A Comparative Analysis of Classification Techniques for Chronic Kidney Disease

S Kaustubh Rao
Department of CSE
B.M.S. College of Engineering
Bengaluru, India
kaustubhrao1702@gmail.com

S Pooja Iyer
Department of CSE
B.M.S. College of Engineering
Bengaluru, India
spoojaiyer15@gmail.com

Sai Pranav
Department of CSE
B.M.S. College of Engineering
Bengaluru, India
saipranav429@gmail.com

Rohit Mahendrakar
Department of Computer Science and Engineering
B.M.S. College of Engineering
Bengaluru, India
rohitmahendrakar129@gmail.com

Umadevi V
Department of Computer Science and Engineering
B.M.S. College of Engineering
Bengaluru, India
umadevi.cse@bmsce.ac.in

***Abstract*—A significant public health issue, Chronic Kidney Disease (CKD) is detected by a gradual decline in kidney function over a period of three months or more. The main contributing factors for CKD are hypertension and diabetes and CKD has systemic effects on the body. Detection of CKD in the early stages is essential for reducing its effects. In order to diagnose CKD, this study facilitates an evaluation of the diagnostic abilities of the algorithms K-Nearest Neighbours (KNN) classifier, Naive Bayes classifier, Decision Tree classifier and Logistic Regression classifier. The application of sampling methodologies, such as undersampling, oversampling and Synthetic Minority Oversampling Technique (SMOTE), shows a significant advancement in results. To ensure that attribute significance is preserved the study also handles missing data. Decision Tree predicts CKD with 99% accuracy demonstrating the power of its diagnostic abilities. Naive Bayes achieves a solid 95%. KNN shows a more modest 80%, while Logistic Regression shows 93%. The range of results highlights the complexity of predicting CKD and the range of algorithm capabilities.**

I. Introduction

Chronic Kidney Disease (CKD) stands as a widespread and critical global health concern, impacting millions worldwide [1]. This ailment involves a steady decline in kidney function, posing severe health risks and even potential fatality. The effective management of CKD, pivotal for improved patient outcomes and reduced healthcare burdens, hinges upon timely and accurate diagnosis.

In recent years, the realm of medical research has witnessed the potency of machine learning(ML) models, particularly in disease diagnosis and prediction. Among these, classification algorithms have garnered substantial attention due to their ability to categorize data based on distinct attributes. However, to ensure the reliability and effectiveness of these algorithms in CKD diagnosis, a comprehensive and thorough analysis is imperative.

In this work, K-Nearest Neighbors (KNN), Naive Bayes, Decision Trees, and Logistic Regression—four well-known classification methods for early CKD identification [2]—are meticulously compared and contrasted. The experiment is expanded to look at how sampling techniques and imputation approaches affect categorization accuracy. The basis for this thorough analysis is a dataset from the Machine Learning Repository of University of California, Irvine(UCI) [3]. This dataset includes vital data, such as medical history and test results from the laboratory , needed for an accurate diagnosis of CKD.

Different accuracy levels are revealed by the results of individual applications of each classification technique: KNN (56%), Naive Bayes (96%), Decision Tree (99%), and Logistic Regression (88%). With the help of a variety of sampling approaches, including undersampling, oversampling, and the Synthetic Minority Over Sampling Technique (SMOTE) for each algorithm, the work recognises and resolves the inherent class imbalance typically identified in CKD datasets. Additionally, missing values are handled by statistical methods like mean, median and mode provide understanding of their impact on classification accuracy.

Medical practitioners and researchers can make informed decisions regarding optimal classification strategies for detection of Chronic Kidney Disease (CKD), benefiting from the valuable insights derived from this comprehensive analysis. This research work deepens the understanding of classification algorithms customized to CKD diagnosis by examining dataset features and handling of missing data.

II. LITERARTURE SURVEY

Chronic Kidney Disease (CKD) has become a serious global health issue, necessitating the urgent requirement for detection in the early stages to lessen its severe negative effects on patient well-being. Conventional diagnostic techniques frequently lead to invasive and postponed diagnoses, emphasizing the importance of early detection. Researchers hope to enhance patient outcomes and lower the incidence of mortality and morbidity rates by utilizing machine learning (ML) approaches in early CKD prediction [4]. It is important to detect CKD in the early stages highlights the necessity of preventing CKD from advancing into end-stage kidney disease (ESKD), which calls for life-saving therapies like dialysis or transplantation [5]. Because CKD ususally develops without major symptoms, its complex nature and a variety of linked risk factors make early identification difficult [6].

Many ML models have been used to create predictive models to detect of CKD [7], including artificial neural networks, C5.0, the Chi-square Automatic Interaction Detector, logistic regression, linear support vector machines (SVM) with penalties L1 and L2, random forests, and more [4] [6] [8]. Using feature selection approaches such analysis of variance, recursive feature elimination, and predictive analytics has improved the performance [4] [5].

These efforts have resulted in notable improvements in diagnostic precision. The various maximum attainable accuracies among investigations are particularly remarkable, with simulated results averaging 99.83%, 99.75%, 99.1%, and 99.6% [6] [9]. This highlights how ML algorithms and predictive analytics have the power to drastically improve the accuracy of early CKD identification.

In conclusion, combining ML methods with predictive analytics offers a viable approach to the problem of early CKD detection. Accurate CKD prediction assumes essential relevance by reducing health problems, improving patient well-being, and lessening the burden of end-stage therapies, as supported by the collective findings reported in the referred studies [10] [11] [12].

III. CLASSIFICATION TECHNIQUES

We employed various predictive models, including K-Nearest Neighbors, Decision Tree, Logistic Regression and Naive Bayes classifier, to forecast the occurrence of CKD.

A. K-Nearest Neighbours (KNN)

A non-parametric machine learning method that is used for both classification and regression applications [13]. It classifies new instances according to how similar they are to the k nearest neighbors in the training set. A key element in KNN is the separation of the instances. The Manhattan distance and the Euclidean distance are two regularly used distance measures. On the basis of feature values of the instances, these metrics quantify the differences or similarities between them. A favorable measurement for continuous data is the Euclidean distance formula, which determines the straight-line distance between two locations in a feature space. The results from KNN is strongly influenced by selecting 'k' and the distance metric that are chosen.

Distance Metric: In order to check the similarity between two instances, the KNN classifier computes their distance from one another. The problem and type of data will determine which distance metric is used. Euclidean distance and Manhattan distance are the two most oftenly used distance metrics. Here is an illustration of the Euclidean distance formula. Euclidean Distance formula:

$$EuclideanDistance = sqrt((x_2 - x_1)^2 + (y_2 - y_1)^2)$$

Here, (x_1, y_1) and (x_2, y_2) represent the coordinates of two instances in a feature space.

B. Decision Tree (DT)

A non-parametric method that employs a structure like a tree to make decisions [14]. For predicting the target class, it divides the data into subsets on the basis of features and their values [15]. Gini index and entropy are two impurity metrics that are used in decision trees to assess how successfully a node separates various classes. Gini index calculates the likelihood that a randomly selected instance at a node would be misclassified, whereas Entropy assesses the disorder or ambiguity of the class distribution. Recursively choosing the features and thresholds that best reduce impurity is how the tree is constructed.

- Gini Index: The Gini index determines a node's impurity by calculating the probability of a randomly chosen instance being misclassified if it were randomly labelled according to the class distribution at that node. The Gini Index is calculated as follows:

  $$GiniIndex = 1 - (p_1^2 + p_2^2 + ... + p_n^2)$$

  where $p_1, p_2, ..., p_n$ are the probabilities of each class at the node.
- Entropy: Entropy is another impurity measure that determines the uncertainty or disorder at a node. It measures the average amount of information needed to find out the class label of a given instance drawn from the node's distribution. Entropy can be calculated as follows:

  $$Entropy = -(p_1 * log_2(p_1) + p_2 * log_2(p_2) + ... + p_n * log_2(p_n))$$

  where $p_1, p_2, ..., p_n$ are the probabilities of each class at the node.

C. Logistic Regression (LR)

A popular classification approach that associates a set of input variables (features) and a binary result. It determines the likelihood that a particular instance belongs to a specific class. The logistic function, also referred to as the sigmoid function, is used in LR to translate a linear combination of feature values to a probability varies from 0 to 1 [16]. The

linear output is converted into a probability estimate using the sigmoid function. When it's necessary to understand how specific factors affect the likelihood of the positive class. The probability of positive class(CKD) can be defined as:

$$P(class = 1) = sigmoid(Z)$$

The sigmoid function is defined as:

$$sigmoid(Z) = 1/(1 + e^{-Z})$$

where Z is the linear combination of inputs and weights

$$Z = w_1x_1 + w_2x_2 + \ldots + w_nx_n + b$$

where,

- $x_1, x_2, \ldots, x_n$ are the input features,
- $w_1, w_2, \ldots, w_n$ are the corresponding weights for the features,
- b is the bias term.

Result of the sigmoid function represents the estimated probability that the instance belongs to class 1 (or the positive class). It could be interpreted as the likelihood of the instance being in the positive class.

D. Naive Bayes (NB)

Based on Bayes' theorem, the Naive Bayes classifier is a probabilistic method. It is frequently employed in tasks involving Natural Language Processing(NLP) and text classification. On the basis of observed attributes, the classifier determines the likelihood that a given instance belongs to a specific class [17]. The Bayes theorem, which links the conditional probability of the class given features to the likelihood of observing the features given the class, serves as the theoretical underpinning of this method. Calculations are made simpler by considering that there is no dependency amongst features. Here is a the explanation of the mathematical formula behind the NB classifier:
Bayes' Theorem: The Naive Bayes classifier is built on Bayes' theorem, which describes the relationship between conditional probabilities. Bayes' theorem can be stated as follows:

$$P(\text{class}|\text{features}) = \frac{P(\text{features}|\text{class}) \cdot P(\text{class})}{P(\text{features})}$$

Here, $P(class|features)$ is the posterior lokelihood of the class given the features, $P(features|class)$ is the likelihood of observing the features given the class, P(class) is the prior likelihood of the class, and P(features) is the probability of observing the features.

IV. METHODOLOGIES

The proposed system for CKD classification is shown in Fig. 1

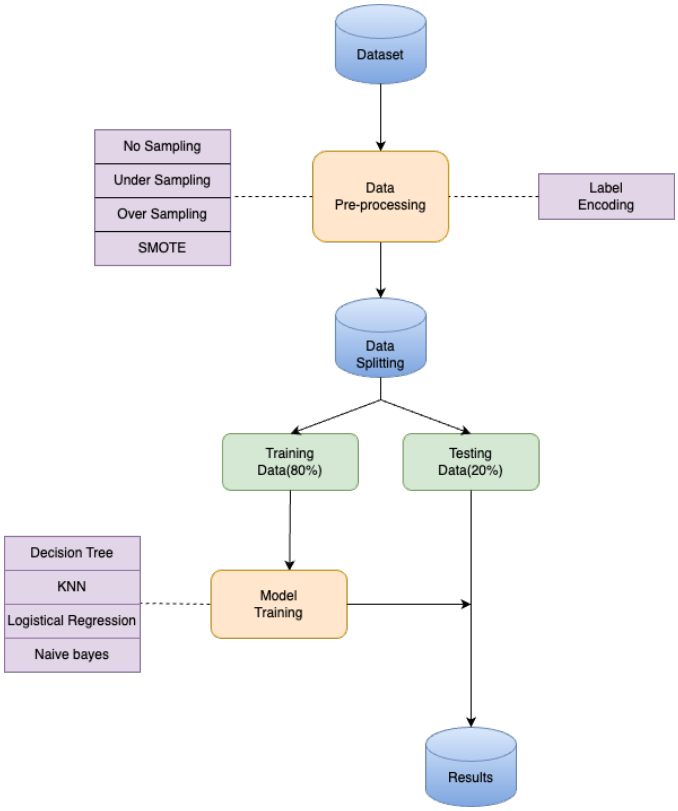

Fig. 1: The system proposed for CKD classification

A. Dataset Description

The CKD dataset [3], available at the Machine Learning Repository, UCI, comprises 400 patient records with 25 features, including both numerical and categorical variables. The objective is to forecast the existence or absence of chronic renal disease using the given characteristics which are as follows.
Blood pressure, age, albumin, specific gravity, sugar, pus cells, rbc, bacteria, pus cell clumps, blood urea, sodium, serum creatinine, potassium, hemoglobin, packed cell value, wbc count, rbc count, diabetes mellitus, hypertension, coronary artery disease, pedal edema, appetite and anemia are among the numerical characteristics from the dataset. Hypertension, diabetes mellitus, coronary artery disease, pedal edema, appetite, and anemia are the categorical variables in the dataset.

B. Data Preprocessing

To maintain data integrity and make it compatible with classification algorithms, data preparation is essential. Several strategies are used in the process to modify, clean, and improve the data, which eventually helps to increase algorithmic performance and produce correct classification results. Several critical preprocessing stages are carried out to diagnose CKD in order to get the dataset ready for efficient classification.

- Label Encoding: Label Encoding for Categorical Variables is the first phase to convert turning category variables into numerical representations. By facilitating the computations and comparisons required for classification, this conversion makes it easier to incorporate categorical data into the algorithmic framework.
- Imputation for Handling Missing Values: Missing values are a typical occurrence in real-world datasets, and they could affect the precision of classification algorithms. Imputation is utilized to handle missing values. To ensure a complete and reliable dataset for analysis, methods like median, mean and mode imputation are used to fill in missing values in numerical features.
- Using Sampling Techniques to Handle Class Imbalance: In medical datasets like CKD, class imbalance—where one class greatly dominates the other is a frequent problem. Under sampling, over sampling, and SMOTE are three sampling techniques that are implemented to reduce class imbalance and boost classification performance.
  - Under Sampling (US): By fewer members of the majority class, Under Sampling seeks to correct the imbalance between classes. Algorithms that are sensitive to class distribution benefit from this rebalancing, which improves their capacity to identify patterns in the minority class.
  - Over Sampling (OS): To balance the distribution of classes, over sampling creates artificial examples of the minority class. Oversampling gives algorithms a more thorough understanding of the features of the minority class by giving the minority class additional training instances
  - SMOTE (Synthetic Minority Over-sampling Technique): SMOTE corrects for class imbalance by interpolating between existing instances to create synthetic instances of the minority class. This method avoids overfitting, which can happen when minority class instances are simply duplicated.

C. Data Splitting for Training the Model

The splitting of testing and training is done in 20-80 ratio respectively followed by preprocessing. The classification model is trained using the bigger training set, which enables it to discover patterns and connections in the data. The testing set, which includes unobserved data, is then used to evaluate the model's performance. This evaluation provides the model's capacity for accurate prediction and generalization to new circumstances.

To ensure the quality and applicability of the dataset for classification tasks, careful data preprocessing is necessary. Techniques like class imbalance addressing through sampling approaches, statistical imputation, label encoding all contribute to an optimized dataset that improves the effectiveness of classification algorithms in the diagnosis of CKD.

Classifer	Metric	Imputation method			
		Mean			
		No sampling	US	OS	SMOTE
Decision Tree	Accuracy	0.95	0.95	**0.98**	0.96
	Recall	0.93	0.90	**0.98**	0.96
	F1 Score	0.93	0.95	**0.96**	0.97
	Precision	0.93	0.91	**0.98**	0.98
	AUC score	0.95	0.95	**0.96**	0.97
KNN	Accuracy	0.62	0.69	0.69	0.75
	Recall	0.68	0.80	0.78	0.88
	F1 Score	0.56	0.73	0.72	0.78
	Precision	0.47	0.63	0.63	0.63
	AUC score	0.72	0.79	0.76	0.79
Logistic Regression	Accuracy	0.86	0.93	0.87	0.91
	Recall	0.93	0.93	0.93	0.98
	Precision	0.83	0.88	0.87	0.87
	F1 Score	0.74	0.93	0.88	0.92
	AUC score	0.92	0.99	0.91	0.94
Naïve Bayes	Accuracy	0.93	0.95	0.95	0.94
	Recall	0.93	0.90	0.96	0.92
	F1 Score	0.91	0.92	0.95	0.94
	Precision	0.90	0.97	0.98	0.98
	AUC score	0.99	0.98	0.97	0.97

TABLE I: Results of Classification by mean imputation method

V. RESULTS

The result of this work on CKD predictive modeling are convincing and shed insight on the complex interactions between imputation methods, sample techniques, and classification algorithms.

Under the mean imputation, the DT showed a high accuracy rate of 98% as shown in Table I. This resulted from the use over sampling and mean filling.

The accuracy of the DT model increased to an amazing 99% by adding both mode imputation and oversampling approaches, as shown in Table III.

On the basis of the outcomes shown in Table II, using median imputation along with SMOTE produced a considerable improvement in accuracy, showing 97% success rate.

The decision tree algorithm performed better when compared to other classification techniques. Mode filling and oversampling together allowed to attain a remarkable 99% accuracy as shown in Table III. Decision tree can therefore be quite effective for identifying CKD trends in data, which can help in better diagnosis.

From Table II the NB algorithm performed well, achieving a respectable 95% accuracy. This occurred when over sampling and median filling was utilized.

Nevertheless, the KNN algorithm's accuracy was just 80% Table II.

Classifier	Metric	Imputation method			
		Median			
		No sampling	US	OS	SMOTE
Decision Tree	Accuracy	0.96	0.98	0.96	**0.97**
	Recall	0.93	0.90	0.98	**0.98**
	F1 Score	0.95	0.93	0.96	**0.98**
	Precision	0.96	0.98	0.94	**0.92**
	AUC score	0.95	0.93	0.96	**0.98**
KNN	Accuracy	0.72	0.70	0.69	0.80
	Recall	0.75	0.80	0.78	0.96
	F1 Score	0.66	0.73	0.72	0.83
	Precision	0.58	0.63	0.67	0.61
	AUC score	0.76	0.79	0.76	0.80
Logistic Regression	Accuracy	0.85	0.93	0.87	0.84
	Recall	0.89	0.93	0.96	0.92
	Precision	0.81	0.88	0.88	0.89
	F1 Score	0.74	0.93	0.88	0.85
	AUC score	0.89	0.99	0.91	0.91
Naïve Bayes	Accuracy	0.92	0.91	0.95	0.93
	Recall	0.96	0.90	0.96	0.92
	F1 Score	0.90	0.92	0.95	0.93
	Precision	0.84	0.97	0.92	0.92
	AUC score	0.99	0.98	0.97	0.97

TABLE II: Results of Classification by median imputation method

Classifer	Metric	Imputation methods			
		Mode			
		No sampling	US	OS	SMOTE
Decision Tree	Accuracy	0.93	0.98	**0.99**	0.96
	Recall	0.89	0.9	**1.00**	0.98
	F1 Score	0.91	0.95	**0.99**	0.96
	Precision	0.93	0.96	**0.98**	0.94
	AUC score	0.95	0.95	**0.96**	0.96
KNN	Accuracy	0.63	0.70	0.63	0.79
	Recall	0.68	0.80	0.63	0.96
	F1 Score	0.57	0.73	0.57	0.82
	Precision	0.49	0.56	0.65	0.64
	AUC score	0.71	0.79	0.73	0.80
Logistic Regression	Accuracy	0.86	0.93	0.87	0.88
	Recall	0.93	0.93	0.96	0.98
	Precision	0.73	0.85	0.88	0.88
	F1 Score	0.96	0.93	0.88	0.89
	AUC score	0.92	0.99	0.91	0.94
Naïve Bayes	Accuracy	0.90	0.91	0.95	0.95
	Recall	0.96	0.90	0.96	0.96
	F1 Score	0.87	0.92	0.95	0.95
	Precision	0.79	0.94	0.96	0.96
	AUC score	0.98	0.98	0.97	0.97

TABLE III: Results of Classification by mode imputation method

In conclusion, this study emphasizes how critical technique selection is for making precise predictions. Doctors and data experts can both profit from this understanding. Making better decisions will allow to anticipate CKD more accurately, which will help to treat patients more quickly and with better outcomes. This research demonstrates how effective data analysis can enhance patient care as healthcare technology develops.

The accuracy obtained for four different classification algorithms with different sampling techniques are depicted in Fig. 2. The results demonstrate that Decision Tree performed well when compared to other three classification techniques.

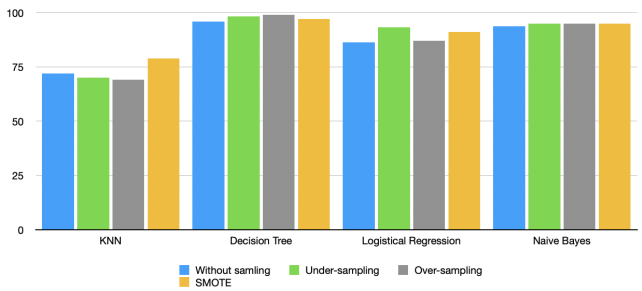

Fig. 2: Statistical Representation of Accuracy

VI. CONCLUSION

The thorough comparison of the classification models for CKD diagnosis emphasizes the importance of early and precise detection in reducing this global health burden. This work exhibit varied levels of accuracy using K-NN, Naive Bayes, DT, and Logistic Regression. DT emerges as the most efficient classifier, closely followed by Naive Bayes, Logistic Regression, and KNN. This research also explores the impact of sampling techniques and imputation methods, demonstrating the critical part they play in enhancing classification performance. While imputation methods (Mean, Mode, and Median) are essential in handling missing data, SMOTE emerges as a potential technique for addressing class imbalance. The selection of relevant classification methods for the diagnosis of CKD is made simpler by the findings proposed in this work, which also takes dataset features and missing data into account. Future research will examine complex algorithms to resolve imbalanced data, and improve the diagnosis and management of CKD.

REFERENCES

[1] A. L. Ammirati, "Chronic kidney disease," *Revista da Associação Médica Brasileira*, vol. 66, pp. s03–s09, 2020.

[2] Y. Amirgaliyev, S. Shamiluulu, and A. Serek, "Analysis of chronic kidney disease dataset by applying machine learning methods," pp. 1–4, 2018.

[3] U. dataset repository, "Chronic kidney disease dataset," 2023, accessed on July 2023. [Online]. Available: https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease

[4] P. Chittora, S. Chaurasia, P. Chakrabarti, G. Kumawat, T. Chakrabarti, Z. Leonowicz, M. Jasiński, Ł. Jasiński, R. Gono, E. Jasińska *et al.*, "Prediction of chronic kidney disease-a machine learning perspective," *IEEE Access*, vol. 9, pp. 17 312–17 334, 2021.

[5] A. J. Aljaaf, D. Al-Jumeily, H. M. Haglan, M. Alloghani, T. Baker, A. J. Hussain, and J. Mustafina, "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," pp. 1–9, 2018.

[6] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A machine learning methodology for diagnosing chronic kidney disease," *IEEE Access*, vol. 8, pp. 20 991–21 002, 2019.

[7] K. Swathi and G. Vamsi Krishna, "Prediction of chronic kidney disease with various machine learning techniques: A comparative study," pp. 257–262, 2023.

[8] D. Baidya, U. Umaima, M. N. Islam, F. J. M. Shamrat, A. Pramanik, and M. S. Rahman, "A deep prediction of chronic kidney disease by employing machine learning method," pp. 1305–1310, 2022.

[9] W. Gunarathne, K. Perera, and K. Kahandawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (ckd)," pp. 291–296, 2017.

[10] Q. Bai, C. Su, W. Tang, and Y. Li, "Machine learning to predict end stage kidney disease in chronic kidney disease," *Scientific reports*, vol. 12, no. 1, p. 8377, 2022.

[11] D. A. Debal and T. M. Sitote, "Chronic kidney disease prediction using machine learning techniques," *Journal of Big Data*, vol. 9, no. 1, pp. 1–19, 2022.

[12] I. U. Ekanayake and D. Herath, "Chronic kidney disease prediction using machine learning methods," pp. 260–265, 2020.

[13] O. Kramer and O. Kramer, "K-nearest neighbors," *Dimensionality reduction with unsupervised nearest neighbors*, pp. 13–23, 2013.

[14] I. Pasadana, D. Hartama, M. Zarlis, A. Sianipar, A. Munandar, S. Baeha, and A. Alam, "Chronic kidney disease prediction by using different decision tree techniques," vol. 1255, no. 1, p. 012024, 2019.

[15] H. Ilyas, S. Ali, M. Ponum, O. Hasan, M. T. Mahmood, M. Iftikhar, and M. H. Malik, "Chronic kidney disease diagnosis using decision tree algorithms," *BMC nephrology*, vol. 22, no. 1, pp. 1–11, 2021.

[16] S. Menard, *Applied logistic regression analysis*. Sage, 2002, no. 106.

[17] D. Berrar, "Bayes' theorem and naive bayes classifier," *Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics*, vol. 403, p. 412, 2018.