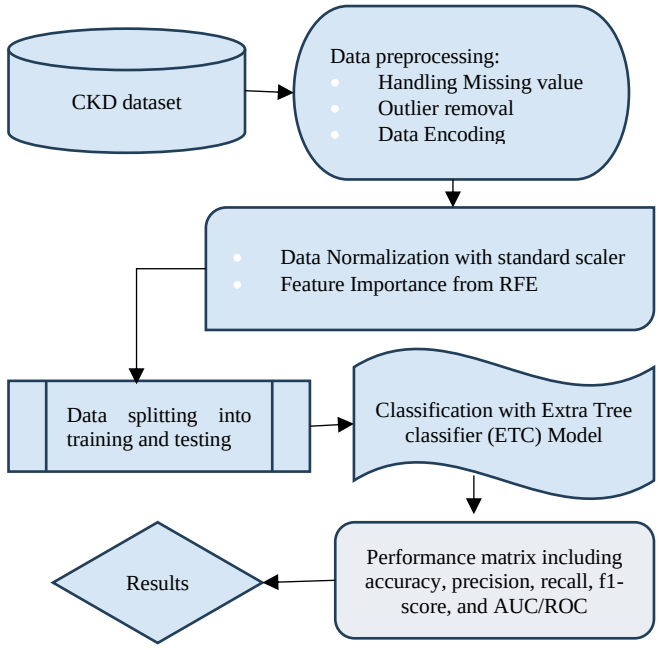

Fig. 1. Data Flowchart Diagram for Chronic Diseases in machine learning

A small description of each step in the data Figure 1 flow diagram is given below:

A. 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 ("not CKD"). While there are 400 records in total, the label "CKD" affected 250 (62.5%) of them and "not CKD" 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:

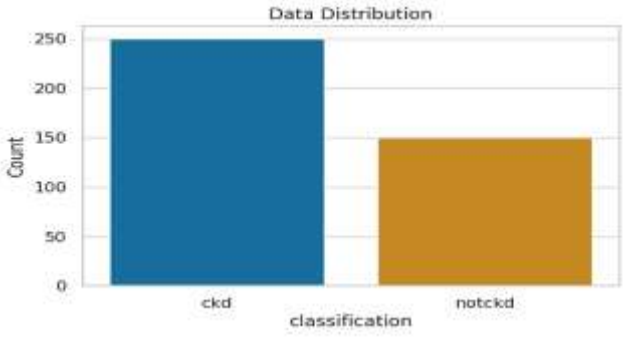

Fig. 2. Class Distribution of CKD Dataset

Figure 2 clearly shows hence there are significantly more samples for CKD than for the not-chronic-kidney-disease (not CKD) category. In all, 250 recent measures were classified as CKD, whereas 150 were classified as non-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.

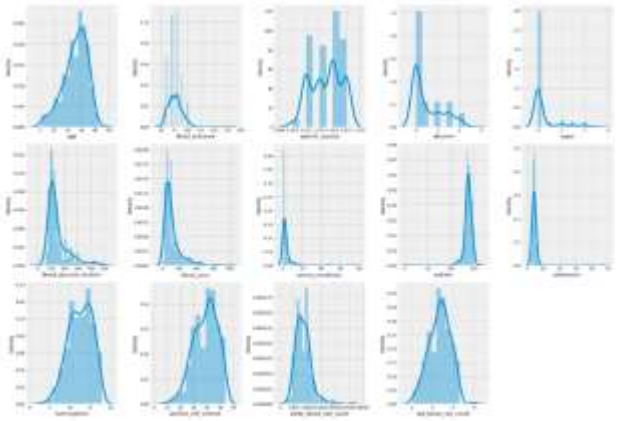

Fig. 3. Distribution of Numerical Features in the CKD Dataset

Figure 3 presents CKD dataset numerical characteristic distribution graphs show that most attributes are highly skewed and have a lot of volatility. 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 haemoglobin concentration and volume of packed cells.

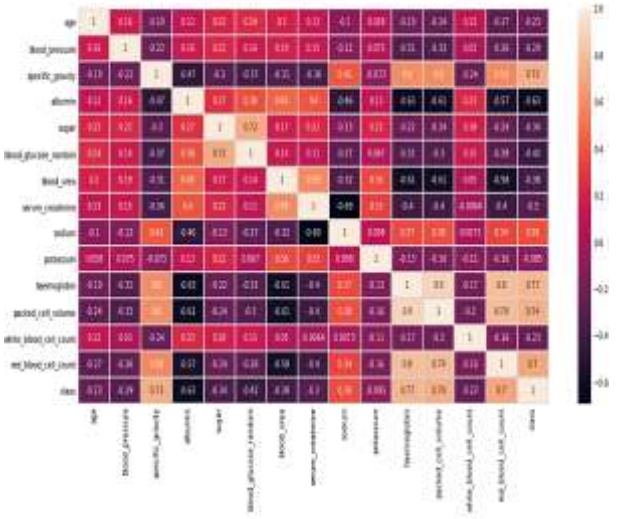

Fig. 4. Correlation Heatmap of CKD Dataset

Figure 4 shows a heatmap of the correlation using Pearson's coefficient for the numerical features in the CKD dataset. Positively, haemoglobin, red blood cell count and packed cell volume are strongly positively correlated, but haemoglobin and serum creatinine are strongly negatively correlated. Serum creatinine and haemoglobin levels are strongly correlated with the class label, showing that these are important features for detecting CKD.

B. 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.
- To remove outliers, the Interquartile Range method was used since it enhances model performance, reduces training time, and prevents data distortions.

C. 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.

D. 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 glomerular filtration rate (GFR). As a result of this choice, model accuracy and dimensionality were both enhanced, which in turn helped to avoid overfitting.

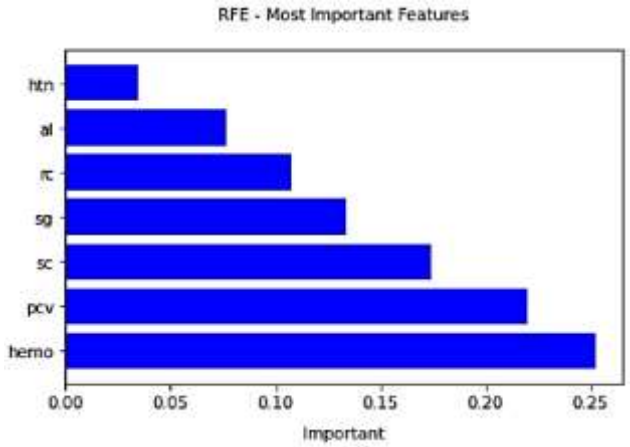

Fig. 5. Important features selected by RFE

Figure 5 shows the top characteristics chosen by RFE for CKD prediction. Although albumin (al) and hypertension (htn) do not substantially impact the prediction process, hemoglobin (hemo), packed cell volume, and serum creatinine (SC) have the greatest significance ratings, suggesting a major effect on model performance.

E. Data Normalization with Standard Scaler

Data normalization modifies the numbers in features to create a consistent distribution for all numerical factors. To boost the accuracy 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 (1) for standardizing data is shown below:

$$\omega = \frac{(X-\bar{x})}{\sigma} \quad (1)$$

where w is the standardized score, x is the observed value, x is the mean, and σ is the standard deviation."

F. Data Splitting

The bulk of the data goes into training, while the rest is set aside for testing. In order to train the model, it utilizes 75% of the data, and to test its performance, it uses 25%.

G. 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 [24]. 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 [25][26]. Applying this approach makes the models more diverse, lowers how much the model changes and slows the chance of overfitting [27][28]. The classification outcome for a sample is decided X in the classification outcome for sample is decided in Equation (2):

$$\hat{y} = mode\left(\{h_t(X)\}_{t=1}^{T}\right) \quad (2)$$

Where: with a total of T trees $\hat{y}$, represents the anticipated class label, 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 [29]. 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:

$$\theta_f \sim u(\min(x_f), \max(x_f)) \quad (3)$$

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 accuracy and reliability of the classifier.

H. 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 accuracy of predictions for every class.

Metrics including recall, accuracy, precision, F1-score, and AUC-ROC are used to assess the efficacy of the ETC model. The primary considerations should be:

- **True Positive (TP):** The percentage of positive samples that the classifier correctly identified as positive.
- **True Negative (TN):** The fraction of false negatives that the classifier properly labelled as false negatives.
- **False Positive (FP):** How many times there are negative even when the classifier thinks, they are positive.
- **False Negative (FN):** A measure of how many positive samples the classifier thinks are negative.

1) Accuracy

The term means how many correct guesses there are when considering all the predictions. Accuracy is the skill of making a right guess about what will unfold. It can show it in an Equation (4):

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4)$$

2) Precision

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

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

3) Recall

Recall agrees with an observation only if the model detects it will be positive, and the value of the variable is found by using a formula that takes into account both the class's observations and predictions Equation (6):

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (6)$$

4) F1-Score

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

$$\text{F1} = \frac{2*(\text{precision}*\text{recall})}{\text{precision+recall}} \quad (7)$$

5) AUC-ROC

The AUC-ROC is very important for measuring the effectiveness of produced classification models. A receiver operating characteristic (ROC) plot illustrates the association between specificity and recall as well as true positives and false positives [30]. 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 Equations (8) and (9) in the picture.

$$FPR = \frac{FP}{TN+FP} \quad (8)$$

$$TPR = \frac{TP}{TP+FN} \quad (9)$$

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

IV. Results And Discussion

The capacity of the DNN model to predict CKD is examined in this research. The tests were conducted using a Windows 10 PC with 16 GB RAM, an Intel i7 CPU (3.60 GHz, four-core), and Python 3. Other tools utilized were TensorFlow and Scikit-Learn. The DNN model's performance results are shown in Table II, and they indicate a reliability of 99.9 per cent for total classification. Because the precision is 99.9%, it may expect some inaccuracies by giving positive results to people who do not truly have CKD. A F1-score of 99.9% proves that both recall and accuracy are adequately addressed by the model. The outcomes that the proposed model generates are explained in the section below:

TABLE II. 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 II shows how the Extremely Randomized Trees Classifier (ETC) performs with the CKD data when looking at large health data. With a 99.4% F1-score, 99.6% recall, 99.9% precision, and 99.5% accuracy, ETC outperformed competing models in terms of prediction. The model's reliability and

usefulness in CKD early on and in supporting healthcare analytics that include massive volumes of medical data are shown by these outcomes.

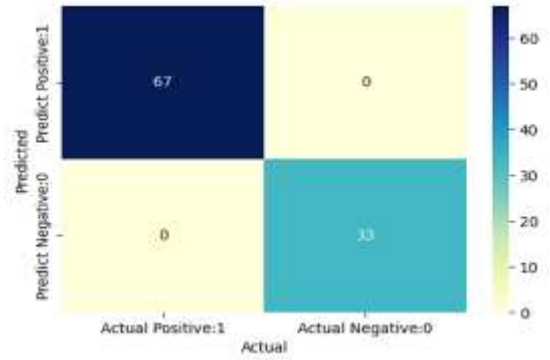

Fig. 6. Performance of Confusion matrices on ETC Model

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 accuracy, recall and precision, proving that the model is highly accurate for use with health big data.

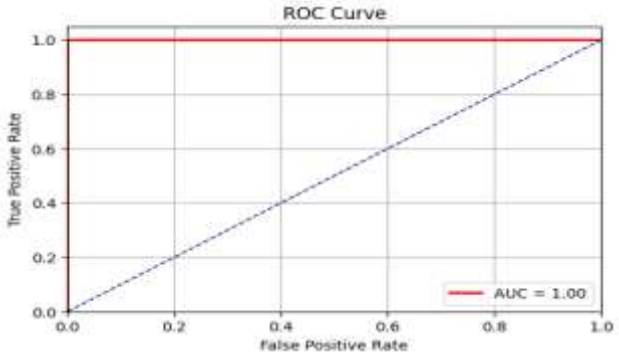

Fig. 7. Performance on ROC Curve of ETC Model

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 accuracy.

A. Comparative Analysis

This part of the document compares the effectiveness of the proposed ETC to several other ML models like DT [31], NB [32], Adaboost [25]. 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 III. ML AND DL MODELS' COMPARISON FOR CHRONIC DISEASES USING BIG DATA HEALTH

Models	**DT**[31]	**NB**[32]	**Adaboost**[25]	**ETC**
Accuracy	96	93	97.5	99.5
Precision	94.28	86.84	98	99.9
Recall	94.28	94.38	97	99.6
F1-score	94.28	90.41	97	99.4

Table III shows how different ML and DL models are used to predict chronic diseases with big data in healthcare. Some of the models employed in the study are DT, SVM, AdaBoost, ETC. The ETC model provides better results than others, having an outstanding accuracy of 99.9% and also reaching 99.9% with respect to recall, precision, and F1-score. However, the DT model has a 96% accuracy rate, balanced recall, and F1-score of 94.2%. Compared to its competitors, the SVM model performs better, with an F1-score of 97.3% and an accuracy of 96.6%. The model's accuracy rating is 97.5%, thanks to its 97% F1-score, 98% recall, and 98% precision. This proves that ETC and AdaBoost can successfully use large healthcare datasets to forecast the occurrence of chronic diseases.

The research points out that forecasting chronic diseases with big healthcare data is better achieved by ensemble learning models than with traditional ML algorithms. From all evaluation angles, models like ETC and AdaBoost present better results than DT and SVM due to their accuracy, toughness, and superior ability to generalize. This demonstrates how ensemble approaches may increase the confidence of chronic illness detection by revealing hidden patterns in large amounts of medical data.

V. 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 accuracy of each algorithm using each disease's related data. For this study, it relys 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 ETC, the model reached an accuracy of 99.5%, demonstrating how useful it could be for clinical use. This finding agrees that using data in healthcare benefits the accuracy of diagnosis and quick response time.

For further improvements, this framework can be upgraded by using CNNs and LSTMs, which are advanced deep learning 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.

[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] S. Pandya, "Integrating Smart IoT and AI-Enhanced Systems for Predictive Diagnostics Disease in Healthcare," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 10, no. 6, Dec. 2024, doi: 10.32628/CSEIT2410612406.

[4] R. P. Mahajan, "Transfer Learning for MRI image reconstruction: Enhancing model performance with pretrained networks," *Int. J. Sci. Res. Arch.*, vol. 15, no. 1, pp. 298–309, Apr. 2025, doi: 10.30574/ijsra.2025.15.1.0939.

[5] R. Alanazi, "Identification and Prediction of Chronic Diseases

Using Machine Learning Approach," *J. Healthc. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/2826127.

[6] K. Batko and A. Ślęzak, "The use of Big Data Analytics in healthcare," *J. Big Data*, 2022, doi: 10.1186/s40537-021-00553-4.

[7] V. Kolluri, "An Innovative Study Exploring Remote Patient Monitoring With AI: Chronic Disease Management Enhancements," *Int. J. Creat. Res. Thoughts*, vol. 10, no. 5, 2022.

[8] V. Kolluri, "Analytics for Patient Care: How AI is Used to Predict Patient Health Outcomes and Improve Care Plans," *TIJER - Int. Res. Journals (TIJER)*, vol. 8, no. 11, pp. 2349–9249.

[9] S. Pandya, "A Systematic Review of Blockchain Technology Use in Protecting and Maintaining Electronic Health Records," *Int. J. Res. Anal. Rev.*, vol. 8, no. 4, 2021.

[10] E. V. Emeihe, E. I. Nwankwo, M. D. Ajegbile, J. A. Olaboye, and C. C. Maha, "The impact of artificial intelligence on early diagnosis of chronic diseases in rural areas," *Comput. Sci. IT Res. J.*, vol. 5, no. 8, pp. 1828–1854, 2024, doi: 10.51594/csitrj.v5i8.1447.

[11] R. S. Raj and M. Kusuma, "A Comprehensive Analysis of Chronic Health Diseases using Big Data," *2023 Int. Conf. Evol. Algorithms Soft Comput. Tech. EASCT 2023*, pp. 1–5, 2023, doi: 10.1109/EASCT59475.2023.10392342.

[12] D. Rao and D. S. Sharma, "Secure and Ethical Innovations: Patenting Ai Models for Precision Medicine, Personalized Treatment, and Drug Discovery in Healthcare," *Int. J. Bus. Manag. Vis.*, vol. 6, no. 2, 2023.

[13] S. Murri, *From Raw to Refined: The Art and Science of Data Engineering*. Notion Press, 2025.

[14] R. Q. Majumder, "Machine Learning for Predictive Analytics: Trends and Future Directions," *Int. J. Innov. Sci. Res. Technol.*, vol. 10, no. 4, pp. 3557–3564, 2025.

[15] I. Preethi and K. Dharmarajan, "Diagnosis of chronic disease in a predictive model using machine learning algorithm," in *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, IEEE, Oct. 2020, pp. 191–196. doi: 10.1109/ICSTCEE49637.2020.9276957.

[16] S. Swaminathan *et al.*, "A machine learning approach to triaging patients with chronic obstructive pulmonary disease.," *PLoS One*, vol. 12, no. 11, Nov. 2017, doi: 10.1371/journal.pone.0188532.

[17] R. Dattangire, R. Vaidya, D. Biradar, and A. Joon, "Exploring the Tangible Impact of Artificial Intelligence and Machine Learning: Bridging the Gap between Hype and Reality," in *2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET)*, IEEE, Aug. 2024, pp. 1–6. doi: 10.1109/ACET61898.2024.10730334.

[18] K. M and S. N, "Early Chronic Kidney Disease Detection Using Machine Learning Techniques," in *2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS)*, IEEE, Mar. 2025, pp. 151–155. doi: 10.1109/ICMLAS64557.2025.10967970.

[19] S. Yogalakshmi and M. Anitha, "A Hybrid Deep Learning Approach for Chronic Kidney Disease Detection Using Convolutional Neural Networks and Support Vector Machines," in *2025 International Conference on Inventive Computation Technologies (ICICT)*, IEEE, Apr. 2025, pp. 620–625. doi: 10.1109/ICICT64420.2025.11004966.

[20] S. Lee *et al.*, "Towards early detection of chronic kidney disease based on gait patterns: IMU-based approach using neural networks," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024, pp. 1–6. doi: 10.1109/EMBC53108.2024.10781594.

[21] S. Kumar S, K. B, S. J, and M. S, "LMA-CNN: Lightweight Multi Attention Convolution Neural Network for Early Diagnosis of the Chronic Kidney Disease," in *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, IEEE, Jul. 2024, pp. 1498–1502. doi: 10.1109/ICSCSS60660.2024.10625428.

[22] R. Bhuria, "Advanced Machine Learning Techniques for Chronic Kidney Disease Prediction and Management," in *2024 International Conference on Artificial Intelligence and Emerging Technology (Global AI Summit)*, 2024, pp. 480–484. doi: 10.1109/GlobalAISummit62156.2024.10947932.

[23] G. Jeyalakshmi, F. Vincy Lloyd, K. Subbulakshmi, and G. Vinudevi, "Application of Deep Learning in Identifying Novel Biomarkers for Chronic Kidney Disease Progression," in *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2024, pp. 353–358. doi: 10.1109/ICACCS60874.2024.10717197.

[24] Anju and A. V. Hazarika, "Extreme Gradient Boosting using Squared Logistics Loss function," *Int. J. Sci. Dev. Res.*, vol. 2, no. 8, pp. 54–61, 2017.

[25] M. A. Islam, M. Z. H. Majumder, and M. A. Hussein, "Chronic kidney disease prediction based on machine learning algorithms," *J. Pathol. Inform.*, vol. 14, 2023, doi: 10.1016/j.jpi.2023.100189.

[26] N. Malali, "AI-Powered Data Preprocessing and Transformation Platform for Autonomous Data Cleaning, Advanced Fea," 202521035175, 2025

[27] A. Balasubramanian, "Improving Air Quality Prediction Using Gradient Boosting," *Int. J. Sci. Technol.*, vol. 13, no. 2, pp. 1–9, 2022.

[28] R. Tarafdar and Y. Han, "Finding Majority for Integer Elements," *J. Comput. Sci. Coll.*, vol. 33, no. 5, pp. 187–191, 2018.

[29] T. U. Roberts, A. Polleri, R. Kumar, R. J. Chacko, J. Stanesby, and K. Yordy, "Directed Trajectories Through Communication Decision Tree using Iterative Artificial Intelligence," 11321614, 2022

[30] M. A. Panza, M. Pota, and M. Esposito, "Anomaly Detection Methods for Industrial Applications: A Comparative Study," *Electron.*, vol. 12, no. 18, 2023, doi: 10.3390/electronics12183971.

[31] M. S. Arif, A. U. Rehman, and D. Asif, "Explainable Machine Learning Model for Chronic Kidney Disease Prediction," *Algorithms*, vol. 17, no. 10, pp. 1–17, 2024, doi: 10.3390/a17100443.

[32] E. M. Senan *et al.*, "Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques," *J. Healthc. Eng.*, pp. 1–10, Jun. 2021, doi: 10.1155/2021/1004767.

