

HSCDT: HYBRID SPECTRAL CLUSTERING DECISION TREE FOR PREDICTING CHRONIC KIDNEY DISEASE IN HEART DISEASE PATIENTS

CHANDRALEKHA E¹, T R SARAVANAN²

¹Research Scholar, Department of Computational Intelligence, SRM Institute of Science and Technology, Kattankulathur, India

²Associate Professor, Department of Computational Intelligence, SRM Institute of Science and Technology, Kattankulathur, India

E-mail: ¹ce7242@srmist.edu.in, ² saravant1@srmist.edu.in

ABSTRACT

Chronic Kidney Disease (CKD) is intertwined with cardiovascular disease (CVD), sharing common risk factors and underlying pathophysiologic processes. However, the early diagnosis of CKD in patients with cardiovascular diseases is crucial for timely intervention and improved patient outcomes. This research introduces a new Hybrid Spectral Clustering Decision Tree (HSCDT) model for predicting CKD in patients with cardiac disease. The method implements spectral clustering and applied decision tree classification for effective discovery of complex patterns in the data set. Various classifiers, including Gradient Boosting, AdaBoost, XGBoost, LightGBM, and Voting Classifier were selected for comparative analysis. The proposed model achieved a peak of 99% cross-validation accuracy and a test accuracy of 99%, outperforming previous classifiers. Important clinical predictors of chronic kidney disease (CKD), such as age, body mass index, smoking, history of coronary heart disease, blood pressure, cholesterol, serum creatinine, hypertension, and eGFR, were analyzed and included in the prediction model. The performance on the experimental data confirmed the model's ability to accurately distinguish the CKD patients from the rest of the population. The findings suggest a significant enhancement in classification effectiveness when spectral clustering and decision trees are consolidated and provide this classification as a possible mode of disease prediction of CKD in the patient suffering from heart disease. Further research can focus on increasing datasets, involving additional biomarkers, and improving models for increased clinical utility.

Keywords: *Chronic Kidney Disease (CKD), Cardiovascular Disease, Hybrid Spectral Clustering Decision Tree (HSCDT), Machine Learning*

1. INTRODUCTION

Chronic Kidney Disease (CKD) and Cardiovascular Diseases (CVD) remain two of the most formidable public health epidemics of the 21st century, posing a substantial burden on worldwide healthcare systems culminating in high morbidity and mortality [1]. CKD is characterized by a gradual and irreversible loss of renal function and affects approximately 10% of the world's population, often undiagnosed until very late stages of the disease [2]. Meanwhile, CVD (comprising heart failure, coronary artery disease, and stroke) is still the most common cause of morbidity globally, accounting for about 18 million deaths each year [3].

There is a powerful bidirectional relationship between CKD and CVD, such that the presence of one glow affects the course of the other [4]. Patients with CVD are at higher risk of CKD by virtue of shared risk factors, namely hypertension, diabetes and atherosclerosis and the nephrotoxicity of some cardiovascular therapies [5]. On the other hand, CKD is a strong risk factor for cardiovascular diseases, such as myocardial infarction, cerebrovascular event, and heart failure, through mechanisms including fluid overload, chronic inflammation, electrolyte disorder, and vascular calcification [6], [7]. Such a pathological interaction sets off a vicious cycle, thereby aggravating the importance to early diagnosis and integrated management in co-morbid patients.

The escalating prevalence of CKD and CVD, enhanced by aging of the population, inactivity, and the worldwide increase in diabetes and obesity, emphasizes the need to address this twin epidemic [8]. Especially, in patients with concomitant cardiac diseases, there is a risk of underdiagnosing early-stage CKD with traditional diagnostic modalities by common overlap of symptoms and misinterpretation as primary cardiovascular disease [9]. Being so, a diagnosis of CKD is often made after substantial renal damage has occurred, which undermines the efficacy of therapeutic strategies and aggravates prognosis [10]. Early diagnosis and prompt intervention are considered as important strategies in order to prevent disease further progression, lower cardiovascular risk, and enhance patient outcome [11]. This work highlights the need for new tools to identify individuals at high risk before symptoms appear clinically [12].

The indication for the current research was decided upon due to the unmet clinical question of early detection of CKD in subjects that have already been diagnosed CVD. To do this, articles that were published in peer-reviewed papers between 2020 and 2024 regarding ML for CKD and CVD diagnosis including those studies using term search like “CKD prediction,” “CVD comorbidity,” “hybrid models,” and “machine learning in clinical” was taken. Screening and shortlisting of the articles was done on relevance, quality of findings and their focus aimed at computational methods in comorbid diagnostics.

ML has proved to be a revolutionary option in healthcare that can discover underlying data patterns in massive and intricate datasets [13], [14]. While several models predicting each disease (CKD or CVD) independently have been published, little is known about the joint risk brought by both diseases. The vast majority of the existing work adopts models like decision trees, random forests, support vector machines and boosting algorithms. Nonetheless, these models frequently fail to make accurate predictions about the presence of comorbidity or to yield interpretable results for clinicians.

To bridge this gap, the following research questions have been formulated:

1. Can a machine learning model accurately predict chronic kidney disease in individuals already diagnosed with cardiovascular disease using clinical, demographic, and laboratory features?

2. Can a hybrid model improve predictive performance and interpretability compared to traditional classifiers?

Contribution:

A Hybrid Spectral Clustering Decision Tree (HSCDT) for better predicting CKD among patients diagnosed with cardiovascular events has been presented. Key to the novelty of the approach is a combination of spectral clustering and decision tree classification which enables a good pattern discovery as well as high interpretability. Compared to previous research which considered only CKD or CVD, the models in the proposed method present a unified framework for the two conditions jointly.

The primary goal of the research is to introduce and assess a new Hybrid Spectral Clustering Decision Tree (HSCDT) model that is designed to enhance the diagnostic capabilities of chronic diseases. The model is specifically designed to address CKD in patients with pre-existing heart failure. The primary contributions of this work are as follows:

- **Formulation of the HSCDT Model:** A Hybrid Spectral Clustering Decision Tree (HSCDT) model is proposed, integrating the advantages of spectral clustering and decision tree methods to improve the precision and dependability of chronic disease diagnostics. This hybrid methodology aims to overcome the shortcomings of conventional classifiers in managing intricate clinical datasets.

- **Enhanced Predictive Accuracy:** The research paper presents empirical evidence indicating that the HSCDT classifier attains more accuracy than conventional ML classifiers. This is evidenced by comprehensive testing and validation utilizing clinical datasets of CKD and heart failure.

- **Thorough Performance Assessment:** The HSCDT model's performance is assessed using multiple metrics, including Recall, Precision, and the F1 score. These criteria are selected to provide a comprehensive evaluation of the model's capacity to accurately identify genuine positives, reduce false positives, and maintain a balance between precision and recall.

- **Identification of Key Risk Factors:** The HSCDT model proficiently detects patients with discernible risk factors, including diabetes, hypertension, hyperlipidemia, and a history of smoking, who are at considerable risk of developing CKD, especially those with pre-existing heart disease. This skill

facilitates early identification and focused intervention for high-risk groups.

• **Improved Early Detection and Management:**

The HSCDT classifier offers a more precise and dependable predictive tool, assisting physicians in the early identification of at-risk patients, hence facilitating prompt interventions that can decelerate disease development and enhance outcomes. This signifies a substantial progression in the domain of precision medicine.

The research underscores the HSCDT model's potential for integration into current clinical processes, thereby strengthening decision-making, decreasing healthcare expenditures, and improving the overall quality of treatment for patients with intricate comorbidities. This research aims to offer a comprehensive, scalable, and interpretable instrument for the early detection and management of CKD in patients with heart disease, thereby enhancing patient outcomes and alleviating the impact of these debilitating conditions on individuals and healthcare systems.

The paper is structured as follows: Section 2 presents a comprehensive literature analysis, addressing the pathophysiology of CKD and CVD, the common risk factors and mechanisms connecting the two disorders, and the recent advancements in research and ML applications in healthcare. Section 3 delineates the approach, encompassing the dataset description, preparation procedures, and the ML techniques utilized in the study. Section 4 presents the suggested Novel Hybrid Spectral Clustering Decision Tree (HSCDT) model, outlining its theoretical underpinnings, implementation, and optimization procedures. Section 5 delineates the experimental outcomes, encompassing the performance metrics of the HSCDT model and a comparative analysis with conventional classifiers. Section 6 closes the study by addressing the clinical significance of the findings, the limits of the research, and prospective avenues for future investigation.

2. LITERATURE REVIEW

ML techniques have been one of the focus areas in CKD for diagnosis, prognosis, and risk prediction. Different CKD diagnosis was explored in various research using different ML models, and other studies focusing on a condition that shares several risk factors with CKD, which is the cardiovascular disease (CVD). CKD is progressive and kidney susceptibility or dysfunction needs to be diagnosed early to prevent or delay the

development towards end-stage kidney disease, the integration of computational intelligence into CKD diagnosis has shown promising results, enhancing early diagnosis and improving patient outcomes. The preceding research are presented below:

2.1 Machine Learning in CKD Diagnosis

Many studies have looked at ML-based methods for CKD diagnosis and prediction. For CKD stage prediction, Ilyas et al. (2020) assessed decision tree algorithms, more especially, J48 and Random Forest. Their results indicated that J48 exceeded Random Forest with an accuracy of 85.5%, therefore highlighting its possible use for automated CKD severity identification [15]. Emphasizing the promise of AI-driven early CKD detection, Senan et al. (2021) examined a dataset of 400 patients and found that the Random Forest method attained 100% accuracy, precision, recall, and F1-score [16].

Sobrinho et al. (2020) conducted a comparative assessment of ML algorithms for CKD diagnosis in devel-oping countries. Their findings demonstrated that the J48 decision tree was the most appropriate strategy, with 95% accuracy, which closely corresponded with expert nephrologist evaluations. This work emphasized the feasibility of decision tree-based models in resource-limited environments [17]. Moreover, Thongprayoon et al. (2021) investigated ML algorithms for forecasting kidney failure risk in CKD patients, illustrating that ML models could improve risk prediction beyond conventional models, such as the Kidney Failure Risk Equation (KFRE), by incorporating various predictive factors [18].

2.2 ML for CKD and CVD Prediction

Given the strong interaction between CKD and CVD, various research has looked at ML models for both dis-ease prediction. With Matthews correlation coefficients of +0.499 and +0.469, Chicco et al. (2021) examined health records using ML approaches to forecast severe CKD in patients with CVD, therefore obtaining great predictive accuracy. Age, estimated glomerular filtration rate (eGFR), and creatinine were key clinical predictors; when temporal data was incorporated, hypertension, smoking, and diabetes became especially important [19].

AUC of 0.946 and AP of 0.678 were obtained when Ye et al. (2023) used Gradient

Boosting Decision Tree (GBDT) models to forecast in-hospital mortality in CKD patients with coronary artery disease. Their results highlighted how ML contributes to mortality risk assessment; SHAP analysis found the top 20 predictive elements [20]. With risk factors including age, hypertension history, sex, antiplatelet drug use, HDL levels, salt, 24-hour urine protein, and eGFR, Zhu et al. (2024) created an ML model for estimating CVD risk in CKD patients that achieved an AUC of 0.89. But their study depended on past data and needed outside confirmation [21].

2.3 ML-Based Prediction of CKD Progression

With an accuracy of 99.5%, Chhabra et al. (2023) predicted CKD using an ensemble learning method including SMOTE. Within their stacking classifier architecture, the research evaluated 12 classifiers and found Support Vector Machine (SVM), Random Forest, and AdaBoost to be the most successful models [22]. Likewise, Islam et al. (2023) selected a subset of 7–8 important features from an initial 25 variables XGBoost was used to predict CKD. Though issues with model selection and generalizability remained [23], their model showed great performance with accuracy, precision, recall, and F1-score approaching 0.98.

With great accuracy and recall, Venkatesan et al. (2023) also used XGBoost for early CKD identification. Their study did not, however, specifically address restrictions [24]. Using health insurance claims data from Taiwan, Krishnamurthy et al. (2020) created ML models projecting CKD onset within 6–12 months. With AUROC scores of 0.957 and 0.954, their convolutional neural network (CNN) model surpassed others. Strong performance notwithstanding identified issues with data generalizability and noise [25].

For estimating end-stage kidney disease (ESKD), Bai et al. (2022) evaluated ML models against the Kidney Failure Risk Equation (KFRE). KFRE maintained better general accuracy even as logistic regression, naïve bayes, and random forest models showed better sensitivity. The study highlighted ML's potential to enhance CKD prediction but also urged outside validation [26].

2.4 ML for Heart Disease Prediction and Its Relation to CKD

Given the strong connection with CKD, multiple research investigations have examined ML models for heart disease prediction. Developing ML models for cardiovascular illness prediction, Bhatt et al. (2023) found that the multilayer perceptron (MLP) model attained the greatest accuracy (87.28%) and an AUC of 0.95. But the study limited itself to one dataset, therefore influencing generalizability [27]. In order to forecast cardiovascular illness with an accuracy of 99.05% [28], Ghosh et al. (2021) also used Relief feature selection technique and Random Forest Bagging Method (RFBM).

Particularly pertinent for CKD patients since serum creatinine is a fundamental biomarker of renal function, Chicco et al. (2020) showed that ejection fraction and serum creatinine by themselves could effectively predict heart failure survival [29]. Using ML models, Tseng et al. (2020) projected acute kidney dam-age (AKI) during cardiac surgery, noting intraoperative urine output, red blood cell transfusion, and hemoglobin level as major predictors. These results highlight the need of ML-driven risk assessment models [30] as well as the interaction between renal and cardiac conditions.

Boosting algorithms and support vector machines (SVM) have great predictive power, especially for coronary artery disease and stroke, according to a meta-analysis on ML-based cardiovascular disease prediction published by Krittanawong et al. (2020). Their results imply that comparable ML models might be used for CKD prediction, therefore underlining the requirement of integrated diagnostic models [31].

Although encouraging outcomes, some difficulties still exist in ML-based CKD diagnosis and prediction. The limited datasets utilized in numerous studies have raised concerns regarding the generalizability of the model. Variations in preprocessing methods and feature selection between studies influenced repeatability. Moreover, most research did not specifically address model interpretability - a key consideration for clinical acceptance.

Integration of CKD and heart disease prediction models should be the main emphasis of

future research to improve diagnosis accuracy for individuals with comorbid diseases. Including environmental, lifestyle, and genetic elements into ML models might help to better estimate risk. Furthermore, crucial for the development of strong and therapeutically relevant solutions will be growing datasets and validating models over various populations.

Although numerous works have applied machine learning models to predict CKD or CVD independently, limited research has been conducted on the joint prediction of both conditions, particularly in patients with pre-existing cardiovascular disease. Traditional classifiers were predominantly used in earlier works, with minimal focus on interpretability, thereby restricting their applicability in clinical environments. Furthermore, the integration of hybrid models that combine unsupervised and supervised learning has rarely been explored. In the present research, these gaps have been addressed through the development of a Hybrid Spectral Clustering Decision Tree (HSCDT) model. Enhanced predictive performance and clinical interpretability have been achieved. By focusing on the early identification of CKD in individuals with heart disease, a novel and targeted direction has been provided in the con-text of managing comorbidities, which remains inadequately addressed in existing literature.

3. PROPOSED METHODOLOGY

The proposed research aims to analyze the complex relationship between CKD and HF through the application of machine learning algorithms for accurate prediction and early diagnosis. The work includes numerous key stages: data extraction from clinical datasets, preprocessing and feature extraction, shoot up spectral clustering to aggregate similar examples. In addition, decision tree-based model will be used for prediction, various classifiers will be evaluated based on accuracy, precision, recall, and F1-score. Figure 1 illustrates the architecture diagram of our proposed work. Details of each of these steps follows.

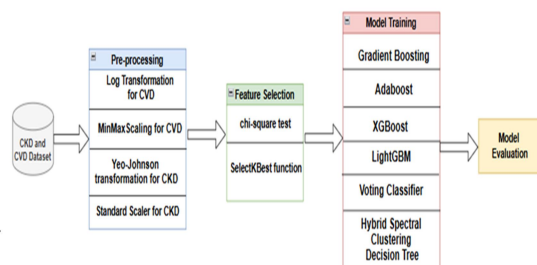


Figure 1: Proposed Work Architecture

3.1 Data Collection

The datasets used for this research were downloaded from Kaggle [32]; a popular platform for data science & ML. The Heart Failure Clinical Records Dataset and the CKD Dataset were the two key datasets used. The Heart Failure Clinical information Dataset consists of data on HF patients and contains 13 variables that describe individual features. It organizes events into deaths from HF and non-fatal events that can progress to fatal ones. The goal of this dataset is to capture important features and clinical signs related to HF to be used for analysis and prediction. The Chronic Kidney Disease Dataset relates to individuals with CKD, and contains 25 attributes that provide relevant information for both those who are affected and those who are not. It divides people into two categories: those with chronic renal disease and those without renal disease. This dataset is designed to aid in identifying important features/factors related to chronic renal disease and cardiac disease. Common features such as age, sex, diabetes, blood pressure, and BMI were aligned across the two datasets.

The dataset contains 32 attributes regarding CKD and CVD. These features include age, sex, cigsPerDayindex, BPLeds, prevalentStroke and prevalentHyp. Additional features include diabetes, total cholesterol (totChol), systolic and diastolic blood pressure (sysBP, diaBP), body mass index (BMI), heart rate, and glucose levels. Data includes ten-year coronary heart disease risk (TenYearCHD) as well as diabetes history (HistoryDiabetes), coronary heart disease history (HistoryCHD), and vascular disease history (HistoryVascular). Other variables included smoking status, history of dyslipidemia, and history of obesity. Other characteristics included medications used to treat dyslipidemia (DLDmeds), diabetic medications (DMmeds), antihypertensives (HTNmeds), and ACEI or ARBs (ACEIARB). Other features include baseline cholesterol (CholesterolBaseline), baseline creatinine (CreatinineBaseline), estimated glomerular filtration rate (eGFRBaseline), baseline systolic and diastolic blood pressure (sBPBaseline, dBPBaseline). It consists of baseline BMI (BMIBaseline), months to event (TimeToEventMonths), and Time in years (TIME_YEAR).

3.2 Data Pre-processing

The pre-processing and scaling of continuous variables were critical steps in preparing the datasets for analysis. In the HF Clinical Records Dataset, a log transformation was applied to reduce skewness in the continuous features, aiming to approximate a log-normal distribution. This transformation is particularly effective for right-skewed data. Following this, MinMaxScaler was used to scale specific variables - creatinine, ejection fraction, platelet counts, and serum sodium - to a range of [0, 1]. This method ensures that all values are proportionally adjusted within the specified range, making it suitable for algorithms sensitive to feature magnitudes. For the CKD dataset, the Yeo-Johnson transformation was employed to handle skewness, which is a more flexible approach than the log transformation as it accommodates zero and negative values. This transformation effectively reduces skewness while maintaining the interpretability of the data. Subsequently, StandardScaler was applied to variables such as red blood cell count, white blood cell count, hemoglobin, serum sodium, random blood glucose, blood pressure, and age. StandardScaler standardizes the data by removing the mean and scaling to unit variance, which is beneficial for algorithms that assume normally distributed features. Despite these transformations, the presence of outliers in the data meant that some skewness persisted, highlighting the need for additional outlier-handling techniques in future work. Overall, these preprocessing steps ensured that the datasets were normalized and standardized, laying a strong foundation for accurate and reliable predictive modeling.

3.3 Feature Selection

Feature selection is done in an iterative way for dimension reduction, refinement of the model, and coming up with important features associated with the target variable which is CKD. First, the dataset was separated into two: the features (X) and the target variable (y). Feature columns are obtained from a predetermined list called `feature_columns` and the target variable is obtained by filtering the main dataset for the 'CKD' column. The chi-square test is used as the feature selection method and is useful to measure the association between two categorical variables. It tests the relationship between each feature and target variable, 'CKD', and determines whether the observed distribution of data is statistically significantly different from the expected distribution of the data under the assumption of

independence. Thus, it is more appropriate for selecting features highly correlated with the target variable. Because it does not make any assumption of linearity between variables, it works perfectly for datasets where continuous features are discretized or in the case of categorical features. Scoring Chi Square method ranks features according to their importance in predicting the target variable. It helps not just in terms of dimensionality reduction, but also facilitates model interpretation as it can show the model, the most important features.

The chi-square test is performed using `SelectKBest` function from the library `scikit-learn`. This function is used to select the top-k non-zero Chi-square features. The `fit_transform` method is called to perform the chi-square test on the features (X) and target variable (y) and returns only the slice of features selected. Lastly, names of the feature columns included in the model are extracted so they can be used for analysis and for training the model.

The chi-square test offers several advantages in the feature selection process. It provides a statistical measure of how a feature relates to the target value (outcome). It reduces dimensionality, thus helping computational efficiency and reducing the risk of overfitting as well. Finally, it also improves the model interpretability by highlighting in the model the most important features predicting target variable, CKD. This technique is especially powerful for datasets where a lot of features are categorical or discretized continuous, since there is no assumption of linearity, making it an extremely effective and robust method for feature selection.

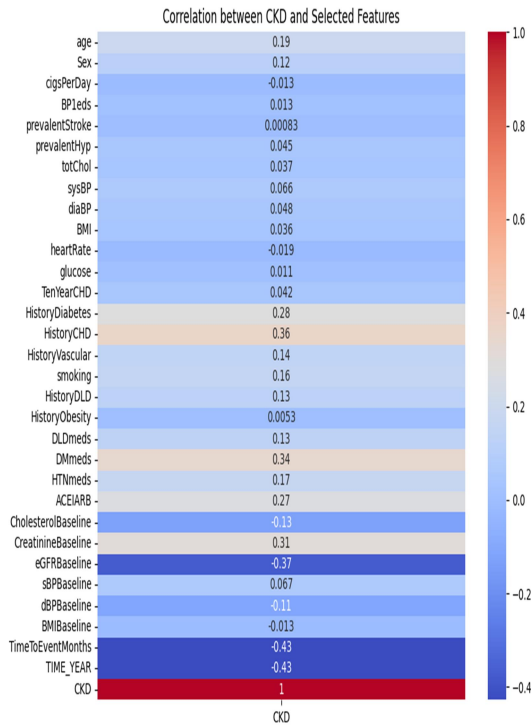


Figure 2: Correlation between CKD and the Selected Features

Figure 2. displays correlation coefficients between CKD and several variables, including age, sex, smoking habits, blood pressure, cholesterol levels, and medical history. Correlation coefficients span from -1 to 1, where values approaching 1 or -1 signify robust positive or negative associations, respectively, while values close to 0 indicate minimal to no association. This information is crucial for identifying characteristics most closely linked to CKD, facilitating diagnosis, prevention, and therapy planning.

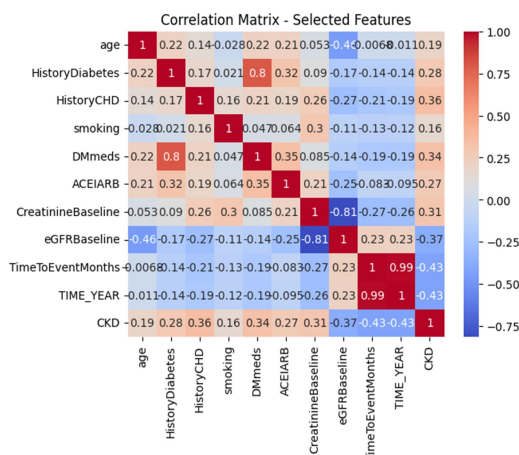


Figure 3: Correlation Matrix of the Selected Features

Based on the performance, the following correlation analysis is made for the correlation of CKD and features selected: HistoryCHD (0.36), eGFRBaseline(0.37), CreatinineBaseline(0.31) & DMmeds(0.34), enabling thereby their positive significant association with CKD. Moderate correlations of HistoryDiabetes (0.28), HTNmeds (0.17), smoking (0.16) and HistoryVascular (0.14) underscore the contributions of diabetes, hypertension, smoking and vascular health to CKD risk. The features age (0.19), BMI (0.036), totChol (0.037), and glucose (0.011) have weak or negligible correlations, indicating that these features do not appear to directly impact CKD. As expected, negative correlations are observed with TimeToEventMonths (- 0.43) and TIME_YEAR (- 0.43), as CKD is temporally progressive in nature. It highlights notable predictors CHD history, markers for kidney function and diabetic specific insights in CKD risk assessment and modelling.

Figure 3. provides an in-depth analysis of the correlations among many variables, extending beyond their association with CKD. Comprehending these interrelationships is essential for developing predictive models and examining the intricate linkages among diverse health indicators. For example, it can indicate if specific conditions or therapies are more likely to coexist.

3.4 Model Training and Evaluation

In model development stage, selected features are used as input to several ML models such as Gradient Boosting, Adaboost, XGBoost, LightGBM and Voting Classifier. These models are evaluated based on their ability to predict and classify the target variable. The prediction performance of the models is evaluated through different assessment criteria, such as accuracy, precision, recall, and f1-score. K-fold cross-validation is used to ensure that the model's performance is not biased on any one data subset. Essentially, this technique divides data into several sets, trains the model on different combination of these subsets while evaluates its performance in the other subsets. Below are the algorithms used in this investigation.

3.4.1 Gradient Boosting

Gradient boosting is a proficient ensemble learning method employed for the identification of CKD in the given dataset. It operates by amalgamating multiple weak classifiers to create a resilient prediction model. In contrast to alternative methods, gradient boosting develops models sequentially, with each subsequent model rectifying

the flaws of its predecessor. This iterative method entails computing residuals, refining forecasts, and modifying weights to improve precision. Gradient boosting can manage several feature types and discern complex patterns within the data. The ultimate value of the gradient boosting model is determined using equation (1).

$$F(x) = \sum_{m=1}^M \gamma_m \overline{h_m(x)} \quad (1)$$

Where,

- $F(x)$ denotes the ultimate anticipated value for the input sample x .
- M is the aggregate quantity of weak learners (trees) inside the ensemble.
- $h_m(x)$ represents the prediction generated by the m^{th} tree for the input sample x .
- γ_m signifies the weight (or learning rate) of the m^{th} tree, reflecting the degree to which this tree's prediction affects the ultimate outcome.

3.4.2 Adaboost

AdaBoost, an acronym for Adaptive Boosting, is an ensemble learning technique employed for CKD diagnosis from the given dataset. The objective is to construct a robust classifier by amalgamating multiple weak learners. AdaBoost allocates weights to each training sample, emphasizing misclassified cases in subsequent iterations. The weak learners are trained sequentially, with each one adjusting its performance according to the outcomes of the preceding models. The ultimate prediction is obtained by consolidating the weighted votes from all weak classifiers. The comprehensive forecast is derived from equation (2).

$$F(x) = \text{sign}(\sum_{m=1}^M \alpha_m \cdot h_m(x)) \quad (2)$$

Where,

- α_m denotes the weight or importance of the prediction generated by the m -th weak learner within the overall ensemble.
- M represents the aggregate count of weak learners within the ensemble.
- $F(x)$ is the ultimate forecast generated for the input sample x .
- $h_m(x)$ denotes the prediction generated by the m -th weak learner for the input sample x .
- The function $\text{sign}(\cdot)$ yields +1 for non-negative arguments and -1 for negative arguments.

3.4.3 XGBoost

The XGBoost classifier was utilized for CKD classification. It employed a gradient boosting approach that constructs an ensemble of decision trees for predictive purposes. To prevent overfitting, regularization techniques like L1 and L2 were implemented, and tree pruning was employed to enhance efficiency. XGBoost effectively handled missing values, provided insights into feature significance, and enabled cross-validation. The goal function for XGBoost can be calculated utilizing equation (3).

$$\text{Obj} = \sum_{i=1}^N l(y_i, F(x_i)) + \sum_{m=1}^M \Omega(f_m) \quad (3)$$

Where,

- N denotes the aggregate quantity of training instances.
- M signifies the quantity of weak learners (trees) inside the ensemble.
- y_i denotes the actual label of the i^{th} training instance.
- x_i is the feature vector associated with the i^{th} training instance.
- $F(x_i)$ is the ultimate prediction generated by the XGBoost model for the i^{th} training instance.
- $l(y_i, F(x_i))$ is the loss function that measures the prediction error between the real label y_i and the predicted value $F(x_i)$.
- f_m denotes the m^{th} weak learner (tree) in the ensemble.
- $\Omega(f_m)$ is the regularization term that penalizes excessively intricate trees to prevent overfitting.

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (4)$$

Where,

- T denotes the aggregate quantity of leaves in the tree f .
- γ is the parameter that governs the complexity of the tree by imposing a penalty on the quantity of leaves.
- λ represents the L2 regularization term that imposes a penalty on the leaf weights to mitigate overfitting.
- w is the vector of weights allocated to the leaves.

3.4.4 LightGBM

LightGBM is an ensemble learning technique employed for CKD identification from the supplied dataset. It employs a leaf-wise tree development

technique, selecting leaf nodes according to the maximum enhancement in the loss function, resulting in improved accuracy. Furthermore, LightGBM naturally accommodates categorical characteristics, obviating the necessity for one-hot encoding and enhancing training efficiency. The formulas for LightGBM are given in equations (5) and (6).

$$Obj = \sum_{i=1}^N \ell(y_i, F(x_i)) + \sum_{m=1}^M \Omega(f_m) \quad (5)$$

where:

- N denotes the quantity of training examples.
- M denotes the quantity of weak learners (trees) in the ensemble.
- y_i represents the true label of the i^{th} training instance.
- x_i denotes the feature vector of the i^{th} training instance.
- $F(x_i)$ is the ultimate output or prediction generated by the LightGBM model for the i^{th} training instance.
- $\ell(y_i, F(x_i))$ is the loss function employed to assess the prediction error between the actual label y_i and the predicted value $F(x_i)$. Prevalent loss functions for binary classification comprise log-loss (cross-entropy) and binary hinge loss.
- f_m denotes the m^{th} weak learner (tree) within the ensemble.
- $\Omega(f_m)$ serves as the regularization term that penalizes intricate trees to mitigate overfitting.

$$\Omega(f) = \gamma \cdot \text{number_leaves} + \frac{1}{2} \lambda \sum_{j=1}^{\text{num_leaves}} w_j^2 \quad (6)$$

Where,

- num_leaves is the number of leaves in the tree f .
- γ is the parameter that controls the complexity of the tree by penalizing the number of leaves.
- λ is the L2 regularization term that penalizes the weights w_j of the leaves to avoid overfitting.

3.4.5 Voting Classifier

The pre-processed data was utilized to train the Voting Classifier, an ensemble ML model that amalgamates multiple individual classifiers to produce predictions. Voting classifier consolidates the predictions from each classifier and establishes the final result based on a majority vote for

classification tasks or averaging for regression tasks. The pre-processed dataset was supplied to the Voting Classifier, which leveraged the variety of its component models to enhance predicted accuracy and generalization on novel, unseen data. The Voting Classifier can integrate multiple classifier types, including Decision Trees, Random Forests, Support Vector Machines, and Logistic Regression. Final prediction formula is given in equation (7).

$$\text{FPred} = \text{mode} \{C_1(x), C_2(x), \dots, C_m(x)\} \quad (7)$$

Where,

- C – is the class

4. NOVEL HYBRID SPECTRAL CLUSTERING DECISION TREE (HSCDT)

Predicting CKD in individuals with CVD mostly depends on a suitable model selection and efficient approach. The Hybrid Spectral Clustering Decision Tree (HSCDT) is the one selected for this aim since it can detect intricate trends and linkages inside high-dimensional medical data. This method forecasts CKD in heart illness patients by combining the strengths of spectral clustering and decision tree classifiers using the proposed Hybrid Spectral Clustering Decision Tree (HSCDT) model. By spotting co-clustering interactions and exposing latent features in the data that other methods might ignore, spectral clustering overcomes the linear restrictions of conventional distance measurements. This approach is especially useful for medical data where interactions, including those involving comorbidities, can be somewhat erratic since it is especially good in catching non-linear relationships.

The major aspect of HSCDT model is its ability to divide the patient population into subgroups with similar characteristics. The grouping this method creates allows us to show the relations between different aspects, creating our understanding of patient risk factors to be more advanced, and also improving the predictive power of the model. Through the identification of these subgroups, HSCDT improves CKD prediction and facilitates development of targeted treatment approaches based upon risk profiles. Unlike conventional models such as logistic regression, random forest, or support vector machines, which mainly focus on linear relationships and existing rules, HSCDT model can adapt to the complex, multi-dimensional data correlated with heart

disease and CKD. By using spectral clustering, the model is suited to deal with different formats of data and provide more generalized predictions. Additionally, this decision tree component offers the model explainability, which is crucial for medical deployments, as understanding the rationale behind decisions made is essential for physicians.

Therefore, the Hybrid Spectral Clustering Decision Tree (HSCDT) model could outperform traditional approaches by: one, delivering improved accuracy; two, guaranteeing advanced adaptability; and three, improving the understanding of CKD prediction in patients with heart disease, which helps to enhance decision making. The algorithm for the proposed work can be seen below.

Algorithm: Hybrid Spectral Clustering Decision Tree (HSCDT)

Input:

- $X = \{x_1, x_2, \dots, x_n\}$: Feature matrix (patients' data with attributes like age, cholesterol, blood pressure, etc.)
- $Y = \{y_1, y_2, \dots, y_n\}$: Ground truth labels (0 for no CKD, 1 for CKD)
- k : Number of clusters for spectral clustering

Step 1: Perform Spectral Clustering

a. Compute Similarity Matrix S :

The similarity matrix captures the relationships between pairs of data points. A Gaussian kernel is commonly used:

$$S(i, j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

where x_i and x_j are feature vectors for patients i and j , respectively, and σ is the parameter controlling the kernel's width.

b. Compute the Degree Matrix D :

The degree matrix is a diagonal matrix where each entry $D(i, j)$ is the sum of the similarities for patient i :

$$D(i, j) = \sum_{j=1}^n S(i, j)$$

c. Compute the Laplacian Matrix L :

The Laplacian matrix L is derived from the degree and similarity matrices:

$$L = D - S$$

d. Eigenvalue Decomposition:

Perform eigenvalue decomposition on the Laplacian matrix to obtain the eigenvalues and eigenvectors:

$$L_v = \lambda_v v$$

where v is the eigenvector and λ is the eigenvalue.

e. Select the Top k Eigenvectors:

Select the first k eigenvectors corresponding to the smallest eigenvalues. These eigenvectors are used to form the feature matrix for clustering:

$$V_k = [v_1, v_2, \dots, v_k]$$

f. Cluster Data Using K-means:

Use the k eigenvectors to perform k -means clustering. This groups patients into k clusters based on their feature similarity:

$$C = KMeans(V_k, k)$$

Step 2: Train a Decision Tree Classifier

a. Prepare the Training Dataset:

Combine the feature matrix X with the cluster labels C from the spectral clustering step to form the training dataset.

$$X_{train} = [x_1, x_2, \dots, x_n], \\ C_{train} = [C_1, C_2, \dots, C_n]$$

b. Train the Decision Tree:

Use the training dataset X_{train} and the corresponding labels C_{train} to train a decision tree classifier. The decision tree uses recursive partitioning based on feature splits that minimize the impurity (e.g., Gini Index or Entropy).

$$Gini(S) = 1 - \sum_{k=1}^m p_k^2$$

where p_k is the proportion of samples from class k in set S .

Step 3: Predict CKD Status

a. Prediction Using the Decision Tree:

After the model has been trained, use it to predict the CKD status for new patient data. The decision tree will output either 0 or 1, indicating the presence or absence of CKD.

$$\hat{y} = PredictDecisionTree(X_{test})$$

where \hat{y} is the predicted CKD status for each test patient.

Step 4: Model Evaluation

Standard metrics including accuracy, precision, recall, and F1-score help one assess the model's performance.

5. EXPERIMENTAL RESULTS

The experiment was performed via Google Colab, a cloud-based Jupyter notebook platform. Essential libraries such as NumPy, Pandas, Matplotlib, Seaborn, and Scikit-learn were installed with `!pip install`. All outcomes and visual representations were produced within the Colab environment.

Experimental results for the research provide a comprehensive comparison among different ML models for CKD prediction. Various models such as Gradient Boosting, Adaboost, XGboost, LightGBM, Voting Classifier and Hybrid Spectral Clustering Decision Tree (HSCDT) which we proposed have been evaluated on different metrics such as cross validation precision, test set accuracy, precision recall and F score. The evaluation aimed to determine the optimal efficient model for CKD forecasting, placing significant weight on accuracy and memory — critical factors in health care settings where false affirmative and negative findings can have grave consequences. The Gradient Boosting achieved CV accuracy of 91% while the test set accuracy was 92%. However, its precision of 0.70 and recall of 0.46 indicates that the model was relatively accurate but failed to identify a large percentage of actual CKD cases (false negatives), leading to a relatively low F1 score of 0.56. That indicates that in this case, no selection should have creation of Gradient Boosting, as it was able to balance precision and recall, resulting in low identification of CKD.

Compared to Adaboost, our naive predictive algorithm performed really bad achieving in the best case a cross-validation accuracy of 93% and an excellent test set accuracy of 99%. A precision of 1.00 suggested there were no false positive predictions, while a recall of 0.95 indicated the model was able to detect most of the actual CKD cases. With an F1 score of 0.98, there was an excellent balance between precision and recall, which confirmed Adaboost as one of the best models for CKD prediction. Its extreme performance is a result of being adaptive; focusing on those data points which were misclassified, and incrementally improving performance. While XGBoost showed a slightly lower cross-validation accuracy of 92%, it still achieved a respectable test set accuracy of 98%. XGBoost yielded 0.95 and 0.89 precision and recall, respectively, suggesting

strong performance with relatively high false negative rate over Adaboost. With the F1 score of 0.92, this was a good performance, indicating that CKD prediction demonstrated effectiveness; however, this was not completely in accordance with the overall performance with Adaboost.

LightGBM struggled with precision and recall and produced a cross-validation accuracy of 90% and test set accuracy of 91%. Happening, the model was only able to identify 29% of true CKD instances with 67% precision, which means it missed many CKD occurrences but predicted many false positive cases. The F1 score of 0.40 alone is the lowest across all models showing the difficulties LightGBM had on this particular classification problem. Although the Voting Classifier, an ensemble built from numerous models, only achieved a 94% cross-validation score, its test set achieved a 99% accuracy equal to Adaboost. The 0.95 precision, 0.89 recall, and 0.92 F1 score indicate that this could be an effective tool for identifying CKD. The Voting Classifier takes advantage from individual models, so overcome the weakness of any one model. Although this ensemble method was effective, it could not outperform the Proposed Hybrid Spectral Clustering Decision Tree (HSCDT).

The proposed HSCDT model outperformed all former models on cross-validation and test set accuracy. Results from the model showed a cross-validation accuracy of 99% and a test set accuracy of 99%, demonstrating its capabilities for generalizing to unseen data. The model achieved a maximum precision (0.99) and recall (0.97), showing that it made very few positive predictions of actual CKD patients and identified almost all people with real CKD. The highest F1 score across scenarios (0.98) suggests that the generated model provides a nearly perfect balance between precision and recall. The outstanding performance is attributed to the unique hybrid method adopted by HSCDT, which combines the benefits of spectral clustering and decision trees. The first part, spectral clustering, offers a way to discover hidden patterns in the data, and the second, decision tree, makes decisions based on these patterns that are interpretable. With this hybridization we develop a more accurate and transferrable model for predicting the presence of CKD than is found with traditional methods.

There are multiple benefits of this model (HSCDT compared to the other models evaluated). This hybrid design makes it easier for complex dependencies to interact, and given that in medical applications, variables such as patient age, medical

history, and lifestyle habits may have non-linear interactions, this is indeed helpful. Secondly, the ability of the model to balance accuracy and recall ensures minimization of both false positives and false negatives. The impact of this is particularly important for CKD prediction, since a false negative may delay CKD diagnosis and treatment, while a false positive may cause unnecessary interventions in a person that is CKD-free. Decision trees provide transparency in their predictions, allowing medical stakeholders to grasp the basis for predictions and, in turn, building trust in automated medical systems.

Due to the HSCDT model's high precision, recall, and F1 score, and of its ability to capture complex relationships in the data, it becomes a powerful tool for CKD prediction. It did outperform Adaboost, XGBoost, and Voting Classifier, which were promising models as well but did not reach the accuracy we achieved in finding CKD patients. In contrast, LightGBM struggled to achieve acceptable performance, highlighting the importance of model selection for medical applications.

The HSCDT model demonstrated superior performance when compared to traditional ML approaches in predicting CKD and represents a potential tool for timely diagnosis and treatment. Its ability to maintain high accuracy, precision, recall, and F1 score, as well as interpretability, make it an ideal choice for medical applications, where understanding and trust in the model's recommendations are essential. Additional insights gained through feature connections add to the overall value of the model and provide a strong instrument for healthcare practitioners in the fight against CKD.

Figure 4. indicates that elevated systolic blood pressure (sysBP) is correlated with those possessing CKD in contrast to those without CKD. In the systolic blood pressure range of 135 to 150, a greater prevalence of persons with CKD is observed, indicating a possible correlation between increased blood pressure and CKD. Figure 5 indicates that patients with CKD are often older, with a concentration of ages in the higher range (e.g., 55 and above), in contrast to those without CKD. This indicates a possible correlation between advancing age and the occurrence of CKD. Figure 6 The graphic indicates that persons with CKD exhibit a greater prevalence of a history of coronary heart disease (CHD) than those without CKD. This indicates a possible association between CKD and heightened risk of coronary heart disease. Figure 7 indicates that individuals with CKD generally have

higher BMI (>25) values than those without CKD. This suggests a possible link between elevated BMI and the presence of CKD, highlighting obesity or overweight as a potential risk factor for CKD. Figure 8 indicates that persons with CKD exhibit a greater prevalence of smoking (more than 2 years) than those without the condition. This suggests a potential correlation between smoking and the occurrence of CKD, identifying smoking as a prospective risk factor for CKD. Table 1 shows the accuracy of the classifiers. Figure 9 shows the classification accuracy. The accuracy, precision, recall and F1 score the classifiers are shown in Table 2. According to Table 3, the value for each factor and parameter was established relative to the occurrence of CKD.

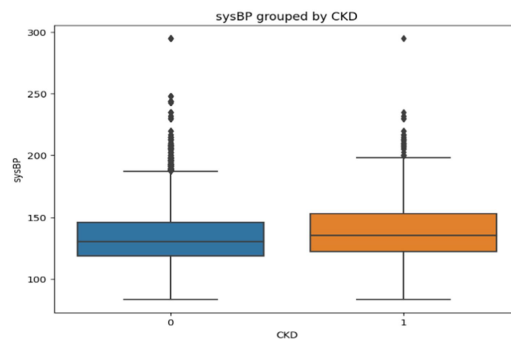


Figure 4: sysBP grouped by CKD

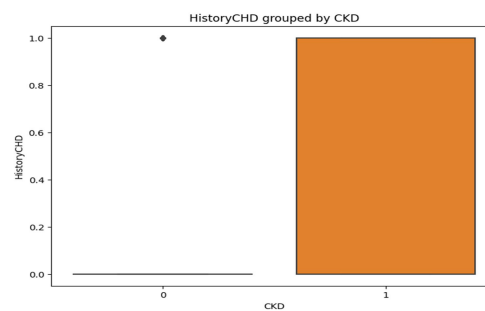


Figure 5: Age grouped by CKD

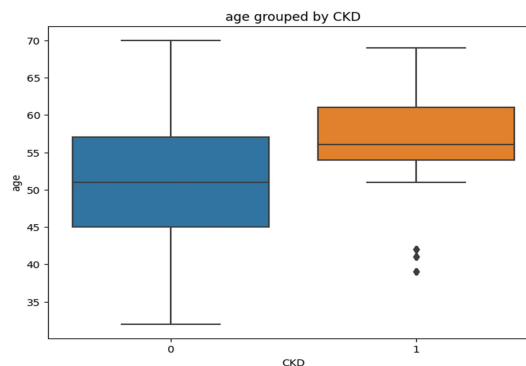


Figure 6: HistoryCHD grouped by CKD

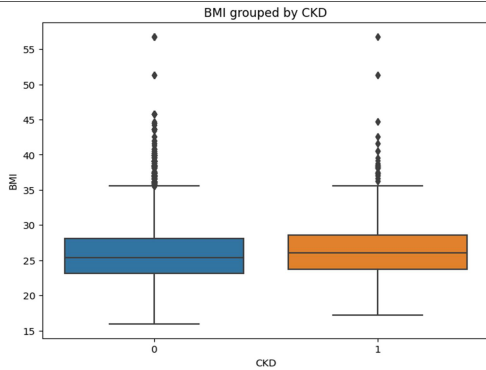


Figure 7: BMI grouped by CKD

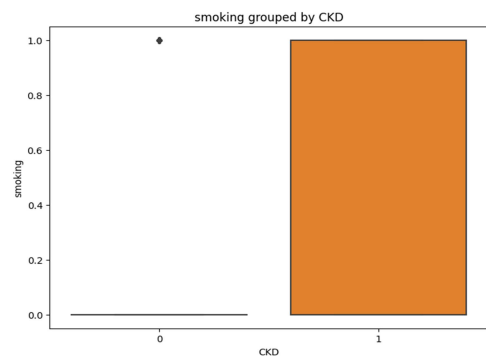


Figure 8: Smoking grouped by CKD

Table 1: Accuracy of the Classifiers.

Model	Average Cross-Validation Accuracy	Accuracy on Test Set
Gradient Boosting	91	92
Adaboost	93	99
XGBoost	92	98
LightGBM	90	91
Voting Classifier	94	99
Proposed Work	99	99

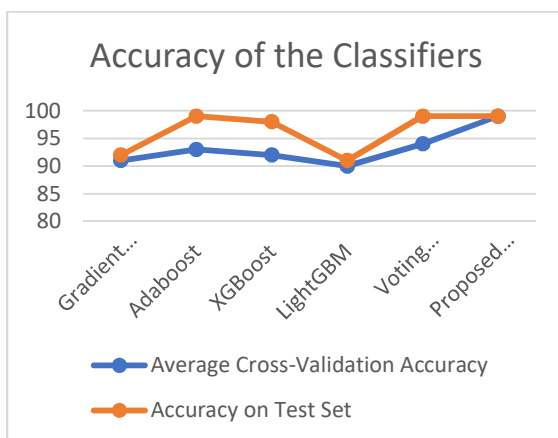


Figure 9: Accuracy of the Classifiers

Table 2: Precision, Recall and F1 Score of the Classifiers.

Model	Precision	Recall	F1 Score
Gradient Boosting	0.70	0.46	0.56
Adaboost	1.00	0.95	0.98
XGBoost	0.95	0.89	0.92
LightGBM	0.67	0.29	0.40
Voting Classifier	0.95	0.89	0.92
Proposed Work	0.99	0.97	0.98

Table 3: Factors and Values for Occurrence of CKD.

Model	Average Cross-Validation Accuracy	Accuracy on Test Set
Factors	Value	Event
Diabetes	Yes	CKD
Coronary Heart disease	Yes	CKD
BMI	greater than 25	CKD
Smoking for more than 2 years	Yes	CKD
Blood Pressure	greater than 130 mm	CKD
Cholesterol	more than 200 mg/dL are	CKD
Serum Creatinine	>1.2 mg/dL	CKD
Hypertension	130/90 mmHg.	CKD
eGFR	< 89 mL/min	CKD

Table 3 delineates the essential components and their respective values linked to the probability of CKD onset. These characteristics, determined through comprehensive medical re-search and clinical observations, operate as critical markers in assessing an individual's risk of CKD. The value of each factor is associated with an elevated likelihood of disease occurrence, rendering them essential in the diagnosis and prediction of CKD.

Diabetes is acknowledged as a substantial risk factor for CKD, since its presence markedly increases the probability of renal damage resulting from sustained hyperglycemia. Individuals with a history of coronary heart disease (CHD) are predisposed to CKD, as cardiovascular problems can adversely affect renal function. An elevated BMI over 25 has been associated with a heightened risk of CKD, as surplus weight imposes extra pressure on the kidneys. Prolonged smoking over two years has been demonstrated to increase risk, as

it diminishes renal blood flow, hence facilitating progressive kidney injury.

Hypertension (exceeding 130 mm Hg) constitutes a significant risk factor, as un-managed high blood pressure can impair renal blood vessels, leading to reduced kidney function. Increased cholesterol levels (exceeding 200 mg/dL) are associated with CKD, as they may lead to the accumulation of fatty deposits in blood arteries, particularly those that supply the kidneys, so compromising renal function. Serum creatinine levels exceeding 1.2 mg/dL signify renal impairment, as creatinine is a metabolic byproduct excreted by the kidneys; an elevation in its concentration indicates diminished renal function. Hypertension (exceeding 130/90 mm Hg) is an additional risk factor, and inadequately managed blood pressure is recognized as a primary contributor to renal disease. An eGFR (estimated Glomerular Filtration Rate) < 89 mL/min indicates compromised kidney function, with results beneath this threshold deemed crucial for diagnosing CKD.

The factors - diabetes, coronary artery disease, body mass index, tobacco use, hypertension, hyperlipidemia, serum creatinine levels, high blood pressure, and estimated glomerular filtration rate - are indispensable for diagnosing CKD. By integrating these key elements into predictive models like the HSCDT, it enables more accurate identification of high-risk individuals, leading to earlier detection and better disease management. This approach provides healthcare providers with a detailed understanding of the factors contributing to CKD, allowing for more powerful treatments and slowing the progression of disease.

6. CONCLUSION

CKD shares common risk factors and pathophysiological characteristics with CVD. Cardiovascular diseases and chronic renal disease is strongly connected, making early detection in patients with cardiovascular disorders crucial for effective treatment and improved patient outcomes. The focus of the research was based on prediction of CKD from heart disease patients, analyzing multiple factors that contribute to the disease and evaluation of various classifiers for prediction purposes. The Hybrid Spectral Clustering Decision Tree (HSCDT) model outperformed the existing classifiers, with an accuracy of 99% in CV and 99% on the test set. Additional evidence for its effectiveness were reinforced by improved precision, memory and F1-score. Key factors

influencing CKD, including age, BMI, smoking status, history of coronary heart disease, blood pressure, cholesterol status, serum creatinine, hypertension, and eGFR showed significant association with CKD prediction. The performance of the classifier was measured by various metrics, and it was confirmed that the proposed model achieved the highest classification accuracy. The prediction accuracy and reliability of the model increased by incorporating spectral clustering with decision trees.

This research has some limitations. The historical data that have been used may not be entirely representative of all patient types. The model needs to be tested on large and recent data points to check the reliability. As well, lifestyle factors and long-term patient data were not included, but could further enhance the predictive model. This is yet to be deployed in real hospitals. In the future, the dataset size can be increased, incorporate additional clinical variables, and fine-tune the optimization process to achieve better clinical performance.

AUTHOR CONTRIBUTION

The conceptualization and design of the research were contributed by Chandralekha E. The data collection, model development, and experimentation were carried out under her supervision. The manuscript draft was prepared and revised by her.

T. R. Saravanan provided guidance on the methodology, reviewed the technical accuracy of the proposed model, and contributed to refining the final version of the manuscript.

REFERENCES:

- [1] Bikbov, Boris, Caroline A. Purcell, Andrew S. Levey, Mari Smith, Amir Abdoli, Molla Abebe, Oladimeji M. Adebayo et al. "Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017." *The lancet* 395, no. 10225 (2020): 709-733.
- [2] World Health Organization. "Cardiovascular disease." http://www.who.int/cardiovascular_diseases/en/ (2017).
- [3] Wheeler, David C., and Wolfgang C. Winkelmayer. "KDIGO 2017 clinical practice guideline update for the diagnosis, evaluation,

- prevention, and treatment of chronic kidney disease-mineral and bone disorder (CKD-MBD) foreword." *Kidney international supplements* 7, no. 1 (2017): 1-59.
- [4] Matsushita, Kunihiro, Shoshana H. Ballew, Josef Coresh, Hisatomi Arima, Johan Ärnlöv, Massimo Cirillo, Natalie Ebert et al. "Measures of chronic kidney disease and risk of incident peripheral artery disease: a collaborative meta-analysis of individual participant data." *The lancet Diabetes & endocrinology* 5, no. 9 (2017): 718-728..
- [5] Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane. "Machine learning in medicine." *New England Journal of Medicine* 380, no. 14 (2019): 1347-1358.
- [6] Topol, Eric J. "High-performance medicine: the convergence of human and artificial intelligence." *Nature medicine* 25, no. 1 (2019): 44-56.
- [7] Johnson, A., L. Bulgarelli, T. Pollard, S. Horng, and L. A. Celi. "Mark." R. MIMIC-IV (version 1.0). *PhysioNet* (2021).
- [8] Ravikumar, S., and E. Kannan. "Analysis on mental stress of professionals and pregnant women using machine learning techniques." *International Journal of Image and Graphics* 23, no. 05 (2023): 2350038.
- [9] Shickel, Benjamin, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. "Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis." *IEEE journal of biomedical and health informatics* 22, no. 5 (2017): 1589-1604.
- [10] David, S. Alex, M. Leelavathi, G. G. Swathika, and N. Ruth Naveena. "Missing child monitoring system using deep learning methods a comparison." In *Artificial Intelligence, Blockchain, Computing and Security Volume 2*, pp. 339-343. CRC Press, 2023.
- [11] Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. "Dermatologist-level classification of skin cancer with deep neural networks." *nature* 542, no. 7639 (2017): 115-118.
- [12] Nusinovici, Simon, Yih Chung Tham, Marco Yu Chak Yan, Daniel Shu Wei Ting, Jialiang Li, Charumathi Sabanayagam, Tien Yin Wong, and Ching-Yu Cheng. "Logistic regression was as good as machine learning for predicting major chronic dis-eases." *Journal of clinical epidemiology* 122 (2020): 56-69.
- [13] Zhang, Yan, Weiwei Xu, Ping Yang, and An Zhang. "Machine learning for the prediction of sepsis-related death: a systematic review and meta-analysis." *BMC Medical Informatics and Decision Making* 23, no. 1 (2023): 283.
- [14] Vermeersch, Gaëlle, Edgard Prihadi, Gilles De Keulenaer, and Paul Vermeersch. "Giant native aortic valve thrombus under non-vitamin K antagonist oral anticoagulant: first manifestation of antiphospholipid syndrome." *European Heart Journal* 42, no. 19 (2021): 1927-1927.
- [15] Ilyas, Hamida, Sajid Ali, Mahvish Ponum, Osman Hasan, Muhammad Tahir Mahmood, Mehwish Ifthikhar, and Mubasher Hussain Malik. "Chronic kidney disease diagnosis using decision tree algorithms." *BMC nephrology* 22, no. 1 (2021): 273.
- [16] Senan, Ebrahim Mohammed, Mosleh Hmoud Al-Adhaileh, Fawaz Waselallah Al-saade, Theyazn HH Aldhyani, Ahmed Abdullah Alqarni, Nizar Alsharif, M. Irfan Uddin, Ahmed H. Alahmadi, Mukti E. Jadhav, and Mohammed Y. Alzahrani. "Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques." *Journal of healthcare engineering* 2021, no. 1 (2021): 1004767.
- [17] Sobrinho, Alvaro, Andressa CM Da S. Queiroz, Leandro Dias Da Silva, Evandro De Barros Costa, Maria Eliete Pinheiro, and Angelo Perkusich. "Computer-aided diagnosis of chronic kidney disease in developing countries: A comparative analysis of machine learning techniques." *IEEE Access* 8 (2020): 25407-25419.
- [18] Thongprayoon, Charat, Wisit Kaewput, Avishek Choudhury, Panupong Hansrivijit, Michael A. Mao, and Wisit Cheungpasitporn. "Is it time for machine learning algorithms to predict the risk of kidney failure in patients with chronic kidney disease?." *Journal of Clinical Medicine* 10, no. 5 (2021): 1121.
- [19] Chicco, Davide, Christopher A. Lovejoy, and Luca Oneto. "A machine learning analysis of health records of patients with chronic kidney disease at risk of cardiovascular disease." *IEEE Access* 9 (2021): 165132-165144.

- [20] Ye, Zixiang, Shuoyan An, Yanxiang Gao, Enmin Xie, Xuecheng Zhao, Ziyu Guo, Yike Li, Nan Shen, Jingyi Ren, and Jingang Zheng. "The prediction of in-hospital mortality in chronic kidney disease patients with coronary artery disease using machine learning models." *European Journal of Medical Research* 28, no. 1 (2023): 33.
- [21] Zhu, He, Shen Qiao, Delong Zhao, Keyun Wang, Bin Wang, Yue Niu, Shunlai Shang et al. "Machine learning model for cardiovascular disease prediction in patients with chronic kidney disease." *Frontiers in Endocrinology* 15 (2024): 1390729.
- [22] Chhabra, Divyanshi, Mamta Juneja, and Gautam Chutani. "An Efficient Ensemble-based Machine Learning approach for Predicting Chronic Kidney Disease." *Current Medical Imaging* 20, no. 1 (2024): e080523216634.
- [23] Islam, Md Ariful, Md Ziaul Hasan Majumder, and Md Alomgeer Hussein. "Chronic kidney disease prediction based on machine learning algorithms." *Journal of pathology informatics* 14 (2023): 100189.
- [24] Venkatesan, Vinoth Kumar, Mahesh Thyluru Ramakrishna, Ivan Izonin, Roman Tkachenko, and Myroslav Havryliuk. "Efficient data preprocessing with ensemble machine learning technique for the early detection of chronic kidney disease." *Applied Sciences* 13, no. 5 (2023): 2885.
- [25] Krishnamurthy, Surya, Kapeleshh Ks, Erik Dovgan, Mitja Luštrek, Barbara Gradišek Piletič, Kathiravan Srinivasan, Yu-Chuan Li, Anton Gradišek, and Shabbir Syed-Abdul. "Machine learning prediction models for chronic kidney disease using national health insurance claim data in Taiwan." In *Healthcare*, vol. 9, no. 5, p. 546. MDPI, 2021.
- [26] Bai, Qiong, Chunyan Su, Wen Tang, and Yike Li. "Machine learning to predict end stage kidney disease in chronic kidney disease." *Scientific reports* 12, no. 1 (2022): 8377.
- [27] Bhatt, Chintan M., Parth Patel, Tarang Ghetia, and Pier Luigi Mazzeo. "Effective heart disease prediction using machine learning techniques." *Algorithms* 16, no. 2 (2023): 88.
- [28] Ghosh, Pronab, Sami Azam, Mirjam Jonkman, Asif Karim, FM Javed Mehedi Shamrat, Eva Ignatious, Shahana Shultana, Abhijith Reddy Beeravolu, and Friso De Boer. "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques." *IEEE Access* 9 (2021): 19304-19326.
- [29] Chicco, Davide, and Giuseppe Jurman. "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone." *BMC medical informatics and decision making* 20 (2020): 1-16.
- [30] Tseng, Po-Yu, Yi-Ting Chen, Chuen-Heng Wang, Kuan-Ming Chiu, Yu-Sen Peng, Shih-Ping Hsu, Kang-Lung Chen, Chih-Yu Yang, and Oscar Kuang-Sheng Lee. "Prediction of the development of acute kidney injury following cardiac surgery by machine learning." *Critical care* 24 (2020): 1-13.
- [31] Krittanawong, Chayakrit, Hafeez Ul Hassan Virk, Sripal Bangalore, Zhen Wang, Kipp W. Johnson, Rachel Pinotti, HongJu Zhang et al. "Machine learning prediction in cardiovascular diseases: a meta-analysis." *Scientific reports* 10, no. 1 (2020): 16057.
- [32] <https://www.kaggle.com/datasets>.