Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 3 (2024)

Discretize-Based Technique and Hybrid Machine Learning Approach for Medical Data Analysis and Mining

Benjamin Asubam Weyori[1], Solomon Antwi Buabeng[2], Lois Azupwah[3], Ben Beklisi Kwame Ayawli[4]

[1]*Department of Computer and Electrical Engineering, University of Energy and Natural Resources, Sunyani, Ghana*

[2]*Department of Computer Science and Informatics, University of Energy and Natural Resources Sunyani, Ghana*

[3]*University Clinic, University of Energy and Natural Resources, Sunyani, Ghana*

[4]*Department of Computer Science, Sunyani Technical University, Sunyani, Ghana*

Abstract

With the extensive growth of data mining, its applications have expanded significantly, encompassing various fields, including healthcare. Despite the abundance of medical data, healthcare providers sometimes rely on personal observations rather than data-driven insights. To address this gap, researchers and medical professionals employ data mining techniques, notably machine learning algorithms, for health data analysis. In this study, three machine learning algorithms, namely K-nearest Neighbor, Random Forest, and Naive Bayes, are employed. Additionally, an ensemble majority voting approach is used in conjunction with a discretization data preprocessing technique, providing flexibility in model selection. Multiple datasets related to Heart disease, Breast tissue, Breast cancer, and Cryotherapy are utilized, offering a diverse range of data for analysis. Among the models tested, the discretized majority voting approach outperforms other classifiers and established state-of-the-art models in the literature. It achieves accuracy rates of 91.9%, 79.7%, 77.1%, and 88.9% for the Heart disease, Breast tissue, Breast cancer, and Cryotherapy datasets, respectively. This proposed methodology presents an effective and suitable learning approach for intelligent healthcare classification systems, especially when precision and model robustness are paramount.

Keywords: Machine Learning (K-Nearest Neighbor; Random Forest; Naive Bayes; Majority Voting); Discretization and Medical Data Mining

Introduction

Data mining has brought about a revolution in various sectors, spanning health [1], education [2], manufacturing [2] [3], fraud detection [4], surveillance [5], economics [6] [7], and criminal investigation [8]. In the realm of healthcare, data mining has emerged as a powerful tool for extracting valuable insights and knowledge from extensive clinical and medical databases, leading to the development of distinct terminologies such as healthcare data mining [9], medical data mining [10], and clinical data mining [11]. The abundance of medical records has generated a wealth of data that holds the potential to enhance patient care and inform evidence-based decision-making [12].

The application of data mining in the healthcare domain plays a pivotal role in converting unstructured health data into actionable information. By employing data mining techniques and algorithms on comprehensive patient data, healthcare professionals and researchers can uncover hidden patterns, correlations, and trends that might elude

traditional analytical methods [13]. Data mining facilitates the identification of risk factors, disease patterns, treatment responses, and other critical insights, thereby significantly influencing clinical practices, policy formulation, and personalized treatment strategies. It encompasses diverse data sources, including electronic health records (EHRs), medical imaging, genetics, wearable technologies, and clinical trial data, providing a comprehensive understanding of a patient's health status and medical history [14].

The utilization of data mining in healthcare offers substantial advantages, such as the detection of disease patterns [15] and the development of predictive models for early diagnosis [16]. This empowers healthcare professionals to identify individuals at risk and implement preventive measures to mitigate the spread of diseases. Moreover, data mining plays a crucial role in the realm of personalized medicine, enabling physicians to tailor treatment regimens to each patient's unique requirements [12]. Furthermore, it has the potential to reduce the number of referrals to medical centers, thereby alleviating the burden on outpatient services and contributing to more efficient healthcare management**.**

Despite the wealth of healthcare data and available data mining techniques, some medical professionals still heavily rely on their firsthand observations and clinical expertise when making patient assessments. Additionally, it's important to recognize that healthcare data mining encounters its fair share of challenges and limitations. Dealing with complex and extensive healthcare datasets involves daunting tasks such as data integration [17], standardization [18], and ensuring data quality, all of which present significant hurdles [19]. Effective data preparation methods become paramount to guarantee trustworthy outcomes. Moreover, the selection and customization of techniques to address specific clinical research inquiries while taking into account unique data characteristics like high dimensionality, class imbalance, and missing values become pivotal considerations.

In light of these challenges, this research endeavors to establish a comprehensive data mining approach by employing data preprocessing techniques and machine learning algorithms on medical datasets. The machine learning classifiers employed in this study encompass random forest (RF), naïve Bayes (NB), and k-nearest neighbor (KNN), all applied to multiple medical datasets. Furthermore, the study adopts an ensemble approach that combines these three algorithms using a majority voting (MJV) scheme, allowing for the flexibility of incorporating or excluding a discretization (D) data preprocessing technique. It's noteworthy that only a limited number of publications have explored the amalgamation of discretization and voting Classifier approaches across multiple medical datasets.

The subsequent sections of this paper are organized as follows: The following section delves into the relevant literature pertaining to heart disease prediction datasets. The "Methodology" section provides a concise overview of the models employed for result comparison and outlines the approach adopted for predicting heart disease. Detailed findings from the models are elucidated in the "Results and Discussion" section. Finally, the "Conclusion and Future Works" section offers a summary of the conclusions drawn and underscores the significance of employing voting Classifier and discretization techniques in the context of heart disease prediction.

Literature Review

Varun et al. [20] developed an Efficient Heart Disease Prediction System utilizing Logistic Regression. They sourced their dataset from Kaggle and achieved an accuracy of 87% with Logistic regression (LR). Other classifiers used in their study included Random Forest (RF) and Naïve Bayes (NB).

Pahwa & Kalra [21] conducted heart disease analysis using a Multi-class Support Vector Machine technique with a dataset from the University of California Irvine (UCI). The classifiers employed were Multi-class SVM, Naïve Bayes, and J48, with Multi-class SVM achieving the highest accuracy at 90.09%.

David & Belcy [22] applied data mining techniques to heart disease prediction using a UCI dataset. They selected Decision Tree, Naïve Bayes, and Random Forest as classifiers, with Random Forest achieving the highest precision at 81%.

Srivenkatesh [23] focused on cardiovascular disease (CVD) prediction using various Machine Learning (ML) classifiers. Support Vector Machines, K-nearest neighbor, Random Forest, Logistic Regression, and Naive Bayes were employed. Logistic regression emerged as the most effective algorithm, boasting an accuracy of 77.06%.

Ware, Rakesh, & Choudhary [24] conducted research on heart attack prediction using fundamental machine learning techniques. They applied Random Forest, Support Vector Machine, K Nearest Neighbor, Logistic Regression, Decision tree, and Naive Bayes classifiers. Support Vector Machine (SVM) demonstrated the highest prediction accuracy at 89.34%.

Vineeth [25] reported on heart disease prediction using Machine learning algorithms with a dataset from the UCI machine learning repository. The algorithms employed included Logistic Regression, K-Nearest Neighbors, Naïve Bayes, Decision tree, and Random Forest, with Random Forest exhibiting the highest efficiency, achieving an accuracy of 88%.

Kamboj [26] conducted research on heart disease prediction using machine learning approaches. The prediction involved a range of machine learning classifiers such as Support Vector Machine, K-Nearest Neighbor (KNN), Random Forest, Naïve Bayes, Logistic Regression, and Decision Tree. Among these, K-Nearest Neighbor (KNN) achieved the highest prediction rate at 87%. The dataset for this research was sourced from the UCI machine learning repository.

Amen & Mahmoud [27] focused on predicting multi-stage heart diseases using machine learning techniques. Their study incorporated machine learning classifiers like Gradient Tree Boosting, Logistic Regression, Support Vector Machine, Random Forest, and Extra Random Forest. Logistic Regression (LR) emerged as a strong performer with an accuracy of 82% in the experimental analysis.

Anggraini [28] aimed to classify breast cancer utilizing data from routine patient examinations found in the Breast Cancer Coimbra (BCC) dataset, available in the UCI machine learning repository since 2018. The research employed backward elimination modeling, the Naïve Bayes algorithm, and split validation to optimize accuracy and validate the model. The study achieved an accuracy rate of 77.14%, demonstrating its effectiveness in breast cancer classification.

Amin & Parvin [29] concentrated on enhancing the performance of the K-NN classification algorithm by using the Gain Ratio to select and reduce dataset attributes. They utilized the Breast Cancer Coimbra and Hepatitis C Virus datasets from the UCI Machine Learning Repository. The results indicated that applying the Gain Ratio improved K-NN's performance in both datasets. Specifically, the Breast Cancer Coimbra dataset achieved an accuracy of 72.85%, while the Hepatitis C Virus dataset demonstrated improved performance with an accuracy of 86.25%.

Chumuang et al. [30] delved into the classification of electrical impedance in breast tissue to analyze breast cancer, employing Bootstrap Aggregating techniques. Their approach involved a comprehensive toolkit and a specialized understanding of pathological changes for in-situ breast cancer detection. The algorithm was rigorously assessed using a UCI dataset specifically focused on breast tissue impedance. Through 10-fold cross-validation involving 106 objects, it achieved an accuracy rate of 74.47%.

Arbawa et al. [31] aimed to classify breast tissue lesions using the Learning Vector Quantization (LVQ) method in conjunction with the Genetic Algorithm (GA) to optimize results. Despite a limited dataset comprising 106 data points across 6 classes, the study revealed that the combination of LVQ with GA yielded a significantly improved accuracy rate of 73%, representing a roughly 50% increase compared to LVQ alone.

Guimarães et al. [32] delved into predicting treatment susceptibility for Human Papillomavirus (HPV) using cryotherapy and immunotherapy. They introduced a hybrid model, the fuzzy neural network (FNN), capable of effectively analyzing complex medical research data. By applying F-score techniques to prune the FNN, the study achieved elevated accuracy rates—84.32% for immunotherapy and 88.64% for cryotherapy—outperforming other

models. This research underscores the potential of combining neural networks and fuzzy systems as valuable tools for predicting HPV treatment outcomes in cryotherapy and immunotherapy.

Rahayu et al. [33] scrutinized the effectiveness of cryotherapy treatment and used a dataset as a benchmark for its evaluation. Their approach employed Neural Network machine learning with specific parameters: 500 training cycles, a learning rate of 0.03, and a momentum of 0.9. This resulted in highly accurate classification, boasting an accuracy rate of 87.78% and an AUC value of 0.955.

In addition to these medical applications, intelligent methods have found extensive use in classification across diverse domains, including energy, finance, agriculture, and transportation. Table 1 provides a summary of recent papers that leverage artificial intelligence for medical classification.

Table 1: Artificial intelligence-based medical classification papers recently published.

Authors	Year	Classifier(s)	Domain	Outcome
Wan et al.	2018	Deep MLP	Parkinson's Disease diagnosis	Analyzing patients' speech and movement patterns to predict levels of Parkinson's Disease.
Alickovic & Subasi	2017	GA feature selection and rotation forest	Breast cancer diagnosis	Using feature extraction and individual or multiple classifiers to propose a diagnostic system that reduces computational complexity and speeds up classification tasks.
Wang et al.	2018	SVM-based ensemble learning algorithm	Breast cancer diagnosis	A weighted receiver operating characteristic curve model is proposed
Saqlain et al.	2017	Probabilistic principal component analysis (PPCA), SVM and RBF	Heart diseases diagnosis	Using PPCA to extract feature subsets and RBF kernel-based SVM to classify them for automatic diagnosis.
Abdar et al.	2019	PSO, SVM and GA	Coronary artery	Development of a new genetic

				training classification model based on feature selection
Javeed et al.	2019	Random search algorithm (RSA), Random Forest model	Heart disease diagnosis	Development of a diagnostic system using RSA for feature selection and a random forest model.
Chen et al.	2020	SVM and GA	Human papillomavirus screening	Analysing cervical secretions using multivariate statistical methods and Raman spectroscopy.
Mahendran et al.	2020	Random Forest, SVM and MLP	Major depressive diagnosis	Developing a stacked generalization model based on machine learning classifiers.
Zhang et al.	2022	MLP, Variational autoencoder	Autism spectrum disorder diagnosis	Proposing a filter feature-selection method and developing a deep learning model with simplified VAE
Chawla et al	2023	Flexible analytic wavelet transforms, KNN	Parkinson's disease detection	Proposing a diagnostic methodology based on flexible analytic wavelet transform

In summary, this paper makes two significant contributions:

- Introduction of a novel learning method centered around majority voting for medical diagnosis.
- Evaluation of the performance of this discrete learning-based technique on four widely recognized medical datasets, with a comparative analysis against existing state-of-the-art methods in the field.

Methodology

This research was conducted to evaluate the effectiveness of employing discretization and voting techniques, as detailed in Figure 1, which outlines the entire research process. The initial step involved the collection of data from credible sources, specifically UCI and Kaggle, to build a diverse dataset. Following this, rigorous preprocessing techniques were implemented, encompassing tasks such as data splitting and transformation. These preprocessing steps were crucial to ensure that the data was in an appropriate format for subsequent modeling.

Once the data was suitably prepared, machine learning algorithms were applied to this preprocessed dataset. To enhance the modeling process and leverage the strength of multiple algorithms, a majority voting approach was employed. This involved combining the outputs of these machine learning models to create an ensemble model, thereby increasing the robustness and reliability of the final predictive model.

In the concluding phase of the study, the performance and accuracy of these models were rigorously assessed. This evaluation served as a critical benchmark to gauge the effectiveness of the discretization and voting techniques in enhancing the overall predictive capability of the models.

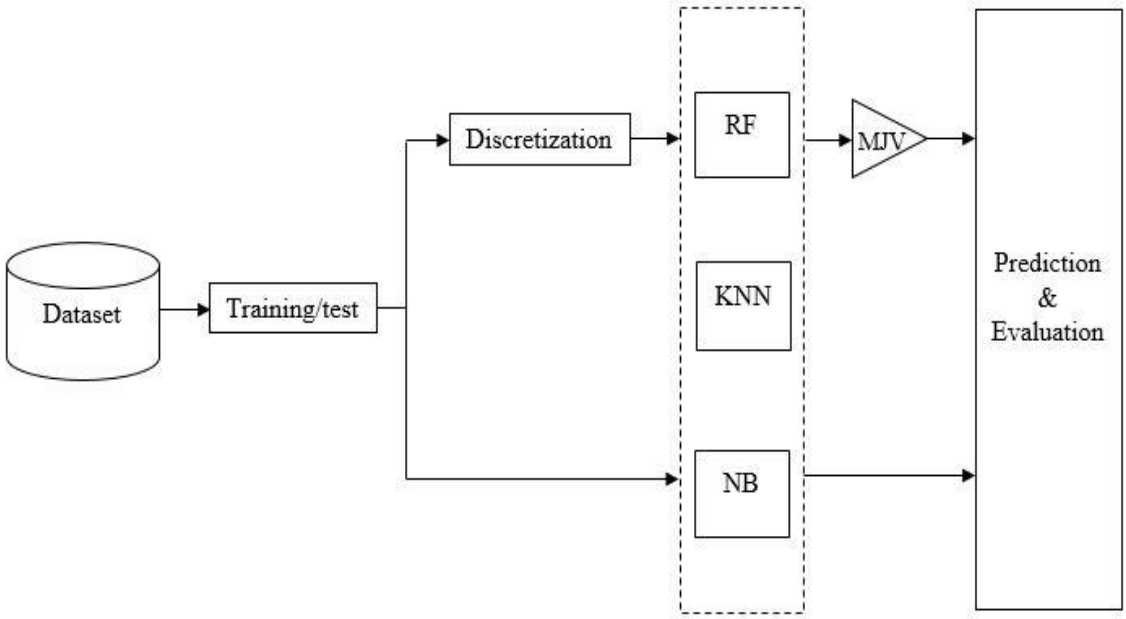

Figure 1: Workflow of experimental design

Dataset Used for the Study

The study made use of four distinct datasets, each serving specific research purposes. A summarized dataset overview is presented in Table 2, which includes crucial information such as feature names, instance counts, and attribute details. The datasets employed in this study encompassed the following:

- Statlog - Cleveland - Hungary - Long Beach – Switzerland (SCHULOS) heart disease dataset [34]: This dataset, derived from a combination of five datasets within the UCI Machine Learning repository, was a primary component of the research.
- Breast tissue (BT) dataset [30]: This dataset, which was also utilized in the study, contains valuable information relevant to breast tissue analysis.
- Breast cancer Coimbra (BCC) dataset [35]: This dataset provided essential data for breast cancer-related investigations, contributing to the comprehensive analysis.
- Cryotherapy (CRY) dataset [36]: The Cryotherapy dataset was another integral component used in the research, specifically tailored for the study's objectives.

Table 2 furnishes an overview of these datasets, offering insights into the features, their respective names, as well as the dimensions represented by rows (R) and columns (C).

Table 1: depicting the total number of records in the dataset

Dataset	Features	Column / Row

SCHULOS	1. Age, 2. Sex, 3. Chest pain type, 4. Resting blood pressure, 6. Cholesterol Serum, 7. Fasting blood sugar, 8. Resting electrocardiographic results, 9. Maximum heart rate achieved, 10. Exercise-induced angina, 11. Old peak ST depression induced by exercise relative to rest, 12. The slope of the peak exercise ST segment, 13. Target	C =12 R = 1190
BT	1. Impedivity (ohm) at zero frequency, 2. Phase angle at 500 KHz, 3. The high-frequency slope of phase angle, 4. Impedance distance between spectral ends, 5. The area under spectrum, area normalised by DA, 6. Maximum of the spectrum, the distance between I0 and real part of the maximum frequency point, 7. Length of the spectral curve	C = 10 R = 106
BCC	1. Age, 2. BMI, 3. Glucose, 4. Insulin, 5. Homeostatic Model Assessment (HOMA), 6. Leptin, 7. Adiponectin, 8. Resistin, 9. MCP.1, 10. Classification	C = 10 R =116
CRY	1. Response to treatment, 2. Gender, 3. Age (year), 4. Time elapsed before treatment (month), 5. The number of warts, 6. Types of wart (Count), 7. The surface area of the wartsc (mm2)	C = 7 R = 90

Data Splitting

In the realm of machine learning, it is a standard practice to divide a dataset into two distinct parts: the training set and the test set. The training set serves the purpose of instructing and fine-tuning the model, while the test set evaluates the model's performance on data it hasn't encountered during training. This partitioning is crucial for assessing how well the model can generalize its learnings and make predictions on new, unseen data. To ensure impartial results, the dataset is methodically shuffled and randomly separated, ensuring an even distribution of instances between the two sets. Typically, about 70% of the data is allocated to the training set, leaving the remaining 30% for testing.

Discretization

Discretization, within the domain of machine learning, is a technique employed to convert continuous variables into discrete ones [37]. It involves breaking down the range of a continuous variable into intervals or bins and then assigning values based on where they fall within these intervals. Various methods can be used for discretization, including equal-width or equal-frequency binning [38].

To illustrate, let's consider a continuous variable X with values $x_1, x_2, x_3, \ldots, x_n$. The discretization process entails dividing the range of X into m intervals or bins, denoted as $[a_1, b_1]$, $[a_2, b_2], \ldots [a_m, b_m]$.. Subsequently, the values of X are substituted with discrete values corresponding to the interval they belong to. This transformation simplifies the data representation and proves beneficial when employing machine learning algorithms that necessitate categorical or ordinal inputs [39]. However, it's important to bear in mind that discretization does result in the loss of information that was initially present in the continuous variables [40].

Machine Learning Models

Within the realm of healthcare, numerous machine learning methods and strategies have demonstrated their potential to revolutionize the industry. This particular study harnessed the power of three supervised learning algorithms: K-Nearest Neighbors (KNN), Naive Bayes (NB), and Random Forest (RF).

Naive Bayes

Naive Bayes stands as a renowned machine learning algorithm grounded in Bayes' theorem, operating under the assumption of feature independence [41]. This theorem is employed to calculate the probability of each class label

given a feature vector. The "naive" aspect arises from the presumption that all features are independent, simplifying the mathematical calculations [42]. Mathematically, the probability of a class label y given a feature vector x is represented by equation (1):

$$P(y \mid x) = \left(P(y) * P(x \mid y)\right) / P(x) \quad (1)$$

In this equation, P(y) denotes the prior probability of class y, P(x | y) signifies the likelihood of observing feature vector x given class y, and P(x) represents the probability of encountering feature vector x. To make a prediction, the algorithm computes the probability for each potential class label and selects the one with the highest probability, as outlined in equation (2):

$$y\_pred = \text{argmax}\_y\, P(y \mid x) \quad P(x|y) \quad (2)$$

Here, y_pred represents the predicted class label. Estimations for the probabilities P(y) and P(x | y) can be derived from the training data using techniques such as maximum likelihood estimation or smoothing methods like Laplace smoothing to handle unobserved feature combinations. The independence assumption allows us to calculate P(x | y) as the product of individual probabilities for each feature given class y.

Naive Bayes is recognized for its efficiency and success, making it suitable for tasks like text classification and spam filtering. Nonetheless, its accuracy may be compromised if the assumption of feature independence is violated or if significant correlations exist among the features.

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) stands out as a prominent machine learning algorithm, well-suited for classification and regression tasks [43]. Its predictive power hinges on the similarity between a new instance and its k closest neighbors within the training data [44]. The underlying assumption is that instances sharing similar feature vectors are likely to belong to the same class label [45]. The choice of k plays a pivotal role, in striking a balance between model flexibility and stability [46]. Smaller k values increase flexibility but can be susceptible to noise, whereas larger k values offer more stable predictions at the cost of local accuracy [47].

Various distance measures come into play, with some of the most commonly used ones being:

$$\text{Euclidean}: \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (3)$$

$$\text{Manhattan}: \sum_{i=1}^{k}|x_i - y_i| \quad (4)$$

$$\text{Minkowski}: \left(\sum_{i=1}^{k}\left(|x_i - y_i|\right)^q\right)^{1/q} \quad (5)$$

Here, k denotes the number of dimensions, xi and yi represent data points, and q signifies the order of the norm. Unlike some other algorithms, KNN dispenses with explicit model training, relying solely on the training data during prediction. While this streamlines training, it can lead to slower predictions, particularly when dealing with large datasets. KNN, characterized by its versatility, excels in delineating complex or nonlinear decision

boundaries. However, it can be computationally demanding, especially with sizable datasets or high-dimensional features, as identifying the nearest neighbors can be time-intensive.

Random Forest (RF)

Random Forest emerges as an ensemble learning algorithm that amalgamates multiple decision trees to make predictions [48]. Its applicability spans both classification and regression tasks [49]. The mathematical representation of a Random Forest prediction for a new instance x is articulated as:

$$y\_pred = majority\_vote(T\_1(x), T\_2(x), ..., T\_n(x)) \quad (6)$$

Here, y_pred denotes the predicted class label, T_i(x) signifies the predicted class label from tree i, and majority_vote() assembles the class label with the most substantial votes.

The Random Forest algorithm confronts overfitting by averaging predictions from multiple decision trees and introduces randomness via bootstrap sampling and feature subsampling [50] [51]. It boasts robustness, performs admirably with high-dimensional data, and offers insights into feature importance. However, if the dataset harbors irrelevant or noisy features, the algorithm might incorporate them into the model, potentially compromising its performance.

Majority Voting

The majority voting technique elevates prediction reliability and accuracy by harnessing the collective knowledge of multiple classifiers [52]. It empowers ensembles to make more dependable decisions by considering consensus among individual models. As elucidated by Raza [53], majority voting is extensively employed in classification problems and can serve as a meta-classifier for amalgamating ensemble technique results.

Implementing majority voting involves deriving the final prediction based on individual model predictions. This can be achieved through a straightforward vote tallying mechanism:

1. Initialise counters for each class: Count_A starts at 0 and Count_B starts at 0.
2. For each model prediction Pi (where i ranges from 1 to N):

- If Pi is 0, increment Count_A by 1.
- If Pi is 1, increment Count_B by 1.

3. Determine the majority class based on the counts:

- If Count_A is greater than Count_B, the final prediction is Class A.
- If Count_B is greater than Count_A, the final prediction is Class B.
- If Count_A is equal to Count_B, the final prediction can be randomly selected or subjected to further processing using tie-breaking rules.

In case of a tie, additional rules or random selection can be applied to finalize the prediction.

In the context of majority voting, when individual models exhibit diverse biases, this approach proves invaluable in mitigating those biases [54]. Conversely, if ensemble models display strong correlations, their predictions may lack substantial variability [55]. In such scenarios, employing majority voting may not yield significant accuracy enhancements and could even lead to a decline.

Evaluation Metric Adopted and Used

To facilitate performance comparisons, we have harnessed four key metrics: accuracy, recall, precision, and F-score. These critical evaluation metrics are encapsulated by equations 9, 11, 12, 13, 14, and 15, delineating their precise mathematical representations.

$$\text{Accuracy (ACC)} = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad (7)$$

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

$$\text{Precision (PRE)} = \frac{TP}{(TP+FP)} \quad (9)$$

$$\text{F-Score (FSC)} = \frac{2*(\text{Re}call*\text{Pr}ecision)}{(\text{Re}call+\text{Pr}ecision)} \quad (10)$$

Results and Discussion

Once the prediction models were meticulously crafted, it became imperative to subject them to rigorous testing to ascertain their aptness for the intended predictive tasks. To comprehensively gauge the models' performance, a diverse array of evaluation metrics was judiciously employed. The ensuing results encapsulated in Table 2 and vividly depicted in Figure 2, offer a panoramic view of the models' effectiveness.

The following insights were garnered by rigorously testing the model. Furthermore, we conducted an in-depth exploration of scenarios where preprocessing, notably discretization, was applied. Remarkably, this was done while keeping the foundational estimator consistent across the classifiers, encompassing KNN, RF, NB, and MJV of KNN-RF-NB.

Table 2: Depict the comparison amongst the Models on various Dataset

Dataset	Metrics	KNN	RF	NB	MJV	D-KNN	D-RF	D-NB	D-MJV
SCHULOS	ACC	76.5	90.2	86.6	90.8	89.6	91.3	86.0	91.9
	PRE	75.5	92.7	88.3	92.7	90.7	93.2	87.8	93.4
	REC	81.2	88.2	85.5	89.2	89.2	89.8	84.9	90.9
	FSC	78.2	90.4	86.9	91.0	90.0	91.5	86.3	92.1
BT	ACC	62.5	62.5	62.5	71.9	53.1	56.3	50.0	79.7
	PRE	58.9	57.2	69.4	71.8	59.4	54.2	53.7	80.2
	REC	58.6	59.0	63.8	70.3	47.4	52.2	49.4	80.2
	FSC	58.6	56.4	63.3	68.7	51.8	52.0	47.6	79.8
BCC	ACC	48.6	60.0	48.6	57.1	68.6	71.4	48.6	77.1
	PRE	33.3	42.9	38.5	42.9	53.8	58.3	40.0	62.5
	REC	50.0	50.0	83.3	75.0	58.3	58.3	10.0	83.3

	FSC	40.0	46.2	52.6	54.5	56.0	58.3	57.1	71.4
CRY	ACC	81.5	70.4	85.2	85.2	77.8	88.9	81.4	88.9
	PRE	73.3	61.1	75.0	75.0	66.7	80.0	73.3	80.0
	REC	91.7	91.7	100.0	100.0	100.0	100.0	91.6	100.0
	FSC	81.5	73.3	85.7	85.7	80.0	88.9	81.5	88.9

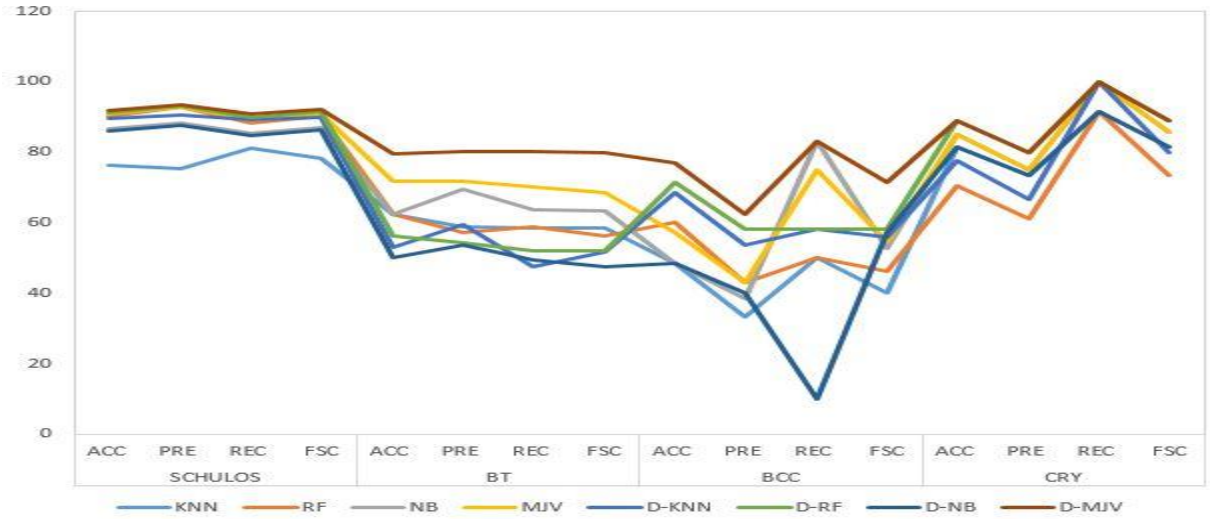

Fig 2: Comparison evaluation performance for the models - subplot

The data presented in Table 2 and Figure 2 offer a compelling insight into the performance of various models, particularly highlighting the superiority of the D-MJV model across multiple evaluation metrics. For instance, when applied to the SCHULOS dataset, the D-MJV model achieved remarkable scores with ACC at 91.9%, PRE at 93.4, REC at 90.9, and FSC at 92.1. Similarly, for the BT dataset, it exhibited impressive performance with ACC at 79.7, PRE at 80.2, REC at 80.2, and FSC at 79.8. The BCC dataset showed ACC at 77.1, PRE at 62.5, REC at 83.3, and FCS at 71.4, while the CRY dataset yielded ACC at 88.9, PRE at 80.0, REC at 100.0, and FCS at 88.9 when the D-MJV model was applied.

These results underscore the effectiveness of the D-MJV model, which leverages both discretization, the process of converting numerical features into discrete variables, and the aggregation of predictions. This combination evidently enhances the overall accuracy and robustness of the model.

However, it's noteworthy that discretization did not uniformly benefit all models. Specifically, NB experienced a negative impact from discretization. For example, when considering the BT dataset, D-NB exhibited a notably lower accuracy value of 50.0% compared to NB's accuracy of 62.5%. This represents a reduction of 12.5% in accuracy due to discretization. Furthermore, for the same BT dataset, all individual algorithms experienced decreased accuracy when subjected to the Discretization dataset.

Additionally, KNN demonstrated the lowest accuracy (76.5%) among the SCHULOS heart dataset models, which can be attributed to its challenges when dealing with large input sizes. In the case of the BCC dataset, both KNN and D-NB exhibited the least accuracy, at 76.5%, largely because these algorithms are sensitive to specific features, and the presence of zeros in the BCC dataset had a detrimental effect on their performance.

Lastly, RF's performance was notably reduced for the CRY dataset, mainly due to its requirement for sufficient data to capture patterns and generalize effectively.

In conclusion, this investigation underscores the significant impact of discretization and majority voting on model performance, with the D-MJV model emerging as the standout performer across various datasets and metrics.

Comparison Analysis

Table 3 provides a comprehensive performance comparison between our research and other existing studies, with a particular emphasis on the accuracy attained by our top-performing model, D-MJV. An in-depth analysis of the statistics presented in Table 3 unmistakably reveals that our proposed model surpasses the performance of the existing models, firmly establishing it as a superior predictor of accuracy.

Table 3: Comparison of the proposed model with other existing hybrid research work

Dataset	Accuracy	Reference
SCHULOS	88.4%	Hybrid (HRFLM) Model [56]
	88.7%	Hybrid model [57]
	91.9%	Propose model D-MJV
BT	74.5%	Bootstrap Aggregating Techniques [30]
	73.0%	LVQ combine with GA [31]
	79.7%	Propose model D-MJV
BCC	72.9%	Gain Ratio on K-NN [29]
	77.1%	Backward elimination modeling - NB algorithm [28]
	77.1%	Propose model D-MJV
CRY	88.6%	Fuzzy neural network (FNN) [32]
	87.9%	Neural Network [33]
	88.9%	Propose model D-MJV

The research analysis results presented in Table 3 underscore a noticeable contrast in prediction rates between existing works and our study. This discrepancy can primarily be attributed to a critical factor: the conversion of the dataset from continuous data to the discrete form. Our research distinguishes itself by demonstrating the effective application of data mining within the medical field, particularly in achieving exceptional classification across multiple datasets.

Our approach leverages the power of dimensionality reduction through discretization. This process simplifies the modeling by diminishing the intricacies associated with continuous variables. Additionally, it enhances the performance of specific algorithms by reducing the influence of extraneous values.

Moreover, our study reveals that the combination of predictions from multiple models yields higher accuracy compared to relying on individual models in isolation. This observation underscores the resilience of ensemble methods to noise and errors within individual models, ensuring more reliable predictions, even when a single model may falter.

Conclusion

This study has a core objective: to enhance the prediction and early detection of various ailments through the application of data mining techniques. Given the vast volume of healthcare data generated, the potential to predict disease occurrences becomes a crucial avenue to explore. Leveraging machine learning algorithms for the analysis of medical records and the identification of patterns can significantly elevate the accuracy of these predictions.

In our research, we worked with multiple datasets, each encompassing distinct aspects of health-related information. Specifically, these datasets comprise the Stat log - Cleveland -Hungary - Long Beach – Switzerland (SCHULOS) heart disease dataset, Breast tissue (BT), Breast cancer Coimbra (BCC), and Cryotherapy (CRY) dataset. Our investigation delves into the analysis of both discretized and non-discretized versions of these datasets.

Discretization, a fundamental concept in machine learning, involves the transformation of continuous features into categorical variables, streamlining the data for analysis. Within this context, we applied three classification algorithms: K-Nearest Neighbors (KNN), Naive Bayes (NB), and Random Forest (RF). To enhance predictive accuracy further, we created a hybrid model that combined all three algorithms. Moreover, this hybrid model offered the flexibility to include or exclude discretization as a preprocessing technique.

Our findings are notable, with the discretized hybrid model (D-MJV) emerging as a standout performer. It achieved impressive accuracy rates of 91.9%, 79.7%, 77.1%, and 88.9% across the aforementioned datasets. This study effectively underscores the pivotal role of discretization in amalgamating multiple models to attain precise predictions. As we look to the future, further research avenues could explore the integration of diverse medical datasets, encompass additional preprocessing techniques such as feature selection, and delve into advanced algorithm combinations, potentially incorporating deep learning. Such endeavors have the potential to significantly enhance the performance of healthcare prediction models. Ultimately, the insights gleaned from this study hold substantial promise for healthcare providers, aiding in treatment, accurate diagnosis, and early disease detection, ultimately contributing to improved patient survival rates.

Reference

[1] H. Zhang and Z. Xu, "The correlation between physical inactivity and students' health based on data mining and related influencing factors," *Math. Biosci. Eng.*, vol. 20, no. 4, pp. 6735–6750, 2023, doi: 10.3934/mbe.2023290.

[2] J. C. Arnold, A. Mühling, and K. Kremer, "Exploring core ideas of procedural understanding in scientific inquiry using educational data mining," *Res. Sci. Technol. Educ.*, vol. 41, no. 1, pp. 372–392, 2021, doi: 10.1080/02635143.2021.1909552.

[3] B. Qin, P. Peng, J. Zhang, H. Wang, and K. Ma, "A framework and prototype system in support of workflow collaboration and knowledge mining for manufacturing value chains," *IET Collab. Intell. Manuf.*, vol. 5, no. 1, pp. 1–11, 2023, doi: 10.1049/cim2.12073.

[4] X. Hu *et al.*, "GAT-COBO : Cost-Sensitive Graph Neural Network for Telecom Fraud Detection," vol. 14, no. 8, pp. 1–16, 2022.

[5] J. Dulan, "Video Surveillance Systems with Presentation of Existing VSS," no. January, 2023, doi: 10.15680/IJIRSET.2023.1201008.

[6] C. Chang and M. Mcaleer, "Informatics , Data Mining , Econometrics and Financial Economics : A Connection *," pp. 1–12, 2015.

[7] W. K. Adu, P. Appiahene, and S. Afrifa, "VAR , ARIMAX and ARIMA models for nowcasting unemployment rate in Ghana using Google trends," *J. Electr. Syst. Inf. Technol.*, pp. 1–16, 2023, doi: 10.1186/s43067-023-00078-1.

[8] S. Swati, "Evaluation of Current Investigations and Future Directions in White- Collar Crime," vol. 7956, pp. 77–82, 2023, doi: 10.36348/sijlcj.2023.v06i02.003.

[9] K. Tsui, V. C. P. Chen, and Y. A. Aslandogan, "Data Mining Methods and Applications," pp. 1–32, 2023.

[10] Y. Ye, M. Wang, and M. Huang, *Construction of Scientific Research Training System and Investigate Its Efficiency on Teaching Reform in the Medical Big Data*, vol. 2. Atlantis Press International BV, 2023.

[11] S. G. Jacob, "Data Mining in Clinical Data Sets : A Review," vol. 4, no. 6, pp. 15–26, 2012.
[12] E. Getzen, L. Ungar, D. Mowery, X. Jiang, and Q. Long, "Mining for Equitable Health : Assessing the Impact of Missing Data in Electronic Health Records," pp. 1–29, 2022.
[13] S. Patel and H. Patel, "Survey of data mining techniques used in healthcare domain," vol. 6, no. 1, pp. 53–60, 2016.
[14] S. Yy, *Diagnostic and Treatment Advances in COVID-19 and SARS-CoV-2*, no. February. 2023.
[15] S. Babu, "Heart Disease Diagnosis Using Data Mining Technique," pp. 750–753, 2017.
[16] B. Luo and Z. Chen, "A CpG-based prediction model for the diagnosis of hepatocellular carcinoma patients," 2023.
[17] M. Hadizadeh, M. A. Islam, W. Viriyasitavat, and M. U. Khandaker, "Federated Learning Approach to Protect Healthcare Data over Big Data Scenario Big Data in Health Care," 2022.
[18] X. Wang, "Influence Model of Analysing the Effect of Mental Health Level Based on Big Data Mining System," vol. 2022, 2022.
[19] I. Harrow, R. Balakrishnan, H. Küçük, and T. Plasterer, "Maximising data value for biopharma through FAIR and quality implementation : FAIR plus Q," *Drug Discov. Today*, vol. 27, no. 5, pp. 1441–1447, 2022, doi: 10.1016/j.drudis.2022.01.006.
[20] S. A. Varun, G. Mounika, P. K. Sahoo, and K. Eswaran, "Efficient System for Heart Disease Prediction by applying Logistic Regression," vol. 8491, pp. 13–16, 2019.
[21] N. Pahwa and A. Kalra, "Heart Disease Analysis Using Crossover with Multi- Class Support Vector Machine Method," vol. 1, pp. 494–501, 2019.
[22] H. B. F. David and S. A. Belcy, "HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES," vol. 6956, no. October, pp. 1817–1823, 2018, doi: 10.21917/ijsc.2018.0253.
[23] M. Srivenkatesh, "Prediction of Cardiovascular Disease using Machine Learning Algorithms," no. 3, pp. 2404–2414, 2020, doi: 10.35940/ijeat.B3986.029320.
[24] S. Ware, S. Rakesh, and B. Choudhary, "Heart Attack Prediction by using Machine Learning Techniques," no. 5, pp. 1577–1580, 2020, doi: 10.35940/ijrte.D9439.018520.
[25] N. Vineeth, "Heart Disease Prediction with Machine Learning Approaches," vol. 9, no. 7, pp. 1454–1458, 2020, doi: 10.21275/SR20724113128.
[26] M. Kamboj, "Heart Disease Prediction with Machine Learning Approaches," vol. 9, no. 7, pp. 1454–1458, 2018, doi: 10.21275/SR20724113128.
[27] K. A. Amen and M. Mahmoud, "Machine Learning for Multiple Stage Heart Disease Prediction MACHINE LEARNING FOR MULTIPLE STAGE," no. September, 2020, doi: 10.5121/csit.2020.101118.
[28] R. A. Anggraini, "Algoritma Naïve Bayes Dengan Backward Elimination Pada Dataset Breast Cancer," vol. 23, no. 1, pp. 87–94, 2023.
[29] A. Al Amin and S. Parvin, "Improving The Performance of K-Nearest Neighbor Algorithm by Reducing The Attributes of Dataset Using Gain Ratio Improving The Performance of K-Nearest Neighbor Algorithm by Reducing The Attributes of Dataset Using Gain Ratio," *J. Phys. Conf. Ser.*, 2020, doi: 10.1088/1742-6596/1566/1/012090.
[30] N. Chumuang, P. Pramkeaw, and A. Farooq, "Electrical Impedance Of Breast' s Tissue Classification By Using Bootstrap Aggregating," *2019 15th Int. Conf. Signal-Image Technol. Internet-Based Syst.*, pp. 551–556, 2019, doi: 10.1109/SITIS.2019.00093.
[31] Y. K. Arbawa, R. A. D. Pisefty, and F. A. Bachtiar, "Wound Classifications Of Breast Tissues with Electrical Impedance Spectroscopy (EIS): Comparison of LVQ and GA-LVQ," in *2019 5th International Conference on Science in Information Technology (ICSITech)*, 2019, pp. 229–234, doi: 10.1109/ICSITech46713.2019.8987447.
[32] A. J. Guimarães, P. Vitor, and D. C. Souza, "Pruning Fuzzy Neural Network Applied to the Construction of Expert Systems to Aid in the Diagnosis of the Treatment of Cryotherapy and Immunotherapy," pp. 1–20, 2019, doi: 10.3390/bdcc3020022.
[33] S. Rahayu, F. S. Nugraha, and I. Komputer, “Analisis tingkat keberhasilan cryoterapy menggunakan

neural network," vol. 15, no. 2, pp. 141–148, 2019, doi: 10.33480/pilar.v15i2.599.
[34] A. I. Hossain, S. Sikder, A. Das, and A. Dey, "Applying Machine Learning Classifiers on ECG Dataset for Predicting Heart Disease," no. July, 2021, doi: 10.1109/ACMI53878.2021.9528169.
[35] Y. D. Austria, M. L. Goh, L. Sta. Maria Jr., J.-A. Lalata, J. E. Goh, and H. Vicente, "Comparison of Machine Learning Algorithms in Breast Cancer Prediction Using the Coimbra Dataset," *Int. J. Simul. Syst. Sci. Technol.*, pp. 1–8, 2019, doi: 10.5013/ijssst.a.20.s2.23.
[36] F. Khozeimeh, R. Alizadehsani, M. Roshanzamir, A. Khosravi, P. Layegh, and S. Nahavandi, "An expert system for selecting wart treatment method," *Comput. Biol. Med.*, vol. 81, no. January, pp. 167–175, 2017, doi: 10.1016/j.compbiomed.2017.01.001.
[37] R. Raja, I. Mukherjee, and B. K. Sarkar, "A Machine Learning-Based Prediction Model for Preterm Birth in Rural India," vol. 2021, 2021.
[38] D. M. Maslove, T. Podchiyska, and H. J. Lowe, "Discretisation of continuous features in clinical datasets," pp. 544–553, 2013, doi: 10.1136/amiajnl-2012-000929.
[39] M. Ilievski *et al.*, "Design Space of Behaviour Planning for Autonomous Driving," 2019.
[40] Z. Yuan, H. Chen, T. Li, Z. Yu, B. Sang, and C. Luo, "Unsupervised attribute reduction for mixed data based on fuzzy rough sets," *Inf. Sci. (Ny).*, vol. 572, pp. 67–87, 2021, doi: 10.1016/j.ins.2021.04.083.
[41] S. Jahangiri, "An Improved Naïve Bayes Approach to Diagnose Cardiovascular Disease : A Case Study," pp. 0–36, 2023.
[42] S. Wang, J. Ren, X. Lian, R. Bai, and X. Jiang, "Boosting the Discriminant Power of Naive Bayes," 2022.
[43] V. K. Prasad, *Intelligent Systems and Sustainable Computing*. 2021.
[44] S. Zhang, S. Member, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN Classification With Different Numbers of Nearest Neighbors," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, 2018, doi: 10.1109/TNNLS.2017.2673241.
[45] T. District, T. District, T. District, and T. District, "Detection of DDoS attack in SDN environment using KNN algorithm," vol. 9, no. 2, pp. 252–257, 2022.
[46] S. Zhang, P. Liao, H. Q. Ye, and Z. Zhou, "Dynamic Marketing Resource Allocation with Two-Stage Decisions," *J. Theor. Appl. Electron. Commer. Res.*, vol. 17, no. 1, pp. 327–344, 2022, doi: 10.3390/jtaer17010017.
[47] L. Yang *et al.*, "Comparative Analysis of the Optimised KNN, SVM, and Ensemble DT Models Using Bayesian Optimization for Predicting Pedestrian Fatalities: An Advance towards Realising the Sustainable Safety of Pedestrians," *Sustain.*, vol. 14, no. 17, 2022, doi: 10.3390/su141710467.
[48] U. e. Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, "An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study," *Sensors*, vol. 22, no. 14, pp. 1–15, 2022, doi: 10.3390/s22145247.
[49] H. Zermane and A. Drardja, "Development of an efficient cement production monitoring system based on the improved random forest algorithm," *Int. J. Adv. Manuf. Technol.*, vol. 120, no. 3–4, pp. 1853–1866, 2022, doi: 10.1007/s00170-022-08884-z.
[50] R. Natras, B. Soja, and M. Schmidt, "Ensemble Machine Learning of Random Forest , AdaBoost and XGBoost for Forecasting Vertical Total Electron Content of the Ionosphere Ensemble Machine Learning of Random Forest , AdaBoost and XGBoost for Forecasting Vertical Total Electron Content of the Io," *J. Geod.*, pp. 1–34, 2022.
[51] M. Tiwari, J. Lee, and M. J. Zhang, "MABSplit : Faster Forest Training Using Multi-Armed Bandits," *Nips*, no. NeurIPS, pp. 1–15, 2022.
[52] O. Stitini, S. Kaloun, and O. Bencharef, "Towards the Detection of Fake News on Social Networks Contributing to the Improvement of Trust and Transparency in Recommendation Systems: Trends and Challenges," *Inf.*, vol. 13, no. 3, 2022, doi: 10.3390/info13030128.
[53] K. Raza, *Chapter 8 - Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule*. Elsevier Inc., 2019.
[54] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *Eng. Appl. Artif. Intell.*, vol. 115, 2022, doi: 10.1016/j.engappai.2022.105151.

[55] I. Castiglioni *et al.*, "AI applications to medical images: From machine learning to deep learning," *Phys. Medica*, vol. 83, no. November 2020, pp. 9–24, 2021, doi: 10.1016/j.ejmp.2021.02.006.

[56] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[57] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021.

[58] S. Wan, Y. Liang, Y. Zhang, M. Guizani, Deep multi-layer perceptron classifier for behavior analysis to estimate Parkinson's disease severity using smartphones, IEEE Access 6 (2018) 36825–36833.

[59] E. Ali˘ckovi'c, A. Subasi, Breast cancer diagnosis using GA feature selection and Rotation Forest, Neural Comput. Applic. 28 (4) (2017) 753–763.

[60] H. Wang, B. Zheng, S.W. Yoon, H.S. Ko, A support vector machine-based ensemble algorithm for breast cancer diagnosis, Europ. J. Oper. Res. 267 (2) (2018) 687–699.

[61] S.M.S. Shah, S. Batool, I. Khan, M.U. Ashraf, S.H. Abbas, S.A. Hussain, Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis, Physica A 482 (2017) 796–807.

[62] M. Abdar, W. Książek, ˙ U.R. Acharya, R.S. Tan, V. Makarenkov, P. Pławiak, A new machine learning technique for an accurate diagnosis of coronary artery disease, Comput. Methods Programs Biomed. 179 (2019) 104992.

[63] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, R. Nour, An intelligent learning system based on random search algorithm and optimised random forest model for improved heart disease detection, IEEE Access 7 (2019) 180235–180243.

[64] C. Chen, J. Wang, C. Chen, J. Tang, X. Lv, C. Ma, Rapid and efficient screening of human papillomavirus by Raman spectroscopy based on GA-SVM, Optik 210 (2020) 164514.

[65] N. Mahendran, P. Vincent, K. Srinivasan, V. Sharma, D. Jayakody, Realising a stacking generalisation model to improve the prediction accuracy of major depressive disorder in adults, IEEE Access 8 (2020) 49509–49522.

[66] F. Zhang, Y. Wei, J. Liu, Y. Wang, W. Xi, Y. Pan, Identification of Autism spectrum disorder based on a novel feature selection method and Variational Autoencoder. arXiv preprint arXiv:2204.03654, 2022.

[67] J. Chawla, P. Rana, S.B. Kaur, H. Singh, K. Yuvaraj, M. Murugappan, A decision support system for automated diagnosis of Parkinson's disease from EEG using FAWT and entropy features, Biomed. Signal Process. Control 79 (2023).